

Package ‘ggmlR’

July 22, 2026

Type Package

Title 'GGML' Tensor Operations for Machine Learning

Version 0.8.2

Description Provides 'R' bindings to the 'GGML' tensor library for machine learning, optimized for 'Vulkan' GPU acceleration with a transparent CPU fallback. The package features a 'Keras'-like sequential API and a 'PyTorch'-style 'autograd' engine for building, training, and deploying neural networks. Key capabilities include high-performance 5D tensor operations, 'f16' precision, and efficient quantization. It supports native 'ONNX' model import (50+ operators) and 'GGUF' weight loading from the 'llama.cpp' and 'Hugging Face' ecosystems. Designed for zero-overhead inference via dedicated weight buffering, it integrates seamlessly as a 'parsnip' engine for 'tidymodels' and provides first-class learners for the 'mlr3' framework.
See <<https://github.com/ggml-org/ggml>> for more information about the underlying library.

Depends R (>= 4.1.0)

Imports generics, R6, methods, stats

License MIT + file LICENSE

URL <https://github.com/Zabis13/ggmlR>

BugReports <https://github.com/Zabis13/ggmlR/issues>

Encoding UTF-8

SystemRequirements C++17, GNU make, libvulkan-dev, glslc (optional, for GPU on Linux), 'Vulkan' 'SDK' (optional, for GPU on Windows)

Suggests testthat (>= 3.0.0), mlr3 (>= 0.21.0), paradox, digest, parsnip, tibble, rlang, dials, lgr, knitr, rmarkdown, Matrix, Seurat, SeuratObject, RSpectra, irlba, uwot, FNN, SingleCellExperiment, SummarizedExperiment, S4Vectors, withr, hardhat, rsample, tune, workflows

VignetteBuilder knitr, rmarkdown

RoxygenNote 7.3.3

Config/testthat/edition 3

Config/build/clean-inst-doc false

NeedsCompilation yes

Author Yuri Baramykov [aut, cre] (ORCID:

<<https://orcid.org/0009-0000-7627-4217>>),

Georgi Gerganov [ctb, cph] (Author of the GGML library),

Jeffrey Quesnelle [ctb, cph] (Contributor to ops.cpp),

Bowen Peng [ctb, cph] (Contributor to ops.cpp),

Mozilla Foundation [ctb, cph] (Author of llamafile/sgemm.cpp)

Maintainer Yuri Baramykov <lbsbmsu@mail.ru>

Repository CRAN

Date/Publication 2026-07-22 11:50:02 UTC

Contents

ag_add	15
ag_batch_norm	15
ag_clamp	16
ag_cross_entropy_loss	16
ag_dataloader	17
ag_default_device	18
ag_default_dtype	18
ag_device	19
ag_dropout	19
ag_dtype	20
ag_embedding	20
ag_eval	21
ag_exp	21
ag_gradcheck	22
ag_linear	23
ag_load_model	23
ag_log	24
ag_matmul	25
ag_mean	25
ag_mse_loss	26
ag_mul	26
ag_multihead_attention	27
ag_param	28
ag_pow	29
ag_relu	29
ag_reshape	30
ag_save_model	30
ag_scale	31
ag_sequential	32
ag_sigmoid	32
ag_softmax	33

ag_softmax_cross_entropy_loss	33
ag_sub	34
ag_sum	34
ag_tanh	35
ag_tensor	35
ag_to_device	36
ag_train	36
ag_transpose	37
as.matrix.ggml_matrix	37
as_gpu_matrix	38
augment.ggmlr_parsnip_model	38
backward	39
clip_grad_norm	40
compile.ggml_sequential_model	41
dequantize_row_iq2_xxs	42
dequantize_row_mxfp4	43
dequantize_row_nvfp4	43
dequantize_row_q1_0	44
dequantize_row_q2_K	45
dequantize_row_q4_0	46
dequantize_row_tq1_0	46
dp_train	47
evaluate.ggml_sequential_model	49
fit.ggml_sequential_model	50
ggml_abort_is_r_enabled	51
ggml_abs	52
ggml_abs_inplace	52
ggml_add	53
ggml_add1	54
ggml_add_inplace	55
ggml_add_rel_pos	55
ggml_apply	56
ggml_arange	57
ggml_are_same_layout	57
ggml_are_same_shape	58
ggml_are_same_stride	58
ggml_argmax	59
ggml_argsort	60
ggml_backend_alloc_ctx_tensors	60
ggml_backend_buffer_clear	61
ggml_backend_buffer_free	62
ggml_backend_buffer_get_size	62
ggml_backend_buffer_get_usage	63
ggml_backend_buffer_is_host	64
ggml_backend_buffer_is_multi_buffer	65
ggml_backend_buffer_name	66
ggml_backend_buffer_reset	66
ggml_backend_buffer_set_usage	67

ggml_backend_buffer_usage_any	68
ggml_backend_buffer_usage_compute	69
ggml_backend_buffer_usage_weights	70
ggml_backend_cpu_init	70
ggml_backend_cpu_set_n_threads	71
ggml_backend_device_register	71
ggml_backend_device_type_accel	72
ggml_backend_device_type_cpu	73
ggml_backend_device_type_gpu	74
ggml_backend_device_type_igpu	75
ggml_backend_dev_by_name	75
ggml_backend_dev_by_type	76
ggml_backend_dev_count	77
ggml_backend_dev_description	78
ggml_backend_dev_get	79
ggml_backend_dev_get_props	80
ggml_backend_dev_init	81
ggml_backend_dev_memory	82
ggml_backend_dev_name	83
ggml_backend_dev_offload_op	84
ggml_backend_dev_supports_bufi	85
ggml_backend_dev_supports_op	86
ggml_backend_dev_type	87
ggml_backend_event_free	88
ggml_backend_event_new	89
ggml_backend_event_record	90
ggml_backend_event_synchronize	91
ggml_backend_event_wait	92
ggml_backend_free	93
ggml_backend_get_device	93
ggml_backend_graph_compute	94
ggml_backend_graph_compute_async	95
ggml_backend_graph_plan_compute	96
ggml_backend_graph_plan_create	97
ggml_backend_graph_plan_free	98
ggml_backend_init_best	99
ggml_backend_init_by_name	99
ggml_backend_init_by_type	100
ggml_backend_load	101
ggml_backend_load_all	102
ggml_backend_meta_device	103
ggml_backend_multi_buffer_alloc_buffer	104
ggml_backend_multi_buffer_set_usage	105
ggml_backend_name	106
ggml_backend_register	107
ggml_backend_reg_by_name	108
ggml_backend_reg_count	109
ggml_backend_reg_dev_count	109

ggml_backend_reg_dev_get	110
ggml_backend_reg_get	111
ggml_backend_reg_name	112
ggml_backend_sched_alloc_graph	113
ggml_backend_sched_free	113
ggml_backend_sched_get_backend	114
ggml_backend_sched_get_n_backends	115
ggml_backend_sched_get_n_copies	115
ggml_backend_sched_get_n_splits	116
ggml_backend_sched_get_tensor_backend	116
ggml_backend_sched_graph_compute	117
ggml_backend_sched_graph_compute_async	118
ggml_backend_sched_new	118
ggml_backend_sched_reserve	119
ggml_backend_sched_reset	120
ggml_backend_sched_set_tensor_backend	121
ggml_backend_sched_synchronize	121
ggml_backend_synchronize	122
ggml_backend_tensor_copy_async	123
ggml_backend_tensor_get_and_sync	124
ggml_backend_tensor_get_async	124
ggml_backend_tensor_get_data	125
ggml_backend_tensor_get_f32_first	126
ggml_backend_tensor_set_async	126
ggml_backend_tensor_set_data	127
ggml_backend_unload	128
ggml_batch_norm	129
ggml_blk_size	129
ggml_build_forward_expand	130
ggml_callback_early_stopping	131
ggml_can_repeat	132
ggml_ceil	132
ggml_ceil_inplace	133
ggml_clamp	133
ggml_compile.ggml_functional_model	134
ggml_concat	135
ggml_cont	136
ggml_conv_1d	136
ggml_conv_1d_dw	137
ggml_conv_2d	138
ggml_conv_2d_direct	138
ggml_conv_2d_dw	139
ggml_conv_2d_dw_direct	140
ggml_conv_transpose_1d	141
ggml_conv_transpose_2d_p0	141
ggml_cos	142
ggml_count_equal	142
ggml_cpu_add	143

ggml_cpu_features	144
ggml_cpu_get_rvv_vlen	145
ggml_cpu_get_sve_cnt	145
ggml_cpu_has_amx_int8	146
ggml_cpu_has_arm_fma	146
ggml_cpu_has_avx	147
ggml_cpu_has_avx2	148
ggml_cpu_has_avx512	148
ggml_cpu_has_avx512_bf16	149
ggml_cpu_has_avx512_vbmi	149
ggml_cpu_has_avx512_vnni	150
ggml_cpu_has_avx_vnni	151
ggml_cpu_has_bmi2	151
ggml_cpu_has_dotprod	152
ggml_cpu_has_f16c	152
ggml_cpu_has_fma	153
ggml_cpu_has_fp16_va	154
ggml_cpu_has_llamafile	154
ggml_cpu_has_matmul_int8	155
ggml_cpu_has_neon	155
ggml_cpu_has_riscv_v	156
ggml_cpu_has_sme	157
ggml_cpu_has_sse3	157
ggml_cpu_has_ssse3	158
ggml_cpu_has_sve	159
ggml_cpu_has_vsx	159
ggml_cpu_has_vxe	160
ggml_cpu_has_wasm_simd	160
ggml_cpu_mul	161
ggml_cpy	162
ggml_crossprod	162
ggml_cycles	163
ggml_cycles_per_ms	163
ggml_default_mlp	164
ggml_dense	165
ggml_diag	166
ggml_diag_mask_inf	167
ggml_diag_mask_inf_inplace	168
ggml_diag_mask_zero	168
ggml_div	169
ggml_div_inplace	170
ggml_dup	170
ggml_dup_inplace	171
ggml_dup_tensor	171
ggml_element_size	172
ggml_elu	172
ggml_elu_inplace	173
ggml_embedding	173

ggml_estimate_memory	174
ggml_evaluate.ggml_functional_model	175
ggml_exp	176
ggml_exp_inplace	177
ggml_extract	177
ggml_fit.ggml_functional_model	179
ggml_fit_opt	181
ggml_flash_attn_back	183
ggml_flash_attn_ext	184
ggml_floor	185
ggml_floor_inplace	186
ggml_free	186
ggml_freeze_weights	187
ggml_ftype_to_ggml_type	188
ggml_gallocr_alloc_graph	188
ggml_gallocr_free	189
ggml_gallocr_get_buffer_size	189
ggml_gallocr_new	190
ggml_gallocr_reserve	191
ggml_geglu	191
ggml_geglu_quick	192
ggml_geglu_split	192
ggml_gelu	193
ggml_gelu_erf	194
ggml_gelu_inplace	194
ggml_gelu_quick	195
ggml_get_f32	196
ggml_get_f32_nd	197
ggml_get_first_tensor	197
ggml_get_i32	198
ggml_get_i32_nd	198
ggml_get_layer	199
ggml_get_max_tensor_size	199
ggml_get_mem_size	200
ggml_get_name	201
ggml_get_next_tensor	201
ggml_get_no_alloc	202
ggml_get_n_threads	202
ggml_get_op_params	203
ggml_get_op_params_f32	203
ggml_get_op_params_i32	204
ggml_get_rel_pos	204
ggml_get_rows	205
ggml_get_rows_back	206
ggml_get_unary_op	206
ggml_glu	207
GGML_GLU_OP_REGLU	208
ggml_glu_split	209

ggml_graph_compute	209
ggml_graph_compute_with_ctx	210
ggml_graph_dump_dot	211
ggml_graph_get_tensor	212
ggml_graph_node	212
ggml_graph_n_nodes	213
ggml_graph_overhead	213
ggml_graph_print	214
ggml_graph_reset	214
ggml_graph_view	215
ggml_group_norm	216
ggml_group_norm_inplace	216
ggml_gru	217
ggml_hardsigmoid	218
ggml_hardswish	219
ggml_im2col	219
ggml_init	220
ggml_init_auto	221
ggml_inject	222
ggml_input	223
ggml_is_available	224
ggml_is_contiguous	224
ggml_is_contiguously_allocated	225
ggml_is_contiguous_0	225
ggml_is_contiguous_1	226
ggml_is_contiguous_2	226
ggml_is_contiguous_channels	227
ggml_is_contiguous_rows	228
ggml_is_permuted	228
ggml_is_quantized	229
ggml_is_transposed	230
ggml_l2_norm	230
ggml_l2_norm_inplace	231
ggml_layer_add	232
ggml_layer_batch_norm	232
ggml_layer_concatenate	233
ggml_layer_conv_1d	234
ggml_layer_conv_2d	235
ggml_layer_dense	236
ggml_layer_dropout	237
ggml_layer_embedding	239
ggml_layer_flatten	240
ggml_layer_global_average_pooling_2d	240
ggml_layer_global_max_pooling_2d	241
ggml_layer_gru	242
ggml_layer_lstm	243
ggml_layer_max_pooling_2d	244
ggml_leaky_relu	245

ggml_load_model	246
ggml_load_weights	246
ggml_log	247
ggml_log_inplace	248
ggml_log_is_r_enabled	248
ggml_log_set_default	249
ggml_log_set_r	249
ggml_lstm	250
ggml_marshal_model	250
ggml_matmul	251
ggml_matmul_f64	252
ggml_matrix-class	253
ggml_mean	254
ggml_model	255
ggml_model_backend	256
ggml_model_sequential	256
ggml_mul	257
ggml_mul_inplace	258
ggml_mul_mat	258
ggml_mul_mat_id	259
ggml_nbytes	260
ggml_neg	261
ggml_neg_inplace	262
ggml_nelements	262
ggml_new_f32	263
ggml_new_i32	264
ggml_new_tensor	264
ggml_new_tensor_1d	265
ggml_new_tensor_2d	266
ggml_new_tensor_3d	266
ggml_new_tensor_4d	267
ggml_norm	268
ggml_norm_inplace	269
ggml_nrows	269
ggml_n_dims	270
ggml_ops_registry	270
ggml_opt_alloc	271
ggml_opt_context_optimizer_type	272
ggml_opt_dataset_data	272
ggml_opt_dataset_free	273
ggml_opt_dataset_get_batch	274
ggml_opt_dataset_init	275
ggml_opt_dataset_labels	276
ggml_opt_dataset_ndata	276
ggml_opt_dataset_shuffle	277
ggml_opt_dataset_weights	278
ggml_opt_default_params	279
ggml_opt_epoch	279

ggml_opt_eval	281
ggml_opt_fit	282
ggml_opt_free	283
ggml_opt_get_lr	284
ggml_opt_grad_acc	285
ggml_opt_init	285
ggml_opt_init_for_fit	287
ggml_opt_inputs	288
ggml_opt_labels	288
ggml_opt_loss	289
ggml_opt_loss_type_cross_entropy	290
ggml_opt_loss_type_mean	290
ggml_opt_loss_type_mse	291
ggml_opt_loss_type_sum	292
ggml_opt_loss_type_weighted_mse	292
ggml_opt_ncorrect	293
ggml_opt_optimizer_name	294
ggml_opt_optimizer_type_adamw	294
ggml_opt_optimizer_type_sgd	295
ggml_opt_outputs	296
ggml_opt_pred	296
ggml_opt_prepare_alloc	297
ggml_opt_reset	298
ggml_opt_result_accuracy	299
ggml_opt_result_free	299
ggml_opt_result_init	300
ggml_opt_result_loss	301
ggml_opt_result_ndata	301
ggml_opt_result_pred	302
ggml_opt_result_reset	303
ggml_opt_set_lr	304
ggml_opt_static_graphs	304
ggml_op_can_inplace	305
ggml_op_desc	306
ggml_op_name	306
ggml_op_symbol	307
ggml_out_prod	307
ggml_pad	308
ggml_pad_reflect_1d	309
ggml_permute	309
ggml_pool_1d	310
ggml_pool_2d	311
ggml_pop_layer	312
ggml_pp_dp_forward	313
ggml_pp_forward	314
ggml_predict_ggml_functional_model	315
ggml_predict_classes	316
ggml_print_mem_status	317

ggml_print_objects	317
ggml_quantize_chunk	318
ggml_quantize_free	319
ggml_quantize_init	319
ggml_quantize_requires_imatrix	320
ggml_quant_block_info	320
ggml_reglu	321
ggml_reglu_split	321
ggml_relu	322
ggml_relu_inplace	323
ggml_repeat	323
ggml_repeat_back	324
ggml_reset	324
ggml_reshape_1d	325
ggml_reshape_2d	326
ggml_reshape_3d	326
ggml_reshape_4d	327
ggml_result	328
ggml_rms_norm	328
ggml_rms_norm_back	329
ggml_rms_norm_inplace	330
ggml_roll	330
ggml_rope	331
ggml_rope_ext	332
ggml_rope_ext_back	334
ggml_rope_ext_inplace	335
ggml_rope_inplace	336
ggml_rope_multi	336
ggml_rope_multi_inplace	338
ggml_round	339
ggml_round_inplace	339
ggml_run	340
ggml_save_model	340
ggml_save_weights	341
ggml_scale	342
ggml_scale_inplace	342
ggml_schedule_cosine_decay	343
ggml_schedule_reduce_on_plateau	344
ggml_schedule_step_decay	345
ggml_set	345
ggml_set_1d	346
ggml_set_2d	346
ggml_set_abort_callback_default	347
ggml_set_abort_callback_r	347
ggml_set_f32	348
ggml_set_f32_nd	349
ggml_set_i32	349
ggml_set_i32_nd	350

ggml_set_input	350
ggml_set_name	351
ggml_set_no_alloc	351
ggml_set_n_threads	352
ggml_set_omp_threads	352
ggml_set_op_params	353
ggml_set_op_params_f32	353
ggml_set_op_params_i32	354
ggml_set_output	355
ggml_set_param	355
ggml_set_seed	356
ggml_set_zero	357
ggml_sgn	357
ggml_sigmoid	358
ggml_sigmoid_inplace	359
ggml_silu	359
ggml_silu_back	360
ggml_silu_inplace	360
ggml_sin	361
ggml_softplus	362
ggml_softplus_inplace	362
ggml_soft_max	363
ggml_soft_max_ext	364
ggml_soft_max_ext_back	365
ggml_soft_max_ext_back_inplace	365
ggml_soft_max_ext_inplace	366
ggml_soft_max_inplace	367
GGML_SORT_ORDER_ASC	367
ggml_sqr	368
ggml_sqrt	369
ggml_sqrt_inplace	370
ggml_sqr_inplace	370
ggml_step	371
ggml_sub	371
ggml_sub_inplace	372
ggml_sum	373
ggml_sum_rows	373
ggml_swiglu	374
ggml_swiglu_split	375
ggml_tanh	375
ggml_tanh_inplace	376
ggml_task	376
ggml_tcrossprod	377
ggml_tensor_copy	378
ggml_tensor_nb	378
ggml_tensor_num	379
ggml_tensor_overhead	379
ggml_tensor_set_f32_scalar	380

ggml_tensor_shape	380
ggml_tensor_type	381
ggml_test	381
ggml_timestep_embedding	382
ggml_time_init	382
ggml_time_ms	383
ggml_time_us	383
ggml_top_k	384
ggml_tp_dp_forward	384
ggml_training_history	386
ggml_transpose	386
GGML_TYPE_F32	387
ggml_type_name	389
ggml_type_size	389
ggml_type_sizef	390
ggml_unary_op_name	390
ggml_unfreeze_weights	391
ggml_unmarshal_model	392
ggml_upscale	392
ggml_used_mem	394
ggml_version	394
ggml_view_1d	395
ggml_view_2d	395
ggml_view_3d	396
ggml_view_4d	397
ggml_view_tensor	397
ggml_vulkan_available	398
ggml_vulkan_backend_name	398
ggml_vulkan_device_caps	399
ggml_vulkan_device_count	399
ggml_vulkan_device_description	400
ggml_vulkan_device_groups	400
ggml_vulkan_device_memory	401
ggml_vulkan_free	402
ggml_vulkan_hard_exit_available	402
ggml_vulkan_init	403
ggml_vulkan_is_backend	404
ggml_vulkan_list_devices	404
ggml_vulkan_p2p_selftest	405
ggml_vulkan_shutdown	406
ggml_vulkan_split_buffer_type	408
ggml_vulkan_split_mul_mat	409
ggml_vulkan_split_row_ranges	410
ggml_vulkan_status	411
ggml_win_part	412
ggml_win_unpart	412
ggml_with_temp_ctx	413
gguf_free	413

gguf_load	414
gguf_metadata	414
gguf_tensor_data	415
gguf_tensor_info	415
gguf_tensor_names	416
glance.ggmlr_parsnip_model	416
iq2xs_free_impl	417
iq2xs_init_impl	418
iq3xs_free_impl	418
iq3xs_init_impl	419
lr_scheduler_cosine	420
lr_scheduler_step	420
nn_topo_sort	421
onnx_device_info	422
onnx_inputs	422
onnx_load	423
onnx_run	424
onnx_summary	424
optimizer_adam	425
optimizer_sgd	425
plot.ggml_history	426
predict.ggml_sequential_model	426
print.ag_tensor	427
print.ggml_functional_model	427
print.ggml_history	428
print.ggml_sequential_model	428
print.onnx_model	429
quantize_iq2_xxs	429
quantize_mxfp4	430
quantize_nvfp4	431
quantize_q1_0	432
quantize_q2_K	433
quantize_q4_0	434
quantize_row_iq3_xxs_ref	435
quantize_row_mxfp4_ref	436
quantize_row_q2_K_ref	436
quantize_row_q4_0_ref	437
quantize_row_tq1_0_ref	438
quantize_tq1_0	439
rope_types	439
RunGGML.SingleCellExperiment	441
summary.ggml_sequential_model	443
tidy.ggmlr_parsnip_model	444
with_grad_tape	444

ag_add	<i>Element-wise addition with broadcasting</i>
--------	--

Description

Computes $A + B$. If B is $[m, 1]$ and A is $[m, n]$, B is broadcast across columns (useful for bias vectors).

Usage

```
ag_add(A, B)
```

Arguments

A	ag_tensor or numeric matrix
B	ag_tensor or numeric matrix (may be $[m, 1]$ or $[1, n]$ for broadcasting)

Value

ag_tensor

ag_batch_norm	<i>Create a Batch Normalisation layer</i>
---------------	---

Description

Normalises each feature (row) over the batch dimension. Learnable scale gamma $[F, 1]$ and shift beta $[F, 1]$.

Usage

```
ag_batch_norm(num_features, eps = 1e-05, momentum = 0.1)
```

Arguments

num_features	Number of features (rows of input)
eps	Numerical stability constant (default 1e-5)
momentum	Running-stats momentum (default 0.1)

Details

Training mode: use batch statistics; update running mean/var. **Eval mode:** use stored running statistics.

Value

An ag_batch_norm environment

Examples

```
bn <- ag_batch_norm(16L)
x  <- ag_tensor(matrix(rnorm(16 * 32), 16, 32))
out <- bn$forward(x)
```

ag_clamp	<i>Element-wise clamp</i>
----------	---------------------------

Description

Clamps values to [lo, hi]. Gradient is 1 inside the interval, 0 at the boundary (straight-through estimator).

Usage

```
ag_clamp(x, lo = -Inf, hi = Inf)
```

Arguments

x	ag_tensor
lo	Lower bound (default -Inf)
hi	Upper bound (default Inf)

Value

ag_tensor

ag_cross_entropy_loss	<i>Categorical Cross-Entropy loss</i>
-----------------------	---------------------------------------

Description

Generic CE: $-\text{sum}(\text{target} * \log(\text{pred})) / \text{batch_size}$. The gradient w.r.t. pred is $-\text{target} / \text{pred} / n$. Use ag_softmax_cross_entropy_loss() for the numerically stable combined softmax + CE (fused gradient $(p - y) / n$).

Usage

```
ag_cross_entropy_loss(pred, target)
```


Arguments

pred	ag_tensor [classes, batch_size] probabilities (any, not just softmax)
target	matrix [classes, batch_size] one-hot (or soft) labels

Value

scalar ag_tensor

ag_dataloader	<i>Create a mini-batch data loader</i>
---------------	--

Description

Returns an iterator environment. Each call to `$next_batch()` returns a named list `list(x, y)` with `ag_tensor` objects of shape `[features, batch_size] / [labels, batch_size]`. After the last batch, `$has_next()` returns `FALSE`; call `$reset()` (or start a new epoch via `$epoch()`) to reshuffle and restart.

Usage

```
ag_dataloader(x, y = NULL, batch_size = 32L, shuffle = TRUE, col_major = TRUE)
```

Arguments

x	Feature matrix [features, n_samples] or [n_samples, features] — see <code>col_major</code> .
y	Label matrix with the same convention.
batch_size	Integer batch size.
shuffle	Logical; if <code>TRUE</code> (default) shuffle at each <code>reset()</code> .
col_major	Logical; if <code>TRUE</code> (default) x and y are already [features, n] (ggml/ag convention). Set <code>FALSE</code> for row-major [n, features] (R/Keras convention) — they will be transposed automatically.

Value

An `ag_dataloader` environment

Examples

```
n <- 128L
x <- matrix(runif(4 * n), 4, n) # [4, 128] col-major
y <- matrix(runif(2 * n), 2, n)
dl <- ag_dataloader(x, y, batch_size = 32L)
dl$reset()
while (dl$has_next()) {
  batch <- dl$next_batch()
```

```
# batch$x: [4, 32], batch$y: [2, 32]
}
```

ag_default_device	<i>Return the current default compute device</i>
-------------------	--

Description

Return the current default compute device

Usage

```
ag_default_device()
```

Value

"cpu" or "gpu"

ag_default_dtype	<i>Return the current default dtype for GPU operations</i>
------------------	--

Description

Return the current default dtype for GPU operations

Usage

```
ag_default_dtype()
```

Value

"f32", "f16", or "bf16"

ag_device	<i>Set the default compute device for ag_* operations</i>
-----------	---

Description

Switches all subsequent ag_tensor / ag_param operations to run on the specified device. Calling ag_device("gpu") initialises the best available ggml backend (Vulkan, Metal, CUDA, or CPU fallback) the first time it is called.

Usage

```
ag_device(device)
```

Arguments

device	"cpu" (default) or "gpu"
--------	--------------------------

Value

Invisibly the previous device string

ag_dropout	<i>Create a Dropout layer</i>
------------	-------------------------------

Description

In training mode applies inverted dropout (random Bernoulli mask, scale by $1/(1-\text{rate})$ to preserve expected values). In eval mode is identity.

Usage

```
ag_dropout(rate)
```

Arguments

rate	Drop probability in [0, 1)
------	----------------------------

Value

An ag_dropout environment

Examples

```
drop <- ag_dropout(0.5)
x   <- ag_tensor(matrix(runif(8), 4, 2))
out <- drop$forward(x) # training mode by default
ag_eval(drop)
out2 <- drop$forward(x) # identity
```

ag_dtype	<i>Set the default floating-point precision for ag_* GPU operations</i>
----------	---

Description

Controls the dtype used when uploading tensors to the ggml backend. "bf16" halves memory usage vs "f32" with minimal accuracy loss. Backward pass always uses f32 R matrices regardless of this setting.

Usage

ag_dtype(dtype)

Arguments

dtype "f32" (default), "f16", or "bf16"

Value

Invisibly the previous dtype string

ag_embedding	<i>Create an Embedding layer</i>
--------------	----------------------------------

Description

Maps 0-based integer indices to dense vectors via table lookup. Input: integer matrix or vector of 0-based indices. Output: float tensor [dim, length(idx)].

Usage

ag_embedding(vocab_size, dim)

Arguments

vocab_size Vocabulary size
dim Embedding dimension

Details

Backward: scatter-add — only the looked-up rows accumulate gradient.

Value

An ag_embedding environment

Examples

```
emb <- ag_embedding(100L, 16L)
idx <- c(0L, 3L, 7L, 2L)
out <- emb$forward(idx)  # [16, 4]
```

ag_eval

Switch a layer or sequential model to eval mode

Description

Switch a layer or sequential model to eval mode

Usage

```
ag_eval(model)
```

Arguments

model An ag_sequential, ag_batch_norm, or ag_dropout layer

Value

The model/layer (invisibly)

ag_exp

Element-wise exponential

Description

Element-wise exponential

Usage

```
ag_exp(x)
```

Arguments

x ag_tensor

Value

ag_tensor

ag_gradcheck	<i>Numerical gradient check (like torch.autograd.gradcheck)</i>
--------------	---

Description

Compares analytical gradients (from backward()) with finite-difference numerical gradients for all input tensors with requires_grad = TRUE.

Usage

```
ag_gradcheck(
  fn,
  inputs,
  eps = 1e-05,
  atol = 1e-04,
  verbose = FALSE,
  quiet = FALSE
)
```

Arguments

fn	A function that takes a list of ag_tensor inputs and returns a scalar ag_tensor loss (must be used inside with_grad_tape).
inputs	Named list of ag_tensor objects. Only those with requires_grad = TRUE are checked.
eps	Finite-difference step size (default 1e-5).
atol	Absolute tolerance for pass/fail (default 1e-4).
verbose	Print per-element comparison (default FALSE).
quiet	Suppress per-parameter and overall status lines (default FALSE). Useful when calling from testthat tests to keep output clean.

Value

Invisibly TRUE if all gradients match, FALSE otherwise. When quiet = FALSE (default), prints a summary report.

Examples

```
W <- ag_param(matrix(runif(6), 2, 3))
x <- ag_tensor(matrix(runif(3), 3, 1))
ag_gradcheck(
  fn = function(ins) ag_mse_loss(ag_relu(ag_matmul(ins$W, ins$x)),
                                matrix(0, 2, 1)),
  inputs = list(W = W, x = x)
)
```

ag_linear	<i>Create a dense layer with learnable parameters</i>
-----------	---

Description

Returns a closure-based layer. Because `ag_param` uses environment semantics, the optimizer updates `W` and `b` in-place, and `forward()` always uses the latest weights.

Usage

```
ag_linear(in_features, out_features, activation = NULL)
```

Arguments

<code>in_features</code>	Input dimension
<code>out_features</code>	Output dimension
<code>activation</code>	"relu", "sigmoid", "tanh", "softmax", or NULL

Value

List with `W`, `b`, `forward(x)`, `params()`

Examples

```
layer <- ag_linear(4L, 8L, activation = "relu")
x      <- ag_tensor(matrix(runif(4 * 16), 4, 16))
out    <- layer$forward(x)
```

ag_load_model	<i>Load an autograd module from a saved state</i>
---------------	---

Description

Reconstructs an `ag_*` module saved with `ag_save_model`. The architecture is rebuilt by calling `model_fn` (either the one passed here, or the one stored inside the container at save time), and the saved parameter and buffer values are copied back by name.

Usage

```
ag_load_model(path, model_fn = NULL, device = NULL)
```

Arguments

path	File path written by ag_save_model .
model_fn	Optional zero-argument rebuild function. Required if no model_fn was stored at save time. If both are present, this argument takes precedence.
device	Optional device for the rebuilt module ("cpu" or "gpu"). If NULL (default), the current ag_device() is used by the rebuild.

Value

The reconstructed module with restored weights, in eval mode.

See Also

[ag_save_model](#)

Examples

```
build <- function() ag_sequential(ag_linear(4L, 8L), ag_linear(8L, 3L))
f <- tempfile(fileext = ".rds")
ag_save_model(build(), f, model_fn = build)
model <- ag_load_model(f)
```

ag_log	<i>Element-wise natural logarithm</i>
--------	---------------------------------------

Description

Element-wise natural logarithm

Usage

```
ag_log(x)
```

Arguments

x	ag_tensor
---	-----------

Value

ag_tensor

ag_matmul	<i>Matrix multiplication</i>
-----------	------------------------------

Description

Computes $A \%* \% B$ and records the operation on the gradient tape.

Usage

ag_matmul(A, B)

Arguments

- A ag_tensor or numeric matrix of shape [m, k]
- B ag_tensor or numeric matrix of shape [k, n]

Value

ag_tensor of shape [m, n]

ag_mean	<i>Mean of elements (or along a dim)</i>
---------	--

Description

Mean of elements (or along a dim)

Usage

ag_mean(x, dim = NULL, keepdim = FALSE)

Arguments

- x ag_tensor
- dim NULL (all), 1 (row-wise), or 2 (col-wise)
- keepdim Logical

Value

ag_tensor

ag_mse_loss	<i>Mean Squared Error loss</i>
-------------	--------------------------------

Description

Mean Squared Error loss

Usage

ag_mse_loss(pred, target)

Arguments

pred	ag_tensor [units, batch_size]
target	ag_tensor or matrix [units, batch_size]

Value

scalar ag_tensor

ag_mul	<i>Element-wise multiplication</i>
--------	------------------------------------

Description

Element-wise multiplication

Usage

ag_mul(A, B)

Arguments

A	ag_tensor or numeric matrix
B	ag_tensor or numeric matrix

Value

ag_tensor

ag_multihead_attention

Create a Multi-Head Attention layer

Description

Implements scaled dot-product multi-head attention as in "Attention Is All You Need" (Vaswani et al., 2017).

Usage

```
ag_multihead_attention(d_model, n_heads, dropout = 0, bias = TRUE)
```

Arguments

d_model	Model (embedding) dimension
n_heads	Number of attention heads. d_model must be divisible by n_heads.
dropout	Attention dropout probability (default 0, applied in training mode only)
bias	Logical: add bias to output projection (default TRUE)

Details

Calling convention (mirrors PyTorch nn.MultiheadAttention):

- layer\$forward(q) — self-attention ($k = v = q$)
- layer\$forward(q, k, v) — cross-attention

Tensor layout: [d_model, seq_len] — columns are tokens, consistent with the rest of the ag_* API.

Forward pass:

```
Q = W_q %%% q           [d_k * n_heads, seq_len]
K = W_k %%% k           [d_k * n_heads, seq_len]
V = W_v %%% v           [d_v * n_heads, seq_len]

for each head h:
  q_h = Q[h*d_k+1 : (h+1)*d_k, ]   [d_k, seq_len]
  k_h = K[h*d_k+1 : (h+1)*d_k, ]   [d_k, seq_len]
  v_h = V[h*d_v+1 : (h+1)*d_v, ]   [d_v, seq_len]
  A_h = softmax(t(q_h) %%% k_h / sqrt(d_k)) [seq_len, seq_len]
  if causal_mask: A_h[i,j] = 0 for j > i
  head_h = v_h %%% A_h             [d_v, seq_len]

concat = rbind(head_1, ..., head_H) [d_v*n_heads, seq_len]
out     = W_o %%% concat + b_o      [d_model, seq_len]
```

Value

An `ag_multihead_attention` environment with `$forward(q, k, v, causal_mask)` and `$parameters()`

Examples

```
# Self-attention
mha <- ag_multihead_attention(64L, 8L)
x  <- ag_tensor(matrix(rnorm(64 * 10), 64, 10)) # [d_model=64, seq_len=10]
out <- mha$forward(x)                          # [64, 10]

# Cross-attention
q  <- ag_tensor(matrix(rnorm(64 * 10), 64, 10))
kv <- ag_tensor(matrix(rnorm(64 * 15), 64, 15))
out <- mha$forward(q, kv, kv)

# Causal (GPT-style)
out <- mha$forward(x, causal_mask = TRUE)
```

ag_param

Create a parameter tensor (gradient tracked)

Description

Create a parameter tensor (gradient tracked)

Usage

```
ag_param(
  data,
  device = .ag_device_state$device,
  dtype = .ag_device_state$dtype
)
```

Arguments

data	Numeric matrix or vector
device	"cpu" (default) or "gpu"
dtype	Floating-point precision: "f32" (default), "f16", or "bf16". Ignored on CPU; controls upload precision on GPU.

Value

An `ag_tensor` with `requires_grad = TRUE`

ag_pow	<i>Element-wise power</i>
--------	---------------------------

Description

Element-wise power

Usage

ag_pow(x, p)

Arguments

- x ag_tensor
- p Numeric exponent (scalar, not tracked for gradients)

Value

ag_tensor

ag_relu	<i>ReLU activation</i>
---------	------------------------

Description

Applies the rectified linear unit: $\max(0, x)$.

Usage

ag_relu(x)

Arguments

- x ag_tensor

Value

ag_tensor

ag_reshape	<i>Reshape tensor</i>
------------	-----------------------

Description

Reshape tensor

Usage

```
ag_reshape(x, nrow, ncol)
```

Arguments

x	ag_tensor
nrow	New number of rows (use -1 to infer)
ncol	New number of columns (use -1 to infer)

Value

ag_tensor with new shape, same data

ag_save_model	<i>Save an autograd module's state to disk</i>
---------------	--

Description

Serializes the trainable parameters and persistent buffers of an ag_sequential (or single ag_* layer) module as a portable state dictionary of plain numeric matrices. This avoids serializing the live module (environments + closures), which is brittle across ggmlR versions and carries non-portable GPU pointers.

Usage

```
ag_save_model(model, path, model_fn = NULL)
```

Arguments

model	An ag_sequential module or a single ag_* layer exposing parameters().
path	File path to write (an RDS container).
model_fn	Optional zero-argument function that rebuilds the module architecture (fresh, untrained). If supplied, it is stored in the container so ag_load_model can rebuild without arguments. Must not capture GPU tensors in its enclosing environment.

Details

Reconstruction requires the architecture. Either pass model_fn here so it is stored in the file, or pass it later to [ag_load_model](#).

Value

path, invisibly.

See Also

[ag_load_model](#)

Examples

```
build <- function() ag_sequential(ag_linear(4L, 8L), ag_linear(8L, 3L))
model <- build()
ag_save_model(model, tempfile(fileext = ".rds"), model_fn = build)
```

ag_scale	<i>Scale tensor by a scalar constant</i>
----------	--

Description

Scale tensor by a scalar constant

Usage

```
ag_scale(x, scalar)
```

Arguments

x	ag_tensor
scalar	Numeric scalar (not tracked for gradients)

Value

ag_tensor

ag_sequential	<i>Create a sequential container of layers</i>
---------------	--

Description

Chains layers so that `forward(x)` passes `x` through each layer in order. `parameters()` collects all trainable params from all layers. `ag_train()` / `ag_eval()` propagate mode to stateful sub-layers.

Usage

```
ag_sequential(...)
```

Arguments

... Layer objects (`ag_linear`, `ag_dropout`, `ag_batch_norm`, `ag_embedding`) or a single list of layers.

Value

An `ag_sequential` environment

Examples

```
model <- ag_sequential(
  ag_linear(4L, 16L, activation = "relu"),
  ag_dropout(0.5),
  ag_linear(16L, 2L, activation = "softmax")
)
x <- ag_tensor(matrix(runif(4 * 8), 4, 8))
out <- model$forward(x)
```

ag_sigmoid	<i>Sigmoid activation</i>
------------	---------------------------

Description

Applies $1/(1 + e^{-x})$.

Usage

```
ag_sigmoid(x)
```

Arguments

x `ag_tensor`

Value

ag_tensor

ag_softmax	<i>Softmax activation (column-wise)</i>
------------	---

Description

Applies numerically stable softmax along rows so that each column (one sample) sums to 1.

Usage

ag_softmax(x)

Arguments

x ag_tensor of shape [classes, batch_size]

Value

ag_tensor of the same shape as x

ag_softmax_cross_entropy_loss	<i>Fused softmax + cross-entropy loss (numerically stable)</i>
-------------------------------	--

Description

Combines softmax and CE in one op using the fused gradient $(p - y) / n$. More numerically stable than chaining ag_softmax + ag_cross_entropy_loss. Use this when your last layer outputs raw logits.

Usage

ag_softmax_cross_entropy_loss(logits, target)

Arguments

logits ag_tensor [classes, batch_size] raw (pre-softmax) scores
target matrix [classes, batch_size] one-hot labels

Value

scalar ag_tensor

ag_sub	<i>Element-wise subtraction</i>
--------	---------------------------------

Description

Element-wise subtraction

Usage

ag_sub(A, B)

Arguments

- A ag_tensor or numeric matrix
- B ag_tensor or numeric matrix

Value

ag_tensor

ag_sum	<i>Sum all elements (or along a dim): out = sum(x)</i>
--------	--

Description

Sum all elements (or along a dim): out = sum(x)

Usage

ag_sum(x, dim = NULL, keepdim = FALSE)

Arguments

- x ag_tensor
- dim NULL (all), 1 (row-wise), or 2 (col-wise)
- keepdim Logical: keep size-1 dimensions

Value

scalar (or reduced) ag_tensor

ag_tanh	<i>Tanh activation</i>
Description	
Tanh activation	
Usage	
ag_tanh(x)	
Arguments	
x	ag_tensor
Value	
ag_tensor	
ag_tensor	<i>Create a dynamic tensor (no gradient tracking)</i>

Description	
ag_tensor is backed by an R environment so all references to the same tensor see updates (like PyTorch tensors).	
Usage	
ag_tensor(data, device = .ag_device_state\$device, dtype = .ag_device_state\$dtype)	
Arguments	
data	Numeric matrix or vector
device	"cpu" (default) or "gpu". When "gpu", compute operations will be dispatched to the ggml backend.
dtype	Floating-point precision: "f32" (default), "f16", or "bf16". Ignored on CPU; controls upload precision on GPU.
Value	
An ag_tensor object (environment)	

ag_to_device	<i>Move a tensor to the specified device</i>
--------------	--

Description

Copies an ag_tensor to the target device, returning a new tensor. The original tensor is not modified.

Usage

```
ag_to_device(tensor, device)
```

Arguments

tensor	An ag_tensor
device	"cpu" or "gpu"

Value

A new ag_tensor on the target device (or the original if already on the target device)

ag_train	<i>Switch a layer or sequential model to training mode</i>
----------	--

Description

Switch a layer or sequential model to training mode

Usage

```
ag_train(model)
```

Arguments

model	An ag_sequential, ag_batch_norm, or ag_dropout layer
-------	--

Value

The model/layer (invisibly)

ag_transpose	<i>Transpose a tensor</i>
--------------	---------------------------

Description

Transpose a tensor

Usage

```
ag_transpose(x)
```

Arguments

x	ag_tensor
---	-----------

Value

ag_tensor with rows and columns swapped

as.matrix.ggml_matrix	<i>Extract the underlying matrix from a ggml_matrix</i>
-----------------------	---

Description

Extract the underlying matrix from a ggml_matrix

Usage

```
## S3 method for class 'ggml_matrix'
as.matrix(x, ...)
```

Arguments

x	A ggml_matrix .
...	Ignored; present for S3 as.matrix compatibility.

Value

The plain numeric matrix.

as_gpu_matrix	Wrap a matrix so multiplies run on the GPU
---------------	--

Description

Tags a matrix as GPU-backed. In an expression like `as_gpu_matrix(A) %*% B`, the multiply dispatches to `ggml_matmul` (Vulkan GPU with CPU fallback) instead of R's built-in. Only one operand needs wrapping. `crossprod()` and `tcrossprod()` on a wrapped matrix likewise use the GPU.

Usage

```
as_gpu_matrix(x, device = "auto", prec = "f32")
```

Arguments

x	A numeric matrix (or something coercible with <code>as.matrix</code>).
device	"auto" (default), "gpu" or "cpu"; see <code>ggml_matmul</code> .
prec	"f32" (default) or "f16"; see <code>ggml_matmul</code> .

Value

A `ggml_matrix`.

Examples

```
A <- matrix(rnorm(9), 3)
B <- matrix(rnorm(9), 3)
as_gpu_matrix(A, device = "cpu") %*% B
```

augment.ggmlr_parsnip_model	Augment new data with predictions from a fitted ggml parsnip model
-----------------------------	--

Description

Adds prediction columns to 'new_data', broom style. For classification this appends '`.pred_class`' plus one '`.pred_<level>`' probability column per class; for regression it appends '`.pred`'. Predictions are produced by the existing '`predict()`' method for ggml parsnip models (no duplicate inference logic).

Usage

```
## S3 method for class 'ggmlr_parsnip_model'
augment(x, new_data, ...)
```

Arguments

x	A fitted 'ggmlr_parsnip_model' (the engine object inside a parsnip fit; e.g. from 'extract_fit_engine()').
new_data	A data frame of predictors (same columns used for fitting).
...	Unused; for generic compatibility.

Value

'new_data' as a tibble with prediction columns appended.

Examples

```
ggml_set_n_threads(1L) # deterministic, single OpenMP pool
spec <- parsnip::mlp(hidden_units = 8L, epochs = 3L) |>
  parsnip::set_engine("ggml", backend = "cpu") |>
  parsnip::set_mode("regression")
fit_obj <- parsnip::fit(spec, mpg ~ ., data = mtcars)
generics::augment(parsnip::extract_fit_engine(fit_obj), mtcars)
```

backward	<i>Run backward pass from a scalar loss tensor</i>
----------	--

Description

Traverses the gradient tape in reverse and accumulates gradients into tensor\$grad for all leaf tensors with requires_grad = TRUE.

Usage

```
backward(loss)
```

Arguments

loss	Scalar ag_tensor
------	------------------

Value

Named environment: tensor id -> gradient matrix (for use by optimizer\$step)

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
x <- ag_tensor(matrix(c(1, 2), 2, 1))
y <- ag_tensor(matrix(c(0, 1), 2, 1))
with_grad_tape({
  out <- ag_matmul(w, x)
  loss <- ag_mse_loss(out, y)
})
grads <- backward(loss)
```

clip_grad_norm

*Clip gradients by global L2 norm***Description**

Rescales all gradients in grads so that their global L2 norm does not exceed max_norm. Modifies the grads environment in-place and returns the pre-clip norm.

Usage

```
clip_grad_norm(params, grads, max_norm)
```

Arguments

params	Named list of ag_param tensors (same as passed to optimizer).
grads	Gradient environment returned by backward().
max_norm	Maximum allowed global L2 norm.

Details

Call this **after** backward() and **before** optimizer\$step().

Value

Numeric: the global L2 norm before clipping (invisibly).

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
x <- ag_tensor(matrix(c(1, 1), 2, 1))
with_grad_tape({
  out <- ag_matmul(w, x)
  loss <- ag_mse_loss(out, matrix(0, 2, 1))
})
grads <- backward(loss)
clip_grad_norm(list(w = w), grads, max_norm = 1.0)
```

```
compile.ggml_sequential_model
```

Compile a Model

Description

Configures the model for training by setting the optimizer, loss function, and metrics. This is the keras-compatible interface; it delegates to [ggml_compile](#).

Usage

```
## S3 method for class 'ggml_sequential_model'
compile(
  object,
  optimizer = "adam",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"),
  ...
)

## S3 method for class 'ggml_functional_model'
compile(
  object,
  optimizer = "adam",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"),
  ...
)
```

Arguments

<code>object</code>	A model object (e.g. <code>ggml_sequential_model</code> or <code>ggml_functional_model</code>).
<code>optimizer</code>	Character: "adam", "adamw", or "sgd".
<code>loss</code>	Character: "categorical_crossentropy" or "mse".
<code>metrics</code>	Character vector of metrics (default "accuracy").
<code>...</code>	Additional arguments passed to ggml_compile .

Value

The compiled model (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(10, activation = "softmax", input_shape = 4)
model <- compile(model, optimizer = "adam",
```

```
loss = "categorical_crossentropy")
```

```
dequantize_row_iq2_xxs
```

Dequantize Row (IQ)

Description

Converts IQ (integer quantization) data back to float values. IQ formats provide high compression with importance-matrix-aware quantization.

Usage

```
dequantize_row_iq2_xxs(raw_data, n_elements)
```

```
dequantize_row_iq2_xs(raw_data, n_elements)
```

```
dequantize_row_iq2_s(raw_data, n_elements)
```

```
dequantize_row_iq3_xxs(raw_data, n_elements)
```

```
dequantize_row_iq3_s(raw_data, n_elements)
```

```
dequantize_row_iq4_nl(raw_data, n_elements)
```

```
dequantize_row_iq4_xs(raw_data, n_elements)
```

```
dequantize_row_iq1_s(raw_data, n_elements)
```

```
dequantize_row_iq1_m(raw_data, n_elements)
```

Arguments

<code>raw_data</code>	Raw vector containing quantized data
<code>n_elements</code>	Number of elements to dequantize

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_mxfp4 *Dequantize Row (MXFP4)*

Description

Converts MXFP4 (microscaling FP4) quantized data back to float values.

Usage

```
dequantize_row_mxfp4(raw_data, n_elements)
```

Arguments

raw_data	Raw vector containing quantized data
n_elements	Number of elements to dequantize

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_nvfp4 *Dequantize NVFP4 Data*

Description

Converts NVFP4 quantized data back to float values.

Usage

```
dequantize_row_nvfp4(raw_data, n_elements)
```

Arguments

raw_data	Raw vector of NVFP4 quantized data
n_elements	Number of dequantized elements (must be multiple of 64)

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_q1_0	<i>Dequantize Q1_0 Data</i>
---------------------	-----------------------------

Description

Converts Q1_0 quantized data back to float values.

Usage

```
dequantize_row_q1_0(raw_data, n_elements)
```

Arguments

- raw_data Raw vector of Q1_0 quantized data
- n_elements Number of dequantized elements (must be multiple of 128)

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_q2_K	<i>Dequantize Row (K-quants)</i>
---------------------	----------------------------------

Description

Converts K-quant quantized data back to float values. K-quants (q2_K through q8_K) provide better quality/size tradeoffs.

Usage

```
dequantize_row_q2_K(raw_data, n_elements)
```

```
dequantize_row_q3_K(raw_data, n_elements)
```

```
dequantize_row_q4_K(raw_data, n_elements)
```

```
dequantize_row_q5_K(raw_data, n_elements)
```

```
dequantize_row_q6_K(raw_data, n_elements)
```

```
dequantize_row_q8_K(raw_data, n_elements)
```

Arguments

raw_data	Raw vector containing quantized data
n_elements	Number of elements to dequantize

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_q4_0	<i>Dequantize Row (Q4_0)</i>
---------------------	------------------------------

Description

Converts Q4_0 quantized data back to float values.

Usage

```
dequantize_row_q4_0(raw_data, n_elements)

dequantize_row_q4_1(raw_data, n_elements)

dequantize_row_q5_0(raw_data, n_elements)

dequantize_row_q5_1(raw_data, n_elements)

dequantize_row_q8_0(raw_data, n_elements)
```

Arguments

raw_data	Raw vector containing quantized data
n_elements	Number of elements to dequantize

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dequantize_row_tq1_0	<i>Dequantize Row (Ternary)</i>
----------------------	---------------------------------

Description

Converts ternary quantized data back to float values. TQ1_0 and TQ2_0 are extreme compression formats.

Usage

```
dequantize_row_tq1_0(raw_data, n_elements)
```

```
dequantize_row_tq2_0(raw_data, n_elements)
```

Arguments

raw_data	Raw vector containing quantized data
n_elements	Number of elements to dequantize

Value

Numeric vector of dequantized values

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

dp_train

Data-parallel training across multiple GPUs

Description

Runs synchronous data-parallel training:

1. `make_model()` is called `n_gpu` times to create one independent model replica per GPU (each with its own parameters).
2. Each iteration: the current data item is forwarded through every replica in parallel; gradients are computed via `backward()`.
3. Gradients are averaged across all replicas (element-wise mean).
4. One optimizer step is taken on replica 0; updated weights are then broadcast to replicas 1 ... N-1 so all replicas stay in sync.

Usage

```
dp_train(
  make_model,
  data,
  loss_fn = NULL,
  forward_fn = NULL,
  target_fn = NULL,
```

```

    n_gpu = NULL,
    n_iter = 10L,
    lr = 0.001,
    max_norm = Inf,
    verbose = 10L
  )

```

Arguments

<code>make_model</code>	A zero-argument function that returns a model object with at least <code>\$forward(x)</code> and <code>\$parameters()</code> methods. Called <code>n_gpu</code> times; each call must produce independent parameters.
<code>data</code>	A list of training samples. Each element is passed directly to <code>forward_fn</code> (or to <code>model\$forward()</code> if <code>forward_fn</code> is <code>NULL</code>).
<code>loss_fn</code>	A function <code>(logits, target) -> scalar ag_tensor</code> . If <code>NULL</code> , <code>forward_fn</code> must return the loss directly.
<code>forward_fn</code>	Optional function <code>(model, sample) -> logits</code> . If <code>NULL</code> , the sample is passed directly as <code>model\$forward(sample)</code> .
<code>target_fn</code>	Optional function <code>(sample) -> target</code> . Used when <code>loss_fn</code> is not <code>NULL</code> to extract the target from a sample. If <code>NULL</code> , sample itself is used as the target.
<code>n_gpu</code>	Number of GPU replicas (default: all available Vulkan devices, minimum 1).
<code>n_iter</code>	Number of training iterations (passes over data).
<code>lr</code>	Learning rate for Adam optimizer (default 1e-3).
<code>max_norm</code>	Gradient clipping threshold (default <code>Inf</code> = no clip).
<code>verbose</code>	Print loss every verbose iterations, or <code>FALSE</code> to suppress output.

Details

Because all replicas live in the same R process and `ag_param` uses environment (reference) semantics, no IPC or NCCL is required — weight synchronisation is a simple in-place copy.

Value

A list with:

`params` Named list of final parameters (from replica 0).
`loss_history` Numeric vector of per-iteration mean loss.
`model` Replica 0 model object.

Examples

```

make_model <- function() {
  W <- ag_param(matrix(rnorm(4), 2, 2))
  list(
    forward = function(x) ag_matmul(W, x),
    parameters = function() list(W = W)
  )
}

```



```

}
data <- lapply(1:8, function(i) matrix(rnorm(2), 2, 1))
result <- dp_train(
  make_model = make_model,
  data       = data,
  loss_fn    = function(out, tgt) ag_mse_loss(out, tgt),
  target_fn  = function(s) s,
  n_gpu      = 1L,
  n_iter     = 10L,
  lr         = 1e-3,
  verbose    = FALSE
)

```

```

evaluate.ggml_sequential_model
      Evaluate a Model

```

Description

Computes loss and metrics on test data. This is the keras-compatible interface; it delegates to [ggml_evaluate](#).

Usage

```

## S3 method for class 'ggml_sequential_model'
evaluate(x, test_x, test_y, batch_size = 32L, ...)

## S3 method for class 'ggml_functional_model'
evaluate(x, test_x, test_y, batch_size = 32L, ...)

```

Arguments

<code>x</code>	A trained model object.
<code>test_x</code>	Test data.
<code>test_y</code>	Test labels.
<code>batch_size</code>	Batch size (default 32).
<code>...</code>	Additional arguments passed to ggml_evaluate .

Value

A named list with loss and metric values.

`fit.ggml_sequential_model`*Train a Model*

Description

Trains the model on data for a fixed number of epochs. This is the keras-compatible interface; it delegates to [ggml_fit](#).

Usage

```
## S3 method for class 'ggml_sequential_model'
fit(
  object,
  x,
  y,
  epochs = 1L,
  batch_size = 32L,
  validation_split = 0,
  validation_data = NULL,
  verbose = 1L,
  callbacks = list(),
  ...
)

## S3 method for class 'ggml_functional_model'
fit(
  object,
  x,
  y,
  epochs = 1L,
  batch_size = 32L,
  validation_split = 0,
  validation_data = NULL,
  verbose = 1L,
  callbacks = list(),
  ...
)
```

Arguments

<code>object</code>	A compiled model object.
<code>x</code>	Training data. Matrix, array, or list of matrices (multi-input).
<code>y</code>	Training labels (matrix, one-hot encoded for classification).
<code>epochs</code>	Number of training epochs (default 1).
<code>batch_size</code>	Batch size (default 32).

`validation_split` Fraction of data for validation (default 0).
`validation_data` Optional list(`x_val`, `y_val`).
`verbose` 0 = silent, 1 = progress (default 1).
`callbacks` List of callback objects (default `list()`).
`...` Additional arguments passed to `ggml_fit`.

Value

The trained model (invisibly), with `model$history`.

Examples

```

model <- ggml_model_sequential() |>
  ggml_layer_dense(10, activation = "softmax", input_shape = 4)
model <- compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")
# model <- fit(model, x_train, y_train, epochs = 5, batch_size = 32)
  
```

`ggml_abort_is_r_enabled`

Check if R Abort Handler is Enabled

Description

Check if R Abort Handler is Enabled

Usage

```
ggml_abort_is_r_enabled()
```

Value

Logical indicating if R-compatible abort handling is active

See Also

Other logging: `ggml_log_is_r_enabled()`, `ggml_log_set_default()`, `ggml_log_set_r()`, `ggml_set_abort_callback`, `ggml_set_abort_callback_r()`

ggml_abs	<i>Absolute Value (Graph)</i>
----------	-------------------------------

Description

Creates a graph node for element-wise absolute value: $|x|$

Usage

```
ggml_abs(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the abs operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(-2, -1, 1, 2))
result <- ggml_abs(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [2, 1, 1, 2]
ggml_free(ctx)
```

ggml_abs_inplace	<i>Absolute Value In-place (Graph)</i>
------------------	--

Description

Creates a graph node for in-place element-wise absolute value.

Usage

```
ggml_abs_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with absolute values

ggml_add	<i>Add tensors</i>
----------	--------------------

Description

Creates a graph node for element-wise addition. Must be computed using ggml_build_forward_expand() and ggml_graph_compute().

Usage

```
ggml_add(ctx, a, b)
```

```
ggml_add(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor
b	Second tensor (same shape as a)

Value

Tensor representing the addition operation

Tensor representing the addition operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(5, 4, 3, 2, 1))
result <- ggml_add(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(5, 4, 3, 2, 1))
result <- ggml_add(ctx, a, b)
```

```
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_add1	Add Scalar to Tensor (Graph)
-----------	------------------------------

Description

Creates a graph node for adding a scalar (1-element tensor) to all elements of a tensor. This is more efficient than creating a full tensor of the same value.

Usage

```
ggml_add1(ctx, a, b)
```

Arguments

ctx	GGML context
a	Input tensor
b	Scalar tensor (1-element tensor)

Value

Tensor representing the operation $a + b$ (broadcasted)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
scalar <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 1)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(scalar, 10)
result <- ggml_add1(ctx, a, scalar)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_add_inplace	<i>Element-wise Addition In-place (Graph)</i>
------------------	---

Description

Creates a graph node for in-place element-wise addition. Result is stored in tensor a, saving memory allocation. Returns a view of the modified tensor.

Usage

```
ggml_add_inplace(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor (will be modified in-place)
b	Second tensor (same shape as a)

Value

View of tensor a with the addition result

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(5, 4, 3, 2, 1))
result <- ggml_add_inplace(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_add_rel_pos	<i>Add Relative Position Bias (Graph)</i>
------------------	---

Description

Adds width and height relative-position bias to a.

Usage

```
ggml_add_rel_pos(ctx, a, pw, ph)
```

Arguments

ctx	GGML context
a	Input tensor
pw	Width relative-position tensor
ph	Height relative-position tensor

Value

Tensor with added relative-position bias

ggml_apply	<i>Apply a Layer Object to a Tensor Node</i>
------------	--

Description

Applies a ggml_layer object (created with ggml_dense(), ggml_lstm(), etc.) to a ggml_tensor_node. Applying the *same* layer object to multiple tensor nodes produces shared weights – the identity of the layer object (layer\$layer_id) is used as the sharing key, not its name.

Usage

```
ggml_apply(tensor, layer)
```

Arguments

tensor	A ggml_tensor_node (e.g. from ggml_input()).
layer	A ggml_layer object.

Value

A new ggml_tensor_node.

Examples

```
encoder <- ggml_dense(64L, activation = "relu")
x1 <- ggml_input(shape = 32L)
x2 <- ggml_input(shape = 32L)
out1 <- x1 |> ggml_apply(encoder)
out2 <- x2 |> ggml_apply(encoder) # shared weights
model <- ggml_model(inputs = list(x1, x2),
                    outputs = list(out1, out2))
```

ggml_arange	<i>Arange (Graph)</i>
-------------	-----------------------

Description

Creates a 1D F32 tensor with values from start (inclusive) to stop (exclusive) in steps of step.

Usage

```
ggml_arange(ctx, start, stop, step = 1)
```

Arguments

ctx	GGML context
start	Start value (inclusive)
stop	Stop value (exclusive)
step	Step between values (default 1)

Value

1D F32 tensor

ggml_are_same_layout	<i>Check if Two Tensors Have the Same Layout</i>
----------------------	--

Description

Compares two tensors to check if they have identical type, shape, and strides. Tensors with the same layout can be used interchangeably for memory operations.

Usage

```
ggml_are_same_layout(a, b)
```

Arguments

a	External pointer to first tensor
b	External pointer to second tensor

Value

Logical indicating if tensors have identical layout

See Also

Other tensor: [ggml_get_op_params\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_f32\(\)](#), [ggml_set_op_params_i32\(\)](#)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 4)
b <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 4)
same <- ggml_are_same_layout(a, b) # TRUE
ggml_free(ctx)
```

`ggml_are_same_shape` *Compare Tensor Shapes*

Description

Checks if two tensors have the same shape.

Usage

```
ggml_are_same_shape(a, b)
```

Arguments

a	First tensor
b	Second tensor

Value

TRUE if shapes are identical, FALSE otherwise

`ggml_are_same_stride` *Compare Tensor Strides*

Description

Check if two tensors have the same stride pattern. Useful for determining if tensors can share operations.

Usage

```
ggml_are_same_stride(a, b)
```

Arguments

- a First tensor
- b Second tensor

Value

Logical indicating if strides are identical

See Also

Other tensor_layout: [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_argmax	<i>Argmax (Graph)</i>
-------------	-----------------------

Description

Creates a graph node that finds the index of the maximum value. CRITICAL for token generation in LLMs.

Usage

```
ggml_argmax(ctx, a)
```

Arguments

- ctx GGML context
- a Input tensor

Value

Tensor with argmax indices

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 5, 3, 2, 4))
result <- ggml_argmax(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_i32(result) # 1 (0-indexed)
ggml_free(ctx)
```

ggml_argsort

*Argsort - Get Sorting Indices (Graph)***Description**

Returns indices that would sort the tensor rows. Each row is sorted independently.

Usage

```
ggml_argsort(ctx, a, order = GGML_SORT_ORDER_ASC)
```

Arguments

ctx	GGML context
a	Input tensor to sort (F32)
order	Sort order: GGML_SORT_ORDER_ASC (0) or GGML_SORT_ORDER_DESC (1)

Value

Tensor of I32 indices that would sort each row

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Create tensor with values to sort
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(3, 1, 4, 1, 5))
# Get indices for ascending sort
indices <- ggml_argsort(ctx, a, GGML_SORT_ORDER_ASC)
graph <- ggml_build_forward_expand(ctx, indices)
ggml_graph_compute(ctx, graph)
result <- ggml_get_i32(indices)
# result: [1, 3, 0, 2, 4] (0-indexed positions for sorted order)
ggml_free(ctx)
```

ggml_backend_alloc_ctx_tensors

*Allocate Context Tensors to Backend***Description**

Allocates all tensors in a GGML context to a specific backend. Returns a buffer that must be freed when no longer needed.

Usage

```
ggml_backend_alloc_ctx_tensors(ctx, backend)
```

Arguments

ctx	GGML context
backend	Backend handle

Value

Backend buffer object

ggml_backend_buffer_clear
<i>Clear buffer memory</i>

Description

Clear buffer memory

Usage

```
ggml_backend_buffer_clear(buffer, value = 0L)
```

Arguments

buffer	External pointer to buffer
value	Byte value to fill with (default 0)

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host\(\)](#), [ggml_backend_buffer_is_mu](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_bufi\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#),

```
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),  
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_u  
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),  
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),  
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),  
ggml_backend_tensor_set_async(), ggml_backend_unload()
```

ggml_backend_buffer_free
<i>Free Backend Buffer</i>

Description

Frees a backend buffer and all associated memory.

Usage

```
ggml_backend_buffer_free(buffer)
```

Arguments

buffer	Backend buffer object
--------	-----------------------

Value

No return value, called for side effects

ggml_backend_buffer_get_size
<i>Get Backend Buffer Size</i>

Description

Returns the total size of a backend buffer.

Usage

```
ggml_backend_buffer_get_size(buffer)
```

Arguments

buffer	Backend buffer object
--------	-----------------------

Value

Size in bytes

ggml_backend_buffer_get_usage
Get buffer usage

Description

Get buffer usage

Usage

```
ggml_backend_buffer_get_usage(buffer)
```

Arguments

buffer	External pointer to buffer
--------	----------------------------

Value

Usage constant

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_is_host\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_buffer_is_host

Check if buffer is host memory

Description

Check if buffer is host memory

Usage

```
ggml_backend_buffer_is_host(buffer)
```

Arguments

buffer	External pointer to buffer
--------	----------------------------

Value

Logical indicating if buffer is in host memory

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_buffer_is_multi_buffer

Check if buffer is a multi-buffer

Description

Check if buffer is a multi-buffer

Usage

```
ggml_backend_buffer_is_multi_buffer(buffer)
```

Arguments

buffer	External pointer to buffer
--------	----------------------------

Value

Logical indicating if buffer is a multi-buffer

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_accessible\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_buffer_name
<i>Get Backend Buffer Name</i>

Description

Returns the name/type of a backend buffer.

Usage

ggml_backend_buffer_name(buffer)

Arguments

buffer	Backend buffer object
--------	-----------------------

Value

Character string with buffer name

ggml_backend_buffer_reset
<i>Reset buffer</i>

Description

Reset buffer

Usage

ggml_backend_buffer_reset(buffer)

Arguments

buffer	External pointer to buffer
--------	----------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_bufi\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_buffer_set_usage
<i>Set buffer usage hint</i>

Description

Set buffer usage hint

Usage

`ggml_backend_buffer_set_usage(buffer, usage)`

Arguments

buffer	External pointer to buffer
usage	Usage constant (use <code>ggml_backend_buffer_usage_*</code> functions)

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#),

```

ggml_backend_dev_get(), ggml_backend_dev_get_props(), ggml_backend_dev_init(), ggml_backend_dev_memory
ggml_backend_dev_name(), ggml_backend_dev_offload_op(), ggml_backend_dev_supports_buft(),
ggml_backend_dev_supports_op(), ggml_backend_dev_type(), ggml_backend_device_register(),
ggml_backend_device_type_accel(), ggml_backend_device_type_cpu(), ggml_backend_device_type_gpu(),
ggml_backend_device_type_igpu(), ggml_backend_event_free(), ggml_backend_event_new(),
ggml_backend_event_record(), ggml_backend_event_synchronize(), ggml_backend_event_wait(),
ggml_backend_get_device(), ggml_backend_graph_compute_async(), ggml_backend_graph_plan_compute(),
ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(), ggml_backend_init_best(),
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_u
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()

```

ggml_backend_buffer_usage_any

Buffer usage: Any

Description

Buffer usage: Any

Usage

ggml_backend_buffer_usage_any()

Value

Integer constant for any buffer usage

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weights\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_u](#)

```
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()
```

ggml_backend_buffer_usage_compute

Buffer usage: Compute

Description

Buffer usage: Compute

Usage

```
ggml_backend_buffer_usage_compute()
```

Value

Integer constant for compute buffer usage

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_weights()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory_size()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_op()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

ggml_backend_buffer_usage_weights

Buffer usage: Weights

Description

Buffer usage: Weights

Usage

ggml_backend_buffer_usage_weights()

Value

Integer constant for weights buffer usage

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_cpu_init *Initialize CPU Backend*

Description

Creates a new CPU backend instance for graph computation.

Usage

```
ggml_backend_cpu_init()
```

Value

Backend pointer

```
ggml_backend_cpu_set_n_threads
```

Set CPU Backend Threads

Description

Sets the number of threads for CPU backend computation.

Usage

```
ggml_backend_cpu_set_n_threads(backend, n_threads)
```

Arguments

backend	CPU backend pointer
n_threads	Number of threads

Value

NULL invisibly

```
ggml_backend_device_register
```

Register a device

Description

Dynamically registers a new device in the global registry. This is an advanced function for custom backend development.

Usage

```
ggml_backend_device_register(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_device_type_accel

Device type: Accelerator

Description

Device type: Accelerator

Usage

`ggml_backend_device_type_accel()`

Value

Integer constant for accelerator device type (e.g. BLAS, AMX)

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#),


```

ggml_backend_dev_init(), ggml_backend_dev_memory(), ggml_backend_dev_name(), ggml_backend_dev_offload_
ggml_backend_dev_supports_bufi(), ggml_backend_dev_supports_op(), ggml_backend_dev_type(),
ggml_backend_device_register(), ggml_backend_device_type_cpu(), ggml_backend_device_type_gpu(),
ggml_backend_device_type_igpu(), ggml_backend_event_free(), ggml_backend_event_new(),
ggml_backend_event_record(), ggml_backend_event_synchronize(), ggml_backend_event_wait(),
ggml_backend_get_device(), ggml_backend_graph_compute_async(), ggml_backend_graph_plan_compute(),
ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(), ggml_backend_init_best(),
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_us
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()

```

ggml_backend_device_type_cpu

Device type: CPU

Description

Device type: CPU

Usage

```
ggml_backend_device_type_cpu()
```

Value

Integer constant for CPU device type

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host`
`ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`,
`ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight`
`ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`,
`ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`,
`ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_`
`ggml_backend_dev_supports_bufi()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`,
`ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_gpu()`,
`ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`,
`ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`,
`ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`,
`ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`,
`ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`,
`ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_us`
`ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`,

```
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()
```

```
ggml_backend_device_type_gpu
    Device type: GPU
```

Description

Device type: GPU

Usage

```
ggml_backend_device_type_gpu()
```

Value

Integer constant for GPU device type

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_cpu()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

ggml_backend_device_type_igpu

Device type: Integrated GPU

Description

Device type: Integrated GPU

Usage

ggml_backend_device_type_igpu()

Value

Integer constant for integrated GPU device type

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_by_name

Get device by name

Description

Get device by name

Usage

```
ggml_backend_dev_by_name(name)
```

Arguments

name	Device name
------	-------------

Value

External pointer to device, or NULL if not found

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_accessible\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_bufi\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

`ggml_backend_dev_by_type`

Get device by type

Description

Get device by type

Usage

```
ggml_backend_dev_by_type(type)
```

Arguments

type	Device type (use <code>ggml_backend_device_type_*</code> functions)
------	---

Value

External pointer to first device of given type, or NULL if not found

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weighted\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_count

Get number of available devices

Description

Get number of available devices

Usage

```
ggml_backend_dev_count()
```

Value

Number of devices

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weighted\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_size\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

```

ggml_backend_dev_name(), ggml_backend_dev_offload_op(), ggml_backend_dev_supports_buft(),
ggml_backend_dev_supports_op(), ggml_backend_dev_type(), ggml_backend_device_register(),
ggml_backend_device_type_accel(), ggml_backend_device_type_cpu(), ggml_backend_device_type_gpu(),
ggml_backend_device_type_igpu(), ggml_backend_event_free(), ggml_backend_event_new(),
ggml_backend_event_record(), ggml_backend_event_synchronize(), ggml_backend_event_wait(),
ggml_backend_get_device(), ggml_backend_graph_compute_async(), ggml_backend_graph_plan_compute(),
ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(), ggml_backend_init_best(),
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_usage(),
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()

```

ggml_backend_dev_description

Get device description

Description

Get device description

Usage

```
ggml_backend_dev_description(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

Device description

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory_size()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_op()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`,

```

ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(), ggml_backend_init_best(),
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_usage(),
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()

```

ggml_backend_dev_get *Get device by index*

Description

Get device by index

Usage

```
ggml_backend_dev_get(index)
```

Arguments

index	Device index (0-based)
-------	------------------------

Value

External pointer to device, or NULL if not found

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_get_props

Get device properties

Description

Get device properties

Usage

```
ggml_backend_dev_get_props(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

List with name, description, memory_free, memory_total, type, device_id, caps

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory_free\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_bufd\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_init *Initialize backend from device*

Description

Initialize backend from device

Usage

```
ggml_backend_dev_init(device, params = NULL)
```

Arguments

device	External pointer to device
params	Optional parameters string

Value

External pointer to backend, or NULL on failure

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_writable\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_memory

Get device memory

Description

Get device memory

Usage

```
ggml_backend_dev_memory(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

Named numeric vector with 'free' and 'total' memory in bytes

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weighted\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supported_ops\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_name	<i>Get device name</i>
-----------------------	------------------------

Description

Get device name

Usage

ggml_backend_dev_name(device)

Arguments

device	External pointer to device
--------	----------------------------

Value

Device name

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_offload_op\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_all\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_offload_op

Check if device should offload operation

Description

Check if device should offload operation

Usage

```
ggml_backend_dev_offload_op(device, op)
```

Arguments

device	External pointer to device
op	External pointer to tensor/operation

Value

Logical indicating if operation should be offloaded

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_supports_bufi

Check if device supports buffer type

Description

Check if device supports buffer type

Usage

```
ggml_backend_dev_supports_bufi(device, bufi)
```

Arguments

device	External pointer to device
bufi	External pointer to buffer type

Value

Logical indicating support

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_supports_op

Check if device supports operation

Description

Check if device supports operation

Usage

```
ggml_backend_dev_supports_op(device, op)
```

Arguments

device	External pointer to device
op	External pointer to tensor/operation

Value

Logical indicating support

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload\(\)](#), [ggml_backend_dev_supports_buf_type\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_dev_type *Get device type*

Description

Get device type

Usage

```
ggml_backend_dev_type(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

Device type constant

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_event_free

Free event

Description

Free event

Usage

```
ggml_backend_event_free(event)
```

Arguments

event	External pointer to event
-------	---------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_event_new

Create new event

Description

Create new event

Usage

```
ggml_backend_event_new(device)
```

Arguments

device	External pointer to device
--------	----------------------------

Value

External pointer to event, or NULL on failure

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_event_record

Record event

Description

Record event

Usage

```
ggml_backend_event_record(event, backend)
```

Arguments

event	External pointer to event
backend	External pointer to backend

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_event_synchronize
Synchronize event

Description

Synchronize event

Usage

```
ggml_backend_event_synchronize(event)
```

Arguments

event	External pointer to event
-------	---------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_accessible\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capabilities\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_event_wait

Wait for event

Description

Wait for event

Usage

```
ggml_backend_event_wait(backend, event)
```

Arguments

backend	External pointer to backend
event	External pointer to event

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_free	<i>Free Backend</i>
-------------------	---------------------

Description

Releases resources associated with a backend.

Usage

```
ggml_backend_free(backend)
```

Arguments

backend	Backend pointer
---------	-----------------

Value

NULL invisibly

ggml_backend_get_device	<i>Get device from backend</i>
-------------------------	--------------------------------

Description

Get device from backend

Usage

```
ggml_backend_get_device(backend)
```

Arguments

backend	External pointer to backend
---------	-----------------------------

Value

External pointer to device

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weighted_avg()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_device()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

`ggml_backend_graph_compute`

Compute Graph with Backend

Description

Executes computation graph using specified backend.

Usage

```
ggml_backend_graph_compute(backend, graph)
```

Arguments

<code>backend</code>	Backend pointer
<code>graph</code>	Graph pointer

Value

Status code (0 = success)

ggml_backend_graph_compute_async
Compute graph asynchronously

Description

Starts graph computation without blocking. Use `ggml_backend_synchronize()` to wait for completion.

Usage

```
ggml_backend_graph_compute_async(backend, graph)
```

Arguments

backend	External pointer to backend
graph	External pointer to computation graph

Value

Integer status code (0 = success)

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

Examples

```

cpu <- ggml_backend_cpu_init()
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
b <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)
ggml_set_f32(a, rnorm(100))
# Start async computation
status <- ggml_backend_graph_compute_async(cpu, graph)
# Do other work while computation runs...
ggml_backend_synchronize(cpu)
ggml_backend_free(cpu)
ggml_free(ctx)

```

ggml_backend_graph_plan_compute

Execute graph plan

Description

Execute graph plan

Usage

```
ggml_backend_graph_plan_compute(backend, plan)
```

Arguments

backend	External pointer to backend
plan	External pointer to plan

Value

Status code (0 = success)

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#),


```

ggml_backend_event_wait(), ggml_backend_get_device(), ggml_backend_graph_compute_async(),
ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(), ggml_backend_init_best(),
ggml_backend_init_by_name(), ggml_backend_init_by_type(), ggml_backend_load(), ggml_backend_load_all(),
ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(), ggml_backend_multi_buffer_set_u
ggml_backend_reg_by_name(), ggml_backend_reg_count(), ggml_backend_reg_dev_count(),
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()

```

ggml_backend_graph_plan_create

Create graph execution plan

Description

Create graph execution plan

Usage

```
ggml_backend_graph_plan_create(backend, graph)
```

Arguments

backend	External pointer to backend
graph	External pointer to computation graph

Value

External pointer to plan, or NULL on failure

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#),

```
ggml_backend_reg_dev_get(), ggml_backend_reg_get(), ggml_backend_reg_name(), ggml_backend_register(),
ggml_backend_synchronize(), ggml_backend_tensor_copy_async(), ggml_backend_tensor_get_async(),
ggml_backend_tensor_set_async(), ggml_backend_unload()
```

ggml_backend_graph_plan_free

Free graph execution plan

Description

Free graph execution plan

Usage

```
ggml_backend_graph_plan_free(backend, plan)
```

Arguments

backend	External pointer to backend
plan	External pointer to plan

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_factor\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_init_best

Initialize best available backend

Description

Initialize best available backend

Usage

```
ggml_backend_init_best()
```

Value

External pointer to backend (GPU if available, otherwise CPU)

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_init_by_name

Initialize backend by name

Description

Initialize backend by name

Usage

```
ggml_backend_init_by_name(name, params = NULL)
```

Arguments

name	Backend name (e.g. "CPU", "Vulkan")
params	Optional parameters string

Value

External pointer to backend, or NULL on failure

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_init_by_type
<i>Initialize backend by type</i>

Description

Initialize backend by type

Usage

```
ggml_backend_init_by_type(type, params = NULL)
```

Arguments

type	Device type constant
params	Optional parameters string

Value

External pointer to backend, or NULL on failure

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_bufmt\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_buffer\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_load	<i>Load backend from dynamic library</i>
-------------------	--

Description

Load backend from dynamic library

Usage

ggml_backend_load(path)

Arguments

path	Path to dynamic library
------	-------------------------

Value

External pointer to registry, or NULL on failure

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weighted_avg()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_size()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

`ggml_backend_load_all` *Load all available backends*

Description

Load all available backends

Usage

```
ggml_backend_load_all()
```

Value

NULL invisibly

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weighted_avg()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_size()`, `ggml_backend_dev_supports_buft()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`,

```
ggml_backend_event_wait(), ggml_backend_get_device(), ggml_backend_graph_compute_async(),
ggml_backend_graph_plan_compute(), ggml_backend_graph_plan_create(), ggml_backend_graph_plan_free(),
ggml_backend_init_best(), ggml_backend_init_by_name(), ggml_backend_init_by_type(),
ggml_backend_load(), ggml_backend_meta_device(), ggml_backend_multi_buffer_alloc_buffer(),
ggml_backend_multi_buffer_set_usage(), ggml_backend_reg_by_name(), ggml_backend_reg_count(),
ggml_backend_reg_dev_count(), ggml_backend_reg_dev_get(), ggml_backend_reg_get(),
ggml_backend_reg_name(), ggml_backend_register(), ggml_backend_synchronize(), ggml_backend_tensor_copy,
ggml_backend_tensor_get_async(), ggml_backend_tensor_set_async(), ggml_backend_unload()
```

ggml_backend_meta_device

Create a Meta Backend Device

Description

Creates a "meta" device that wraps multiple "simple" backend devices for tensor parallelism. Each tensor is split across the wrapped devices according to the result of `split_fn`, which is called by ggml when weight buffers are allocated.

Usage

```
ggml_backend_meta_device(devs, split_fn, env = environment(split_fn))
```

Arguments

<code>devs</code>	A list of <code>ggml_backend_dev_t</code> external pointers.
<code>split_fn</code>	A function <code>function(tensor_info, n_devs)</code> returning the split state as described above.
<code>env</code>	An environment in which to evaluate <code>split_fn</code> ; defaults to the function's enclosing environment.

Details

The split function is invoked with two arguments:

tensor_info a named list with fields `name` (character), `type` (integer `ggml_type` enum), `ne` (numeric vector of dimensions), `op` (integer `op` enum), `flags` (integer).

n_devs the number of simple devices wrapped by the meta backend.

It must return a named list with:

axis integer; one of 0..3 to split along a tensor axis, 10 for `MIRRORED` (full copy on each device), 11 for `PARTIAL` (each device has a partial sum), or 98/99 for `NONE/UNKNOWN`.

ne integer or numeric vector of length `n_segments * n_devs` giving the per-segment, per-device slice size along the split axis.

n_segments integer; usually 1, larger for fused tensors like QKV.

If `split_fn` errors or returns an unparseable result, the meta backend silently falls back to `MIRRORED` for that tensor and stops calling the callback (sticky error). This is intentional: a misbehaving callback would otherwise spray errors for every tensor in the model.

Note: with a single device this is a degenerate (no-op) configuration — useful for testing but provides no parallelism benefit. The feature is **experimental** and the API may change.

Value

External pointer to the meta `ggml_backend_dev_t`.

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_device()`, `ggml_backend_dev_supports_buf_t()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_multi_buffer_alloc_buffer()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

`ggml_backend_multi_buffer_alloc_buffer`

Allocate multi-buffer

Description

Creates a buffer that combines multiple backend buffers into one. Useful for managing memory across different backends.

Usage

```
ggml_backend_multi_buffer_alloc_buffer(buffers)
```

Arguments

<code>buffers</code>	List of backend buffer external pointers
----------------------	--

Value

External pointer to multi-buffer

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_capable()`, `ggml_backend_dev_supports_buf_type()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

Examples

```
cpu <- ggml_backend_cpu_init()
ctx1 <- ggml_init(1024, no_alloc = TRUE)
ctx2 <- ggml_init(2048, no_alloc = TRUE)
a <- ggml_new_tensor_1d(ctx1, GGML_TYPE_F32, 10)
b <- ggml_new_tensor_1d(ctx2, GGML_TYPE_F32, 20)
buf1 <- ggml_backend_alloc_ctx_tensors(ctx1, cpu)
buf2 <- ggml_backend_alloc_ctx_tensors(ctx2, cpu)
multi <- ggml_backend_multi_buffer_alloc_buffer(list(buf1, buf2))
ggml_backend_buffer_free(multi)
ggml_backend_free(cpu)
ggml_free(ctx1)
ggml_free(ctx2)
```

`ggml_backend_multi_buffer_set_usage`

Set usage for all buffers in a multi-buffer

Description

Set usage for all buffers in a multi-buffer

Usage

```
ggml_backend_multi_buffer_set_usage(buffer, usage)
```

Arguments

buffer	External pointer to multi-buffer
usage	Usage constant (from ggml_backend_buffer_usage_*)

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_bufi64\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_device\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_name	<i>Get Backend Name</i>
-------------------	-------------------------

Description

Returns the name of the backend (e.g., "CPU").

Usage

```
ggml_backend_name(backend)
```

Arguments

backend	Backend pointer
---------	-----------------

Value

Character string name

ggml_backend_register *Register a backend*

Description

Dynamically registers a new backend in the global registry. This is an advanced function for custom backend development.

Usage

```
ggml_backend_register(reg)
```

Arguments

reg	External pointer to backend registry
-----	--------------------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_data\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_data\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_reg_by_name

Get backend registry by name

Description

Get backend registry by name

Usage

```
ggml_backend_reg_by_name(name)
```

Arguments

name	Registry name
------	---------------

Value

External pointer to registry, or NULL if not found

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_buf_type\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_reg_count

Get number of registered backends

Description

Get number of registered backends

Usage

```
ggml_backend_reg_count()
```

Value

Number of registered backends

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_size\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_reg_dev_count

Get number of devices in registry

Description

Get number of devices in registry

Usage

```
ggml_backend_reg_dev_count(reg)
```

Arguments

reg External pointer to registry

Value

Number of devices

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_buf_type\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_reg_dev_get
<i>Get device from registry</i>

Description

Get device from registry

Usage

```
ggml_backend_reg_dev_get(reg, index)
```

Arguments

reg External pointer to registry
index Device index (0-based)

Value

External pointer to device

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_c`, `ggml_backend_dev_supports_bufi`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_al`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

ggml_backend_reg_get	<i>Get backend registry by index</i>
----------------------	--------------------------------------

Description

Get backend registry by index

Usage

`ggml_backend_reg_get(index)`

Arguments

index	Registry index (0-based)
-------	--------------------------

Value

External pointer to registry, or NULL if not found

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weighted_avg()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_capable()`, `ggml_backend_dev_supports_bufi64()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_size()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

`ggml_backend_reg_name` *Get registry name*

Description

Get registry name

Usage

```
ggml_backend_reg_name(reg)
```

Arguments

<code>reg</code>	External pointer to registry
------------------	------------------------------

Value

Registry name

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host_buffer()`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weighted_avg()`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_capable()`, `ggml_backend_dev_supports_bufi64()`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_alloc_size()`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy_async()`, `ggml_backend_tensor_get_async()`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

<code>ggml_backend_dev_supports_bufi()</code> , <code>ggml_backend_dev_supports_op()</code> , <code>ggml_backend_dev_type()</code> , <code>ggml_backend_device_register()</code> , <code>ggml_backend_device_type_accel()</code> , <code>ggml_backend_device_type_cpu()</code> , <code>ggml_backend_device_type_gpu()</code> , <code>ggml_backend_device_type_ipu()</code> , <code>ggml_backend_event_free()</code> , <code>ggml_backend_event_new()</code> , <code>ggml_backend_event_record()</code> , <code>ggml_backend_event_synchronize()</code> , <code>ggml_backend_event_wait()</code> , <code>ggml_backend_get_device()</code> , <code>ggml_backend_graph_compute_async()</code> , <code>ggml_backend_graph_plan_compute()</code> , <code>ggml_backend_graph_plan_create()</code> , <code>ggml_backend_graph_plan_free()</code> , <code>ggml_backend_init_best()</code> , <code>ggml_backend_init_by_name()</code> , <code>ggml_backend_init_by_type()</code> , <code>ggml_backend_load()</code> , <code>ggml_backend_load_all()</code> , <code>ggml_backend_meta_device()</code> , <code>ggml_backend_multi_buffer_al</code> , <code>ggml_backend_multi_buffer_set_usage()</code> , <code>ggml_backend_reg_by_name()</code> , <code>ggml_backend_reg_count()</code> , <code>ggml_backend_reg_dev_count()</code> , <code>ggml_backend_reg_dev_get()</code> , <code>ggml_backend_reg_get()</code> , <code>ggml_backend_register()</code> , <code>ggml_backend_synchronize()</code> , <code>ggml_backend_tensor_copy_async()</code> , <code>ggml_backend_tensor_get_async()</code> , <code>ggml_backend_tensor_set_async()</code> , <code>ggml_backend_unload()</code>

<code>ggml_backend_sched_alloc_graph</code> <i>Allocate graph on scheduler</i>

Description

Allocates memory for a graph across the scheduler’s backends. Must be called before computing the graph.

Usage

```
ggml_backend_sched_alloc_graph(sched, graph)
```

Arguments

<code>sched</code>	Scheduler pointer
<code>graph</code>	Graph pointer

Value

Logical indicating success

<code>ggml_backend_sched_free</code> <i>Free backend scheduler</i>

Description

Releases resources associated with the backend scheduler.

Usage

```
ggml_backend_sched_free(sched)
```

Arguments

 sched Scheduler pointer from ggml_backend_sched_new()

Value

 NULL (invisible)

Examples

```
cpu <- ggml_backend_cpu_init()
sched <- ggml_backend_sched_new(list(cpu))
ggml_backend_sched_free(sched)
ggml_backend_free(cpu)
```

`ggml_backend_sched_get_backend`
Get backend from scheduler

Description

 Returns a specific backend from the scheduler by index.

Usage

```
ggml_backend_sched_get_backend(sched, index = 0L)
```

Arguments

 sched Scheduler pointer

 index Backend index (0-based)

Value

 Backend pointer

`ggml_backend_sched_get_n_backends`*Get number of backends in scheduler*

Description

Returns the number of backends managed by the scheduler.

Usage

```
ggml_backend_sched_get_n_backends(sched)
```

Arguments

<code>sched</code>	Scheduler pointer
--------------------	-------------------

Value

Integer count of backends

`ggml_backend_sched_get_n_copies`*Get number of tensor copies*

Description

Returns the number of tensor copies made in the last computed graph. Copies occur when data needs to be transferred between backends.

Usage

```
ggml_backend_sched_get_n_copies(sched)
```

Arguments

<code>sched</code>	Scheduler pointer
--------------------	-------------------

Value

Integer count of copies

ggml_backend_sched_get_n_splits
<i>Get number of graph splits</i>

Description

Returns the number of splits in the last computed graph. Higher numbers indicate more distribution across backends.

Usage

ggml_backend_sched_get_n_splits(sched)

Arguments

sched	Scheduler pointer
-------	-------------------

Value

Integer count of splits

ggml_backend_sched_get_tensor_backend
<i>Get tensor backend assignment</i>

Description

Returns which backend a tensor is assigned to.

Usage

ggml_backend_sched_get_tensor_backend(sched, tensor)

Arguments

sched	Scheduler pointer
tensor	Tensor pointer

Value

Backend pointer or NULL if not assigned

ggml_backend_sched_graph_compute
Compute graph using scheduler

Description

Computes a graph by distributing work across multiple backends. This is the main function for multi-GPU computation.

Usage

```
ggml_backend_sched_graph_compute(sched, graph)
```

Arguments

sched	Scheduler pointer
graph	Graph pointer

Value

Status code (0 = success)

Examples

```
# Multi-GPU example
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  gpu1 <- ggml_vulkan_init(0)
  gpu2 <- ggml_vulkan_init(1)
  sched <- ggml_backend_sched_new(list(gpu1, gpu2))

  ctx <- ggml_init(64 * 1024 * 1024)
  a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10000)
  b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10000)
  ggml_set_f32(a, rnorm(10000))
  ggml_set_f32(b, rnorm(10000))

  c <- ggml_add(ctx, a, b)
  graph <- ggml_build_forward_expand(ctx, c)

  # Reserve memory
  ggml_backend_sched_reserve(sched, graph)

  # Compute using both GPUs
  ggml_backend_sched_graph_compute(sched, graph)

  result <- ggml_get_f32(c)

  cat("Splits:", ggml_backend_sched_get_n_splits(sched), "\n")
  cat("Copies:", ggml_backend_sched_get_n_copies(sched), "\n")
}
```

```

    ggml_free(ctx)
    ggml_backend_sched_free(sched)
    ggml_vulkan_free(gpu1)
    ggml_vulkan_free(gpu2)
}

```

```
ggml_backend_sched_graph_compute_async
```

Compute graph asynchronously

Description

Computes a graph asynchronously across backends. Use `ggml_backend_sched_synchronize()` to wait for completion.

Usage

```
ggml_backend_sched_graph_compute_async(sched, graph)
```

Arguments

<code>sched</code>	Scheduler pointer
<code>graph</code>	Graph pointer

Value

Status code (0 = success)

```
ggml_backend_sched_new
```

Create a new backend scheduler

Description

Creates a scheduler that can distribute computation across multiple backends (GPUs, CPU). A CPU backend is automatically added as a fallback. Backends with lower index have higher priority.

Usage

```
ggml_backend_sched_new(backends, parallel = TRUE, graph_size = 2048)
```

Arguments

backends	List of backend pointers (from ggml_vulkan_init() or ggml_backend_cpu_init()). Note: A CPU backend is automatically added, so you only need to specify GPU backends.
parallel	Logical, whether to run backends in parallel (default: TRUE)
graph_size	Expected maximum graph size (default: 2048)

Value

Scheduler pointer

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  # Create two GPU backends (CPU is added automatically)
  gpu1 <- ggml_vulkan_init(0)
  gpu2 <- ggml_vulkan_init(1)

  # Create scheduler with both GPUs + CPU (automatic)
  sched <- ggml_backend_sched_new(list(gpu1, gpu2), parallel = TRUE)

  # The scheduler now has 3 backends: GPU1, GPU2, CPU
  cat("Backends:", ggml_backend_sched_get_n_backends(sched), "\\n")

  # Use scheduler...

  # Cleanup
  ggml_backend_sched_free(sched)
  ggml_vulkan_free(gpu1)
  ggml_vulkan_free(gpu2)
}
```

ggml_backend_sched_reserve

Reserve memory for scheduler

Description

Pre-allocates memory based on a measurement graph. This should be called before using the scheduler to compute graphs.

Usage

```
ggml_backend_sched_reserve(sched, graph)
```

Arguments

sched	Scheduler pointer
graph	Graph pointer to measure memory requirements

Value

Logical indicating success

Examples

```
cpu <- ggml_backend_cpu_init()
sched <- ggml_backend_sched_new(list(cpu))
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 1000)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 1000)
c <- ggml_add(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, c)
ggml_backend_sched_reserve(sched, graph)
ggml_backend_sched_free(sched)
ggml_backend_free(cpu)
ggml_free(ctx)
```

ggml_backend_sched_reset

Reset scheduler

Description

Resets the scheduler, deallocating all tensors. Must be called before changing node backends or allocating a new graph.

Usage

```
ggml_backend_sched_reset(sched)
```

Arguments

sched	Scheduler pointer
-------	-------------------

Value

NULL (invisible)

ggml_backend_sched_set_tensor_backend
<i>Set tensor backend assignment</i>

Description

Manually assigns a specific tensor to run on a specific backend. This overrides automatic scheduling.

Usage

ggml_backend_sched_set_tensor_backend(sched, tensor, backend)

Arguments

sched	Scheduler pointer
tensor	Tensor pointer
backend	Backend pointer to assign tensor to

Value

NULL (invisible)

ggml_backend_sched_synchronize
<i>Synchronize scheduler</i>

Description

Waits for all asynchronous operations to complete.

Usage

ggml_backend_sched_synchronize(sched)

Arguments

sched	Scheduler pointer
-------	-------------------

Value

NULL (invisible)

ggml_backend_synchronize

Synchronize backend

Description

Synchronize backend

Usage

```
ggml_backend_synchronize(backend)
```

Arguments

backend	External pointer to backend
---------	-----------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_accessible\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_buf_type\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_size\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_tensor_copy_async\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_tensor_copy_async

Copy tensor asynchronously between backends

Description

Copy tensor asynchronously between backends

Usage

```
ggml_backend_tensor_copy_async(backend_src, backend_dst, src, dst)
```

Arguments

backend_src	Source backend
backend_dst	Destination backend
src	Source tensor
dst	Destination tensor

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_capable\(\)](#), [ggml_backend_dev_supports_buf_type\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_get_data_ptr\(\)](#), [ggml_backend_tensor_set_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_tensor_get_and_sync
<i>Backend Tensor Get and Sync</i>

Description

Gets tensor data from a backend with synchronization.

Usage

ggml_backend_tensor_get_and_sync(backend, tensor, offset = 0, size)

Arguments

backend	Backend pointer (or NULL for CPU)
tensor	Tensor pointer
offset	Byte offset (default 0)
size	Number of bytes to read

Value

Raw vector with tensor data

ggml_backend_tensor_get_async
<i>Get tensor data asynchronously</i>

Description

Get tensor data asynchronously

Usage

ggml_backend_tensor_get_async(backend, tensor, offset = 0, size)

Arguments

backend	External pointer to backend
tensor	External pointer to tensor
offset	Byte offset (default 0)
size	Number of bytes to read

Value

Numeric vector with data

See Also

Other backend: `ggml_backend_buffer_clear()`, `ggml_backend_buffer_get_usage()`, `ggml_backend_buffer_is_host`, `ggml_backend_buffer_is_multi_buffer()`, `ggml_backend_buffer_reset()`, `ggml_backend_buffer_set_usage()`, `ggml_backend_buffer_usage_any()`, `ggml_backend_buffer_usage_compute()`, `ggml_backend_buffer_usage_weight`, `ggml_backend_dev_by_name()`, `ggml_backend_dev_by_type()`, `ggml_backend_dev_count()`, `ggml_backend_dev_description()`, `ggml_backend_dev_get()`, `ggml_backend_dev_get_props()`, `ggml_backend_dev_init()`, `ggml_backend_dev_memory()`, `ggml_backend_dev_name()`, `ggml_backend_dev_offload_c`, `ggml_backend_dev_supports_bufi`, `ggml_backend_dev_supports_op()`, `ggml_backend_dev_type()`, `ggml_backend_device_register()`, `ggml_backend_device_type_accel()`, `ggml_backend_device_type_cpu()`, `ggml_backend_device_type_gpu()`, `ggml_backend_device_type_igpu()`, `ggml_backend_event_free()`, `ggml_backend_event_new()`, `ggml_backend_event_record()`, `ggml_backend_event_synchronize()`, `ggml_backend_event_wait()`, `ggml_backend_get_device()`, `ggml_backend_graph_compute_async()`, `ggml_backend_graph_plan_compute()`, `ggml_backend_graph_plan_create()`, `ggml_backend_graph_plan_free()`, `ggml_backend_init_best()`, `ggml_backend_init_by_name()`, `ggml_backend_init_by_type()`, `ggml_backend_load()`, `ggml_backend_load_all()`, `ggml_backend_meta_device()`, `ggml_backend_multi_buffer_al`, `ggml_backend_multi_buffer_set_usage()`, `ggml_backend_reg_by_name()`, `ggml_backend_reg_count()`, `ggml_backend_reg_dev_count()`, `ggml_backend_reg_dev_get()`, `ggml_backend_reg_get()`, `ggml_backend_reg_name()`, `ggml_backend_register()`, `ggml_backend_synchronize()`, `ggml_backend_tensor_copy`, `ggml_backend_tensor_set_async()`, `ggml_backend_unload()`

ggml_backend_tensor_get_data
<i>Get Tensor Data via Backend</i>

Description

Gets tensor data using the backend API. This works with tensors allocated on any backend, not just CPU.

Usage

`ggml_backend_tensor_get_data(tensor, offset = 0, n_elements = NULL)`

Arguments

tensor	Tensor pointer
offset	Byte offset (default: 0)
n_elements	Number of elements to retrieve (NULL for all)

Value

R vector with tensor data

ggml_backend_tensor_get_f32_first
<i>Get First Float from Backend Tensor</i>

Description

Reads the first f32 element from a backend tensor.

Usage

ggml_backend_tensor_get_f32_first(tensor)

Arguments

tensor	Tensor pointer
--------	----------------

Value

Float value

ggml_backend_tensor_set_async
<i>Set tensor data asynchronously</i>

Description

Set tensor data asynchronously

Usage

ggml_backend_tensor_set_async(backend, tensor, data, offset = 0, size = NULL)

Arguments

backend	External pointer to backend
tensor	External pointer to tensor
data	Numeric or integer vector
offset	Byte offset (default 0)
size	Number of bytes to copy

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_buffer\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_alloc_size\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_unload\(\)](#)

ggml_backend_tensor_set_data
<i>Set Tensor Data via Backend</i>

Description

Sets tensor data using the backend API. This works with tensors allocated on any backend, not just CPU.

Usage

`ggml_backend_tensor_set_data(tensor, data, offset = 0)`

Arguments

tensor	Tensor pointer
data	R vector with data to set
offset	Byte offset (default: 0)

Value

No return value, called for side effects

ggml_backend_unload	<i>Unload backend</i>
---------------------	-----------------------

Description

Unload backend

Usage

```
ggml_backend_unload(reg)
```

Arguments

reg	External pointer to registry
-----	------------------------------

Value

NULL invisibly

See Also

Other backend: [ggml_backend_buffer_clear\(\)](#), [ggml_backend_buffer_get_usage\(\)](#), [ggml_backend_buffer_is_host_accessible\(\)](#), [ggml_backend_buffer_is_multi_buffer\(\)](#), [ggml_backend_buffer_reset\(\)](#), [ggml_backend_buffer_set_usage\(\)](#), [ggml_backend_buffer_usage_any\(\)](#), [ggml_backend_buffer_usage_compute\(\)](#), [ggml_backend_buffer_usage_weight\(\)](#), [ggml_backend_dev_by_name\(\)](#), [ggml_backend_dev_by_type\(\)](#), [ggml_backend_dev_count\(\)](#), [ggml_backend_dev_description\(\)](#), [ggml_backend_dev_get\(\)](#), [ggml_backend_dev_get_props\(\)](#), [ggml_backend_dev_init\(\)](#), [ggml_backend_dev_memory\(\)](#), [ggml_backend_dev_name\(\)](#), [ggml_backend_dev_offload_device\(\)](#), [ggml_backend_dev_supports_buft\(\)](#), [ggml_backend_dev_supports_op\(\)](#), [ggml_backend_dev_type\(\)](#), [ggml_backend_device_register\(\)](#), [ggml_backend_device_type_accel\(\)](#), [ggml_backend_device_type_cpu\(\)](#), [ggml_backend_device_type_gpu\(\)](#), [ggml_backend_device_type_igpu\(\)](#), [ggml_backend_event_free\(\)](#), [ggml_backend_event_new\(\)](#), [ggml_backend_event_record\(\)](#), [ggml_backend_event_synchronize\(\)](#), [ggml_backend_event_wait\(\)](#), [ggml_backend_get_device\(\)](#), [ggml_backend_graph_compute_async\(\)](#), [ggml_backend_graph_plan_compute\(\)](#), [ggml_backend_graph_plan_create\(\)](#), [ggml_backend_graph_plan_free\(\)](#), [ggml_backend_init_best\(\)](#), [ggml_backend_init_by_name\(\)](#), [ggml_backend_init_by_type\(\)](#), [ggml_backend_load\(\)](#), [ggml_backend_load_all\(\)](#), [ggml_backend_meta_device\(\)](#), [ggml_backend_multi_buffer_all_device\(\)](#), [ggml_backend_multi_buffer_set_usage\(\)](#), [ggml_backend_reg_by_name\(\)](#), [ggml_backend_reg_count\(\)](#), [ggml_backend_reg_dev_count\(\)](#), [ggml_backend_reg_dev_get\(\)](#), [ggml_backend_reg_get\(\)](#), [ggml_backend_reg_name\(\)](#), [ggml_backend_register\(\)](#), [ggml_backend_synchronize\(\)](#), [ggml_backend_tensor_copy\(\)](#), [ggml_backend_tensor_get_async\(\)](#), [ggml_backend_tensor_set_async\(\)](#)

ggml_batch_norm	Create a Batch Normalization Layer Object
-----------------	---

Description

Create a Batch Normalization Layer Object

Usage

```
ggml_batch_norm(eps = 1e-05, name = NULL, trainable = TRUE)
```

Arguments

eps	Small constant for numerical stability (default 1e-5).
name	Optional character name.
trainable	Logical.

Value

A ggml_layer object.

ggml_blk_size	Get Block Size
---------------	----------------

Description

Returns the block size for a GGML type. Quantized types process data in blocks (e.g., 32 elements for Q4_0).

Usage

```
ggml_blk_size(type)
```

Arguments

type	GGML type constant
------	--------------------

Value

Integer block size

See Also

Other type_system: [ggml_ftype_to_ggml_type\(\)](#), [ggml_is_quantized\(\)](#), [ggml_type_name\(\)](#), [ggml_type_sizef\(\)](#)

Examples

```
ggml_blk_size(GGML_TYPE_F32) # 1
ggml_blk_size(GGML_TYPE_Q4_0) # 32
```

```
ggml_build_forward_expand
```

Build forward expand

Description

Builds a computation graph from the output tensor, expanding backwards to include all dependencies.

Creates a computation graph by expanding backwards from the output tensor

Usage

```
ggml_build_forward_expand(ctx, tensor)

ggml_build_forward_expand(ctx, tensor)
```

Arguments

ctx	GGML context
tensor	Output tensor of the computation

Value

Graph pointer
Graph object (external pointer)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(5, 4, 3, 2, 1))
result <- ggml_add(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_set_f32(a, 1:10)
```

```
ggml_set_f32(b, 11:20)
c <- ggml_add(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, c)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(c)
ggml_free(ctx)
```

ggml_callback_early_stopping
Early stopping callback

Description

Stops training when the monitored metric does not improve.

Usage

```
ggml_callback_early_stopping(  
  monitor = "val_loss",  
  patience = 5,  
  min_delta = 0,  
  mode = "auto"  
)
```

Arguments

monitor	Metric to monitor: "val_loss", "val_accuracy", "train_loss", "train_accuracy"
patience	Number of epochs with no improvement before stopping
min_delta	Minimum change to qualify as improvement
mode	"min" (lower is better) or "max" (higher is better). "auto" infers from monitor name.

Value

List with on_epoch_end function

See Also

Other callbacks: [ggml_schedule_cosine_decay\(\)](#), [ggml_schedule_reduce_on_plateau\(\)](#), [ggml_schedule_step_deca](#)

ggml_can_repeat	<i>Check If Tensor Can Be Repeated</i>
-----------------	--

Description

Check if tensor a can be repeated (broadcast) to match tensor b. Used for broadcasting operations.

Usage

```
ggml_can_repeat(a, b)
```

Arguments

- a Source tensor (smaller)
- b Target tensor (larger or same size)

Value

Logical indicating if a can be repeated to match b

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
b <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 8)
ggml_can_repeat(a, b) # TRUE - a can broadcast along dim 1
ggml_free(ctx)
```

ggml_ceil	<i>Ceiling (Graph)</i>
-----------	------------------------

Description

Creates a graph node for element-wise ceiling: ceil(x)

Usage

```
ggml_ceil(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the ceil operation

ggml_ceil_inplace	<i>Ceiling In-place (Graph)</i>
-------------------	---------------------------------

Description

Creates a graph node for in-place element-wise ceiling.

Usage

```
ggml_ceil_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with ceiling values

ggml_clamp	<i>Clamp (Graph)</i>
------------	----------------------

Description

Creates a graph node for clamping values to a range: clamp(x, min, max)

Usage

```
ggml_clamp(ctx, a, min_val, max_val)
```

Arguments

ctx	GGML context
a	Input tensor
min_val	Minimum value
max_val	Maximum value

Value

Tensor with values clamped to [min_val, max_val]

ggml_compile.ggml_functional_model

Compile a Sequential Model

Description

Configures the model for training: infers shapes, creates backend. Weight tensors are created at training time when batch_size is known.

Usage

```
## S3 method for class 'ggml_functional_model'
ggml_compile(
  model,
  optimizer = "adam",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"),
  backend = "auto"
)

ggml_compile(
  model,
  optimizer = "adam",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"),
  backend = "auto"
)

## S3 method for class 'ggml_sequential_model'
ggml_compile(
  model,
  optimizer = "adam",
  loss = "categorical_crossentropy",
  metrics = c("accuracy"),
  backend = "auto"
)
```

Arguments

model	A ggml_sequential_model object
optimizer	Optimizer name: "adam" or "sgd"
loss	Loss function name: "categorical_crossentropy" or "mse"
metrics	Character vector of metrics (currently "accuracy")
backend	Backend to use: "auto" (GPU if available, else CPU), "cpu", or "vulkan"

Value

The compiled model (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_max_pooling_2d(c(2, 2)) |>
  ggml_layer_flatten() |>
  ggml_layer_dense(10, activation = "softmax")
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")
```

ggml_concat	<i>Concatenate Tensors (Graph)</i>
-------------	------------------------------------

Description

Concatenates two tensors along a specified dimension. CRITICAL for KV-cache operations in transformers.

Usage

```
ggml_concat(ctx, a, b, dim = 0)
```

Arguments

ctx	GGML context
a	First tensor
b	Second tensor (must match a in all dimensions except the concat dim)
dim	Dimension along which to concatenate (0-3)

Value

Concatenated tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 3)
b <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 2)
ggml_set_f32(a, rnorm(12))
ggml_set_f32(b, rnorm(8))
# Concatenate along dimension 1: result is 4x5
c <- ggml_concat(ctx, a, b, 1)
graph <- ggml_build_forward_expand(ctx, c)
```

```
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_cont	<i>Make Contiguous (Graph)</i>
-----------	--------------------------------

Description

Makes a tensor contiguous in memory. Required after permute/transpose before some operations.

Usage

```
ggml_cont(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Contiguous tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 3, 4)
ggml_set_f32(a, 1:12)
transposed <- ggml_transpose(ctx, a)
contiguous <- ggml_cont(ctx, transposed)
ggml_free(ctx)
```

ggml_conv_1d	<i>1D Convolution (Graph)</i>
--------------	-------------------------------

Description

Applies 1D convolution to input data.

Usage

```
ggml_conv_1d(ctx, a, b, s0 = 1L, p0 = 0L, d0 = 1L)
```


Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0	Stride (default 1)
p0	Padding (default 0)
d0	Dilation (default 1)

Value

Convolved tensor

ggml_conv_1d_dw	<i>Depthwise 1D Convolution (Graph)</i>
-----------------	---

Description

Applies depthwise 1D convolution: each input channel is convolved with its own kernel.

Usage

ggml_conv_1d_dw(ctx, a, b, s0 = 1L, p0 = 0L, d0 = 1L)

Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0	Stride (default 1)
p0	Padding (default 0)
d0	Dilation (default 1)

Value

Convolved tensor

ggml_conv_2d	<i>2D Convolution (Graph)</i>
--------------	-------------------------------

Description

Applies 2D convolution to input data.

Usage

ggml_conv_2d(ctx, a, b, s0 = 1L, s1 = 1L, p0 = 0L, p1 = 0L, d0 = 1L, d1 = 1L)

Arguments

ctx	GGML context
a	Convolution kernel tensor [KW, KH, IC, OC]
b	Input data tensor [W, H, C, N]
s0	Stride dimension 0 (default 1)
s1	Stride dimension 1 (default 1)
p0	Padding dimension 0 (default 0)
p1	Padding dimension 1 (default 0)
d0	Dilation dimension 0 (default 1)
d1	Dilation dimension 1 (default 1)

Value

Convolved tensor

ggml_conv_2d_direct	<i>Direct 2D Convolution (Graph)</i>
---------------------	--------------------------------------

Description

Applies 2D convolution using the direct algorithm (no im2col).

Usage

```
ggml_conv_2d_direct(  
  ctx,  
  a,  
  b,  
  s0 = 1L,  
  s1 = 1L,  
  p0 = 0L,
```

```
    p1 = 0L,  
    d0 = 1L,  
    d1 = 1L  
)
```

Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0	Stride dimension 0 (default 1)
s1	Stride dimension 1 (default 1)
p0	Padding dimension 0 (default 0)
p1	Padding dimension 1 (default 0)
d0	Dilation dimension 0 (default 1)
d1	Dilation dimension 1 (default 1)

Value

Convolved tensor

ggml_conv_2d_dw	<i>Depthwise 2D Convolution (Graph)</i>
-----------------	---

Description

Applies depthwise 2D convolution: each input channel is convolved with its own kernel. Uses the im2col-based path.

Usage

```
ggml_conv_2d_dw(  
    ctx,  
    a,  
    b,  
    s0 = 1L,  
    s1 = 1L,  
    p0 = 0L,  
    p1 = 0L,  
    d0 = 1L,  
    d1 = 1L  
)
```

Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0, s1	Strides along dim 0 and 1 (default 1)
p0, p1	Padding along dim 0 and 1 (default 0)
d0, d1	Dilation along dim 0 and 1 (default 1)

Value

Convolved tensor

ggml_conv_2d_dw_direct
<i>Depthwise 2D Convolution, direct (Graph)</i>

Description

Direct depthwise 2D convolution without an explicit im2col intermediate.

Usage

```
ggml_conv_2d_dw_direct(  
    ctx,  
    a,  
    b,  
    s0 = 1L,  
    s1 = 1L,  
    p0 = 0L,  
    p1 = 0L,  
    d0 = 1L,  
    d1 = 1L  
)
```

Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0, s1	Strides along dim 0 and 1 (default 1)
p0, p1	Padding along dim 0 and 1 (default 0)
d0, d1	Dilation along dim 0 and 1 (default 1)

Value

Convolved tensor

`ggml_conv_transpose_1d`*Transposed 1D Convolution (Graph)*

Description

Applies transposed 1D convolution (deconvolution) to input data.

Usage

```
ggml_conv_transpose_1d(ctx, a, b, s0 = 1L, p0 = 0L, d0 = 1L)
```

Arguments

<code>ctx</code>	GGML context
<code>a</code>	Convolution kernel tensor
<code>b</code>	Input data tensor
<code>s0</code>	Stride (default 1)
<code>p0</code>	Padding (default 0)
<code>d0</code>	Dilation (default 1)

Value

Transposed convolved tensor

`ggml_conv_transpose_2d_p0`*Transposed 2D Convolution, zero padding (Graph)*

Description

Applies transposed 2D convolution (deconvolution) with zero padding.

Usage

```
ggml_conv_transpose_2d_p0(ctx, a, b, stride = 1L)
```

Arguments

<code>ctx</code>	GGML context
<code>a</code>	Convolution kernel tensor
<code>b</code>	Input data tensor
<code>stride</code>	Stride (default 1)

Value

Transposed convolved tensor

ggml_cos	<i>Cosine (Graph)</i>
----------	-----------------------

Description

Creates a graph node for element-wise cosine: cos(x)

Usage

```
ggml_cos(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the cos operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(0, pi/3, pi/2, pi))
result <- ggml_cos(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [1, 0.5, 0, -1]
ggml_free(ctx)
```

ggml_count_equal	<i>Count Equal Elements (Graph)</i>
------------------	-------------------------------------

Description

Creates a graph node that counts equal elements between two tensors. Useful for accuracy computation.

Usage

```
ggml_count_equal(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor
b	Second tensor (same shape as a)

Value

Tensor containing the count of equal elements

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
pred <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 100)
labels <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 100)
# ... set values ...
correct <- ggml_count_equal(ctx, pred, labels)
graph <- ggml_build_forward_expand(ctx, correct)
ggml_graph_compute(ctx, graph)
# correct now contains count of matching elements
ggml_free(ctx)
```

ggml_cpu_add	<i>Element-wise Addition (CPU Direct)</i>
--------------	---

Description

Performs element-wise addition of two tensors using direct CPU computation. Returns the result as an R numeric vector. Does NOT use computation graphs.

Usage

```
ggml_cpu_add(a, b)
```

Arguments

a	First tensor (must be F32 type)
b	Second tensor (must be F32 type, same size as a)

Value

Numeric vector containing the element-wise sum

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(5, 4, 3, 2, 1))
ggml_cpu_add(a, b)
ggml_free(ctx)
```

ggml_cpu_features	<i>Get All CPU Features</i>
-------------------	-----------------------------

Description

Returns a named list of all CPU feature detection results. Useful for diagnostics and optimizing computation.

Usage

```
ggml_cpu_features()
```

Value

Named list with feature names and logical values

See Also

Other cpu_features: [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

Examples

```
features <- ggml_cpu_features()
print(features)
# On typical x86-64: sse3=TRUE, avx=TRUE, avx2=TRUE, ...
```

ggml_cpu_get_rvv_vlen	<i>Get RISC-V Vector Length</i>
-----------------------	---------------------------------

Description

Returns the RISC-V RVV vector length in bytes (0 if not supported).

Usage

ggml_cpu_get_rvv_vlen()

Value

Integer vector length in bytes

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_get_sve_cnt	<i>Get SVE Vector Length (ARM)</i>
----------------------	------------------------------------

Description

Returns the SVE vector length in bytes (0 if not supported).

Usage

ggml_cpu_get_sve_cnt()

Value

Integer vector length in bytes

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_amx_int8 *CPU Feature Detection - AMX INT8*

Description

Check if the CPU supports AMX INT8 (Advanced Matrix Extensions). AMX provides hardware acceleration for matrix operations on Intel CPUs.

Usage

```
ggml_cpu_has_amx_int8()
```

Value

Logical indicating AMX INT8 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_arm_fma *CPU Feature Detection - ARM FMA*

Description

Check if the CPU supports ARM FMA (Fused Multiply-Add).

Usage

```
ggml_cpu_has_arm_fma()
```

Value

Logical indicating ARM FMA support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx	<i>CPU Feature Detection - AVX</i>
------------------	------------------------------------

Description

Check if the CPU supports AVX instructions.

Usage

`ggml_cpu_has_avx()`

Value

Logical indicating AVX support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx2	CPU Feature Detection - AVX2
-------------------	------------------------------

Description

Check if the CPU supports AVX2 instructions. AVX2 provides 256-bit SIMD operations for faster matrix math.

Usage

```
ggml_cpu_has_avx2()
```

Value

Logical indicating AVX2 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx512	CPU Feature Detection - AVX-512
---------------------	---------------------------------

Description

Check if the CPU supports AVX-512 instructions. AVX-512 provides 512-bit SIMD for maximum throughput.

Usage

```
ggml_cpu_has_avx512()
```

Value

Logical indicating AVX-512 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx512_bf16

CPU Feature Detection - AVX-512 BF16

Description

Check if the CPU supports AVX-512 BF16 (bfloat16) instructions.

Usage

```
ggml_cpu_has_avx512_bf16()
```

Value

Logical indicating AVX-512 BF16 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx512_vbmi

CPU Feature Detection - AVX-512 VBMI

Description

Check if the CPU supports AVX-512 VBMI instructions.

Usage

```
ggml_cpu_has_avx512_vbmi()
```

Value

Logical indicating AVX-512 VBMI support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx512_vnni

CPU Feature Detection - AVX-512 VNNI

Description

Check if the CPU supports AVX-512 VNNI instructions. VNNI accelerates neural network inference with int8/int16 dot products.

Usage

`ggml_cpu_has_avx512_vnni()`

Value

Logical indicating AVX-512 VNNI support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_avx_vnni	CPU Feature Detection - AVX-VNNI
-----------------------	----------------------------------

Description

Check if the CPU supports AVX-VNNI instructions.

Usage

ggml_cpu_has_avx_vnni()

Value

Logical indicating AVX-VNNI support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_bmi2	CPU Feature Detection - BMI2
-------------------	------------------------------

Description

Check if the CPU supports BMI2 (Bit Manipulation Instructions 2).

Usage

ggml_cpu_has_bmi2()

Value

Logical indicating BMI2 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_dotprod	<i>CPU Feature Detection - Dot Product (ARM)</i>
----------------------	--

Description

Check if the CPU supports ARM dot product instructions. Accelerates int8 matrix multiplication common in quantized models.

Usage

```
ggml_cpu_has_dotprod()
```

Value

Logical indicating dot product support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_f16c	<i>CPU Feature Detection - F16C</i>
-------------------	-------------------------------------

Description

Check if the CPU supports F16C instructions for float16 conversion.

Usage

```
ggml_cpu_has_f16c()
```


Value

Logical indicating F16C support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_fma	CPU Feature Detection - FMA
------------------	-----------------------------

Description

Check if the CPU supports FMA (Fused Multiply-Add) instructions. FMA allows matrix operations to run faster by combining operations.

Usage

`ggml_cpu_has_fma()`

Value

Logical indicating FMA support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_fp16_va *CPU Feature Detection - FP16 Vector Arithmetic (ARM)*

Description

Check if the CPU supports ARM half-precision FP16 vector arithmetic.

Usage

```
ggml_cpu_has_fp16_va()
```

Value

Logical indicating FP16 VA support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_llamafile
CPU Feature Detection - Llamafile

Description

Check if llamafile optimizations are available.

Usage

```
ggml_cpu_has_llamafile()
```

Value

Logical indicating llamafile support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_matmul_int8	<i>CPU Feature Detection - INT8 Matrix Multiply (ARM)</i>
--------------------------	---

Description

Check if the CPU supports ARM INT8 matrix multiplication.

Usage

`ggml_cpu_has_matmul_int8()`

Value

Logical indicating INT8 MATMUL support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_neon	<i>CPU Feature Detection - NEON (ARM)</i>
-------------------	---

Description

Check if the CPU supports ARM NEON instructions. NEON is ARM’s SIMD extension for vectorized operations.

Usage

`ggml_cpu_has_neon()`

Value

Logical indicating NEON support

See Also

Other `cpu_features`: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

<code>ggml_cpu_has_riscv_v</code>	<i>CPU Feature Detection - RISC-V Vector</i>
-----------------------------------	--

Description

Check if the CPU supports RISC-V Vector extension.

Usage

`ggml_cpu_has_riscv_v()`

Value

Logical indicating RISC-V V support

See Also

Other `cpu_features`: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_sme	CPU Feature Detection - SME (ARM)
------------------	-----------------------------------

Description

Check if the CPU supports ARM SME (Scalable Matrix Extension).

Usage

ggml_cpu_has_sme()

Value

Logical indicating SME support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_sse3	CPU Feature Detection - SSE3
-------------------	------------------------------

Description

Check if the CPU supports SSE3 instructions.

Usage

ggml_cpu_has_sse3()

Value

Logical indicating SSE3 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

Examples

```
ggml_cpu_has_sse3()
```

ggml_cpu_has_sse3	<i>CPU Feature Detection - SSSE3</i>
-------------------	--------------------------------------

Description

Check if the CPU supports SSSE3 instructions.

Usage

```
ggml_cpu_has_sse3()
```

Value

Logical indicating SSSE3 support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_sve	CPU Feature Detection - SVE (ARM)
------------------	-----------------------------------

Description

Check if the CPU supports ARM SVE (Scalable Vector Extension).

Usage

ggml_cpu_has_sve()

Value

Logical indicating SVE support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

ggml_cpu_has_vsx	CPU Feature Detection - VSX (PowerPC)
------------------	---------------------------------------

Description

Check if the CPU supports PowerPC VSX instructions.

Usage

ggml_cpu_has_vsx()

Value

Logical indicating VSX support

See Also

Other `cpu_features`: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vxe\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

<code>ggml_cpu_has_vxe</code>	<i>CPU Feature Detection - VXE (IBM z/Architecture)</i>
-------------------------------	---

Description

Check if the CPU supports IBM z/Architecture VXE instructions.

Usage

`ggml_cpu_has_vxe()`

Value

Logical indicating VXE support

See Also

Other `cpu_features`: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_wasm_simd\(\)](#)

<code>ggml_cpu_has_wasm_simd</code>	<i>CPU Feature Detection - WebAssembly SIMD</i>
-------------------------------------	---

Description

Check if the CPU/environment supports WebAssembly SIMD.

Usage

`ggml_cpu_has_wasm_simd()`

Value

Logical indicating WASM SIMD support

See Also

Other cpu_features: [ggml_cpu_features\(\)](#), [ggml_cpu_get_rvv_vlen\(\)](#), [ggml_cpu_get_sve_cnt\(\)](#), [ggml_cpu_has_amx_int8\(\)](#), [ggml_cpu_has_arm_fma\(\)](#), [ggml_cpu_has_avx\(\)](#), [ggml_cpu_has_avx2\(\)](#), [ggml_cpu_has_avx512\(\)](#), [ggml_cpu_has_avx512_bf16\(\)](#), [ggml_cpu_has_avx512_vbmi\(\)](#), [ggml_cpu_has_avx512_vnni\(\)](#), [ggml_cpu_has_avx_vnni\(\)](#), [ggml_cpu_has_bmi2\(\)](#), [ggml_cpu_has_dotprod\(\)](#), [ggml_cpu_has_f16c\(\)](#), [ggml_cpu_has_fma\(\)](#), [ggml_cpu_has_fp16_va\(\)](#), [ggml_cpu_has_llamafile\(\)](#), [ggml_cpu_has_matmul_int8\(\)](#), [ggml_cpu_has_neon\(\)](#), [ggml_cpu_has_riscv_v\(\)](#), [ggml_cpu_has_sme\(\)](#), [ggml_cpu_has_sse3\(\)](#), [ggml_cpu_has_ssse3\(\)](#), [ggml_cpu_has_sve\(\)](#), [ggml_cpu_has_vsx\(\)](#), [ggml_cpu_has_vxe\(\)](#)

ggml_cpu_mul	<i>Element-wise Multiplication (CPU Direct)</i>
--------------	---

Description

Performs element-wise multiplication of two tensors using direct CPU computation. Returns the result as an R numeric vector. Does NOT use computation graphs.

Usage

```
ggml_cpu_mul(a, b)
```

Arguments

- a First tensor (must be F32 type)
- b Second tensor (must be F32 type, same size as a)

Value

Numeric vector containing the element-wise product

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(2, 2, 2, 2, 2))
ggml_cpu_mul(a, b)
ggml_free(ctx)
```

ggml_cpy

*Copy Tensor with Type Conversion (Graph)***Description**

Copies tensor a into tensor b, performing type conversion if needed. The tensors must have the same number of elements. CRITICAL for type casting operations (e.g., F32 to F16).

Usage

```
ggml_cpy(ctx, a, b)
```

Arguments

ctx	GGML context
a	Source tensor
b	Destination tensor (defines output type and shape)

Value

Tensor representing the copy operation (returns b with a's data)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Create F32 tensor
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
ggml_set_f32(a, rnorm(100))
# Create F16 tensor for output
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F16, 100)
# Copy with F32 -> F16 conversion
result <- ggml_cpy(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_crossprod

*GPU cross product (drop-in for crossprod)***Description**

Computes $t(A) \%*\% B$ (or $t(A) \%*\% A$ when B is NULL) on the Vulkan GPU, returning an R matrix, with a CPU fallback. See [ggml_matmul](#) for the precision and dispatch notes.

Usage

```
ggml_crossprod(A, B = NULL, device = "auto", prec = "f32")
```

Arguments

A A numeric matrix.

B A numeric matrix with `nrow(B) == nrow(A)`, or `NULL` (default) for `t(A) %*% A`.

device, prec See [ggml_matmul](#).

Value

`t(A) %*% B` as a numeric matrix.

ggml_cycles	<i>Get CPU Cycles</i>
-------------	-----------------------

Description

Returns the current CPU cycle count. Useful for low-level benchmarking.

Usage

```
ggml_cycles()
```

Value

Numeric value representing CPU cycles

Examples

```
ggml_cycles()
```

ggml_cycles_per_ms	<i>Get CPU Cycles per Millisecond</i>
--------------------	---------------------------------------

Description

Returns an estimate of CPU cycles per millisecond. Useful for converting cycle counts to time.

Usage

```
ggml_cycles_per_ms()
```

Value

Numeric value representing cycles per millisecond

Examples

```
ggml_cycles_per_ms()
```

ggml_default_mlp	<i>Default MLP builder for classification and regression</i>
------------------	--

Description

Constructs an uncompiled sequential multi-layer perceptron suitable as a starting point for tabular classification or regression. This is the default model_fn used by LearnerClassifGGML and LearnerRegrGGML when the user does not supply a custom builder, and it is also exported for direct use or as a template for user-defined builders.

Usage

```
ggml_default_mlp(  
  n_features,  
  n_out,  
  task_type = c("classif", "regr"),  
  hidden_layers = c(128L, 64L),  
  activation = "relu",  
  dropout = 0.2  
)
```

Arguments

n_features	Integer. Number of input features. Required.
n_out	Integer. Number of output units. For classification this is the number of classes; for regression this is typically 1.
task_type	Character. One of "classif" or "regr". Controls the final layer's activation.
hidden_layers	Integer vector. Widths of the hidden dense layers. Default c(128L, 64L). Pass integer(0) for a linear model.
activation	Character. Activation applied to each hidden layer. Default "relu". Passed through to ggml_layer_dense .
dropout	Numeric in [0, 1). Dropout rate applied after each hidden layer. Set to 0 to disable dropout. Default 0.2.

Details

The returned model is **not compiled**: the caller is responsible for calling [ggml_compile](#) with the appropriate loss ("categorical_crossentropy" for classification, "mse" for regression) before training.

The final layer is chosen based on task_type:

- "classif" — dense with units = n_out and softmax activation.
- "regr" — dense with units = n_out and no activation (identity / linear output).

Value

An uncompiled ggml_sequential_model object. Call [ggml_compile](#) before [ggml_fit](#).

See Also

[ggml_model_sequential](#), [ggml_layer_dense](#), [ggml_layer_dropout](#), [ggml_compile](#)

Examples

```
## Not run:
# 3-class classifier on 20 features
model <- ggml_default_mlp(
  n_features = 20L,
  n_out      = 3L,
  task_type  = "classif",
  hidden_layers = c(64L, 32L),
  dropout    = 0.1
)
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")

# Single-output regressor
reg <- ggml_default_mlp(
  n_features = 10L,
  n_out      = 1L,
  task_type  = "regr"
)
reg <- ggml_compile(reg, optimizer = "adam", loss = "mse")

## End(Not run)
```

ggml_dense

Create a Dense Layer Object

Description

Returns a reusable layer object for use with [ggml_apply\(\)](#). Applying the same object to multiple tensor nodes shares weights.

Usage

```
ggml_dense(units, activation = NULL, name = NULL, trainable = TRUE)
```

Arguments

- units Number of output units.
- activation Activation function name or NULL.
- name Optional character name.
- trainable Logical; whether weights are updated during training.

Value

A ggml_layer object.

Examples

```
encoder <- ggml_dense(64L, activation = "relu")
x1 <- ggml_input(shape = 32L)
x2 <- ggml_input(shape = 32L)
out1 <- x1 |> ggml_apply(encoder)
out2 <- x2 |> ggml_apply(encoder) # shared weights
```

ggml_diag	<i>Diagonal Matrix (Graph)</i>
-----------	--------------------------------

Description

Creates a diagonal matrix from a vector. For vector $a[n]$, produces matrix with a on the diagonal.

Usage

```
ggml_diag(ctx, a)
```

Arguments

- ctx GGML context
- a Input vector tensor

Value

Diagonal matrix tensor

Examples

```

ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 3)
ggml_set_f32(a, c(1, 2, 3))
d <- ggml_diag(ctx, a) # 3x3 diagonal matrix
graph <- ggml_build_forward_expand(ctx, d)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)

```

ggml_diag_mask_inf	<i>Diagonal Mask with -Inf (Graph)</i>
--------------------	--

Description

Creates a graph node that sets elements above the diagonal to -Inf. This is used for causal (autoregressive) attention masking.

Usage

```
ggml_diag_mask_inf(ctx, a, n_past)
```

Arguments

ctx	GGML context
a	Input tensor (typically attention scores)
n_past	Number of past tokens (shifts the diagonal). Use 0 for standard causal masking where position i can only attend to positions $\leq i$.

Details

In causal attention, we want each position to only attend to itself and previous positions. Setting future positions to -Inf ensures that after softmax, they contribute 0 attention weight.

The `n_past` parameter allows for KV-cache scenarios where the diagonal needs to be shifted to account for previously processed tokens.

Value

Tensor with same shape as input, elements above diagonal set to -Inf

Examples

```

ctx <- ggml_init(16 * 1024 * 1024)
# Create attention scores matrix
scores <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 4)
ggml_set_f32(scores, rep(1, 16))
# Apply causal mask

```

```
masked <- ggml_diag_mask_inf(ctx, scores, 0)
graph <- ggml_build_forward_expand(ctx, masked)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_diag_mask_inf_inplace	<i>Diagonal Mask with -Inf In-place (Graph)</i>
----------------------------	---

Description

In-place version of ggml_diag_mask_inf. Returns a view of the input tensor.

Usage

```
ggml_diag_mask_inf_inplace(ctx, a, n_past)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
n_past	Number of past tokens

Value

View of input tensor with elements above diagonal set to -Inf

ggml_diag_mask_zero	<i>Diagonal Mask with Zero (Graph)</i>
---------------------	--

Description

Creates a graph node that sets elements above the diagonal to 0. Alternative to -Inf masking for certain use cases.

Usage

```
ggml_diag_mask_zero(ctx, a, n_past)
```

Arguments

ctx	GGML context
a	Input tensor
n_past	Number of past tokens

Value

Tensor with same shape as input, elements above diagonal set to 0

ggml_div	<i>Element-wise Division (Graph)</i>
----------	--------------------------------------

Description

Creates a graph node for element-wise division.

Usage

```
ggml_div(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor (numerator)
b	Second tensor (denominator, same shape as a)

Value

Tensor representing the division operation (a / b)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(10, 20, 30, 40, 50))
ggml_set_f32(b, c(2, 2, 2, 2, 2))
result <- ggml_div(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_div_inplace	<i>Element-wise Division In-place (Graph)</i>
------------------	---

Description

Creates a graph node for in-place element-wise division. Result is stored in tensor a, saving memory allocation.

Usage

```
ggml_div_inplace(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor (will be modified in-place)
b	Second tensor (same shape as a)

Value

View of tensor a with the division result

ggml_dup	<i>Duplicate Tensor (Graph)</i>
----------	---------------------------------

Description

Creates a graph node that copies a tensor. This is a graph operation that must be computed using ggml_build_forward_expand() and ggml_graph_compute(). Unlike ggml_dup_tensor which just allocates, this creates a copy operation in the graph.

Usage

```
ggml_dup(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the copy operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
b <- ggml_dup(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)
ggml_graph_compute(ctx, graph)
ggml_get_f32(b)
ggml_free(ctx)
```

ggml_dup_inplace	<i>Duplicate Tensor In-place (Graph)</i>
------------------	--

Description

Creates a graph node for in-place tensor duplication. Returns a view of the input tensor.

Usage

```
ggml_dup_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

View of tensor a

ggml_dup_tensor	<i>Duplicate Tensor</i>
-----------------	-------------------------

Description

Creates a copy of a tensor with the same shape and type

Usage

```
ggml_dup_tensor(ctx, tensor)
```

Arguments

ctx	GGML context
tensor	Tensor to duplicate

Value

New tensor pointer with same shape

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
b <- ggml_dup_tensor(ctx, a)
ggml_nelements(b)
ggml_free(ctx)
```

<code>ggml_element_size</code>	<i>Get Element Size</i>
--------------------------------	-------------------------

Description

Returns the size of a single element in the tensor.

Usage

```
ggml_element_size(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Element size in bytes

<code>ggml_elu</code>	<i>ELU Activation (Graph)</i>
-----------------------	-------------------------------

Description

Creates a graph node for ELU (Exponential Linear Unit) activation. $ELU(x) = x$ if $x > 0$, else $\alpha * (\exp(x) - 1)$ where $\alpha = 1$.

Usage

```
ggml_elu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the ELU operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
r <- ggml_elu(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r)
ggml_free(ctx)
```

ggml_elu_inplace	<i>ELU Activation In-place (Graph)</i>
------------------	--

Description

Creates a graph node for in-place ELU (Exponential Linear Unit) activation.

Usage

```
ggml_elu_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with ELU applied

ggml_embedding	<i>Create an Embedding Layer Object</i>
----------------	---

Description

Create an Embedding Layer Object

Usage

```
ggml_embedding(vocab_size, dim, name = NULL, trainable = TRUE)
```

Arguments

vocab_size	Number of distinct tokens.
dim	Embedding dimension.
name	Optional character name.
trainable	Logical.

Value

A ggml_layer object.

ggml_estimate_memory	<i>Estimate Required Memory</i>
----------------------	---------------------------------

Description

Helper function to estimate memory needed for a tensor

Usage

```
ggml_estimate_memory(type = GGML_TYPE_F32, ne0, ne1 = 1, ne2 = 1, ne3 = 1)
```

Arguments

type	Tensor type (GGML_TYPE_F32, etc)
ne0	Size of dimension 0
ne1	Size of dimension 1 (optional)
ne2	Size of dimension 2 (optional)
ne3	Size of dimension 3 (optional)

Value

Estimated memory in bytes

Examples

```
# For 1000x1000 F32 matrix
ggml_estimate_memory(GGML_TYPE_F32, 1000, 1000)
```

ggml_evaluate.ggml_functional_model

Evaluate a Trained Model

Description

Evaluate a Trained Model

Usage

```
## S3 method for class 'ggml_functional_model'
ggml_evaluate(model, x, y, batch_size = 32L, ...)

ggml_evaluate(model, ...)

## S3 method for class 'ggml_sequential_model'
ggml_evaluate(
  model,
  x,
  y,
  batch_size = 32,
  sample_weight = NULL,
  class_weight = NULL,
  ...
)
```

Arguments

model	A trained ggml_sequential_model
x	Test data
y	Test labels (one-hot encoded)
batch_size	Batch size for evaluation
...	Additional arguments (ignored).
sample_weight	Numeric vector of per-sample weights (length = nrow(x)).
class_weight	Named vector of weights per class, e.g. c("0"=1, "1"=10). Cannot be used with sample_weight.

Value

Named list with loss and accuracy.

Examples

```

ggml_set_n_threads(1L) # deterministic, single OpenMP pool
n <- 128
x <- matrix(runif(n * 4), nrow = n, ncol = 4)
y <- matrix(0, nrow = n, ncol = 2)
for (i in seq_len(n)) { y[i, if (sum(x[i,]) > 2) 1L else 2L] <- 1 }

model <- ggml_model_sequential() |>
  ggml_layer_dense(8, activation = "relu") |>
  ggml_layer_dense(2, activation = "softmax")
model$input_shape <- 4L
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")
model <- ggml_fit(model, x, y, epochs = 5, batch_size = 32, verbose = 0)

# Basic evaluation
result <- ggml_evaluate(model, x, y, batch_size = 32)

# With sample_weight
sw <- runif(n, 0.5, 1.5)
result <- ggml_evaluate(model, x, y, batch_size = 32, sample_weight = sw)

# With class_weight
result <- ggml_evaluate(model, x, y, batch_size = 32,
  class_weight = c("0" = 1, "1" = 2))

```

ggml_exp

*Exponential (Graph)***Description**

Creates a graph node for element-wise exponential: $\exp(x)$

Usage

```
ggml_exp(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the exp operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 3)
ggml_set_f32(a, c(0, 1, 2))
result <- ggml_exp(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [1, e, e^2]
ggml_free(ctx)
```

ggml_exp_inplace	<i>Exponential In-place (Graph)</i>
------------------	-------------------------------------

Description

Creates a graph node for in-place element-wise exponential: e^x

Usage

```
ggml_exp_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with exponential values

ggml_extract	<i>Extract a feature-by-cell matrix from a single-cell container</i>
--------------	--

Description

Pulls an expression matrix out of a Seurat object, a SingleCellExperiment, a sparse dgCMatrx or a plain matrix, returning a dense numeric matrix with features in rows and cells in columns — the layout the GPU engine expects. Optional genes/cells subsetting happens before any sparse-to-dense materialisation.

Usage

```
ggml_extract(x, assay = NULL, layer = "data", genes = NULL, cells = NULL, ...)

## S3 method for class 'matrix'
ggml_extract(x, assay = NULL, layer = "data", genes = NULL, cells = NULL, ...)

## S3 method for class 'dgCMatrx'
ggml_extract(
  x,
  assay = NULL,
  layer = "data",
  genes = NULL,
  cells = NULL,
  keep_sparse = FALSE,
  ...
)

## S3 method for class 'Seurat'
ggml_extract(
  x,
  assay = NULL,
  layer = "data",
  genes = NULL,
  cells = NULL,
  keep_sparse = FALSE,
  ...
)

## S3 method for class 'SingleCellExperiment'
ggml_extract(
  x,
  assay = NULL,
  layer = "logcounts",
  genes = NULL,
  cells = NULL,
  keep_sparse = FALSE,
  ...
)
```

Arguments

<code>x</code>	A Seurat, SingleCellExperiment, dgCMatrx or matrix object.
<code>assay</code>	Assay to read from. Seurat: defaults to the object's default assay. Ignored for bare matrices.
<code>layer</code>	Layer / slot to read. Seurat v5: a layer name (default "data"); Seurat v4: mapped to the slot argument of GetAssayData. Ignored for bare matrices.
<code>genes</code>	Optional character/integer vector selecting feature rows.

cells	Optional character/integer vector selecting cell columns.
...	Passed to methods.
keep_sparse	If TRUE, return the (subset) dgCMatrix without densifying. Only the LogNormalize path (op = "normalize") sets this, since it transforms the stored non-zeros in place; every other op leaves it FALSE and gets a dense matrix as before.

Details

For a SingleCellExperiment the layer argument names an assay (default "logcounts", the log-normalised matrix); it is read with SummarizedExperiment::assay().

Value

A dense numeric matrix, features x cells.

```
ggml_fit.ggml_functional_model
      Train a Model (dispatcher)
```

Description

Dispatcher: if the first argument is a ggml_sequential_model, delegates to the Keras-style high-level API (ggml_fit_sequential); otherwise delegates to the low-level optimizer loop (ggml_fit_opt).

Usage

```
## S3 method for class 'ggml_functional_model'
ggml_fit(
  model,
  x,
  y,
  epochs = 1L,
  batch_size = 32L,
  validation_split = 0,
  validation_data = NULL,
  verbose = 1L,
  ...
)

ggml_fit(model, ...)

## S3 method for class 'ggml_sequential_model'
ggml_fit(model, ...)

## Default S3 method:
ggml_fit(model, ...)
```

Arguments

<code>model</code>	A compiled model object.
<code>x</code>	Training data (matrix or array).
<code>y</code>	Training labels (matrix, one-hot encoded).
<code>epochs</code>	Number of training epochs (default: 1).
<code>batch_size</code>	Batch size (default: 32).
<code>validation_split</code>	Fraction of data for validation (default: 0).
<code>validation_data</code>	Optional list(<code>x_val</code> , <code>y_val</code>). Overrides <code>validation_split</code> .
<code>verbose</code>	0 = silent, 1 = progress (default: 1).
<code>...</code>	Arguments passed to the appropriate implementation.

Details**Keras-style (Sequential model):**

model A compiled `ggml_sequential_model`

x Training data (matrix or array)

y Training labels (matrix, one-hot encoded for classification)

epochs Number of training epochs (default: 1)

batch_size Batch size (default: 32)

validation_split Fraction of data for validation (default: 0)

validation_data Optional list(`x_val`, `y_val`) for validation. Overrides `validation_split`.

class_weight Named vector of weights per class, e.g. `c("0"=1, "1"=10)`. Cannot be used with `sample_weight`.

sample_weight Numeric vector of per-sample weights (length = `nrow(x)`). Cannot be used with `class_weight`.

verbose 0 = silent, 1 = progress (default: 1)

Low-level (optimizer loop):

sched Backend scheduler

ctx_compute Compute context

inputs Input tensor

outputs Output tensor

dataset Dataset from `ggml_opt_dataset_init()`

loss_type Loss type (default: MSE)

optimizer Optimizer type (default: AdamW)

nepoch Number of epochs (default: 10)

nbatch_logical Logical batch size (default: 32)

val_split Validation fraction (default: 0)

callbacks List of callback objects

silent Suppress output (default: FALSE)

Value

For Sequential models: the trained model (invisibly). For the low-level API: a data frame with columns epoch, train_loss, train_accuracy, val_loss, val_accuracy.

See Also

[ggml_fit_opt](#), [ggml_compile](#)

Examples

```
ggml_set_n_threads(1L) # deterministic, single OpenMP pool
n <- 128
x <- matrix(runif(n * 4), nrow = n, ncol = 4)
y <- matrix(0, nrow = n, ncol = 2)
for (i in seq_len(n)) { y[i, if (sum(x[i,]) > 2) 1L else 2L] <- 1 }

model <- ggml_model_sequential() |>
  ggml_layer_dense(8, activation = "relu") |>
  ggml_layer_dense(2, activation = "softmax")
model$input_shape <- 4L
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")

# Basic training
model <- ggml_fit(model, x, y, epochs = 5, batch_size = 32, verbose = 0)

# With validation_data
x_val <- matrix(runif(32 * 4), nrow = 32, ncol = 4)
y_val <- matrix(0, nrow = 32, ncol = 2)
for (i in seq_len(32)) { y_val[i, if (sum(x_val[i,]) > 2) 1L else 2L] <- 1 }
model <- ggml_fit(model, x, y, epochs = 3, batch_size = 32,
  validation_data = list(x_val, y_val), verbose = 0)

# With class_weight (useful for imbalanced classes)
model <- ggml_fit(model, x, y, epochs = 3, batch_size = 32,
  class_weight = c("0" = 1, "1" = 2), verbose = 0)

# With sample_weight
sw <- runif(n, 0.5, 1.5)
model <- ggml_fit(model, x, y, epochs = 3, batch_size = 32,
  sample_weight = sw, verbose = 0)
```

ggml_fit_opt

Fit model with R-side epoch loop and callbacks

Description

Trains a model epoch by epoch in R, allowing callbacks for early stopping and learning rate scheduling. Optimizer state (momentum) is preserved across all epochs.

Usage

```
ggml_fit_opt(
  sched,
  ctx_compute,
  inputs,
  outputs,
  dataset,
  loss_type = ggml_opt_loss_type_mse(),
  optimizer = ggml_opt_optimizer_type_adamw(),
  nepoch = 10L,
  nbatch_logical = 32L,
  val_split = 0,
  callbacks = list(),
  silent = FALSE
)
```

Arguments

<code>sched</code>	Backend scheduler
<code>ctx_compute</code>	Compute context (for temporary tensors)
<code>inputs</code>	Input tensor with shape <code>[ne_datapoint, batch_size]</code>
<code>outputs</code>	Output tensor with shape <code>[ne_label, batch_size]</code>
<code>dataset</code>	Dataset created with <code>'ggml_opt_dataset_init()'</code>
<code>loss_type</code>	Loss type (default: MSE)
<code>optimizer</code>	Optimizer type (default: AdamW)
<code>nepoch</code>	Number of epochs
<code>nbatch_logical</code>	Logical batch size (for gradient accumulation)
<code>val_split</code>	Fraction of data for validation (0.0 to 1.0)
<code>callbacks</code>	List of callback lists. Each element may have <code>'on_epoch_begin(epoch, logs, state)'</code> and/or <code>'on_epoch_end(epoch, logs, state)'</code> . Built-in factories: <code>'ggml_callback_early_stopping()'</code> , <code>'ggml_schedule_step_decay()'</code> , <code>'ggml_schedule_cosine_decay()'</code> , <code>'ggml_schedule_reduce_on_plateau()'</code> . <code>'state'</code> is a mutable environment with fields: <code>'stop'</code> (set TRUE to stop training), <code>'lr_ud'</code> , <code>'nepoch'</code> .
<code>silent</code>	Whether to suppress per-epoch progress output

Value

Data frame with columns `epoch`, `train_loss`, `train_accuracy`, `val_loss`, `val_accuracy`

See Also

Other optimization: [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#),

```
ggml_opt_get_lr(), ggml_opt_grad_acc(), ggml_opt_init(), ggml_opt_init_for_fit(), ggml_opt_inputs(),
ggml_opt_labels(), ggml_opt_loss(), ggml_opt_loss_type_cross_entropy(), ggml_opt_loss_type_mean(),
ggml_opt_loss_type_mse(), ggml_opt_loss_type_sum(), ggml_opt_loss_type_weighted_mse(),
ggml_opt_ncorrect(), ggml_opt_optimizer_name(), ggml_opt_optimizer_type_adamw(), ggml_opt_optimizer_type_adagrad(),
ggml_opt_outputs(), ggml_opt_pred(), ggml_opt_prepare_alloc(), ggml_opt_reset(), ggml_opt_result_accuracy(),
ggml_opt_result_free(), ggml_opt_result_init(), ggml_opt_result_loss(), ggml_opt_result_ndata(),
ggml_opt_result_pred(), ggml_opt_result_reset(), ggml_opt_set_lr(), ggml_opt_static_graphs()
```

Examples

```
if (FALSE) {
  history <- ggml_fit_opt(sched, ctx_compute, inputs, outputs, dataset,
    nepoch = 50, val_split = 0.2,
    callbacks = list(
      ggml_callback_early_stopping(monitor = "val_loss", patience = 5),
      ggml_schedule_cosine_decay()
    ))
}
```

ggml_flash_attn_back	Flash Attention Backward (Graph)
----------------------	----------------------------------

Description

Backward pass for Flash Attention. Used during training to compute gradients through attention.

Usage

```
ggml_flash_attn_back(ctx, q, k, v, d, masked = TRUE)
```

Arguments

ctx	GGML context
q	Query tensor (same as forward pass)
k	Key tensor (same as forward pass)
v	Value tensor (same as forward pass)
d	Gradient tensor from upstream (same shape as forward output)
masked	Logical: whether causal masking was used in forward pass

Value

Gradient tensor

ggml_flash_attn_ext *Flash Attention (Graph)*

Description

Creates a graph node for Flash Attention computation. This is a memory-efficient implementation of scaled dot-product attention.

Usage

```
ggml_flash_attn_ext(
    ctx,
    q,
    k,
    v,
    mask = NULL,
    scale,
    max_bias = 0,
    logit_softcap = 0
)
```

Arguments

ctx	GGML context
q	Query tensor of shape [head_dim, n_head, n_tokens, batch]
k	Key tensor of shape [head_dim, n_head_kv, n_kv, batch]
v	Value tensor of shape [head_dim, n_head_kv, n_kv, batch]
mask	Optional attention mask tensor (NULL for no mask). For causal attention, use ggml_diag_mask_inf instead.
scale	Attention scale factor, typically $1/\sqrt{\text{head_dim}}$
max_bias	Maximum ALiBi bias (0.0 to disable ALiBi)
logit_softcap	Logit soft-capping value (0.0 to disable). Used by some models like Gemma 2.

Details

Flash Attention computes: $\text{softmax}(Q * K^T / \text{scale} + \text{mask}) * V$

Key features: - Memory efficient: $O(n)$ instead of $O(n^2)$ memory for attention matrix - Supports grouped-query attention (GQA) when $n_head_kv < n_head$ - Supports multi-query attention (MQA) when $n_head_kv = 1$ - Optional ALiBi (Attention with Linear Biases) for position encoding - Optional logit soft-capping for numerical stability

Value

Attention output tensor of shape [head_dim, n_head, n_tokens, batch]

Examples

```
ctx <- ggml_init(64 * 1024 * 1024)
head_dim <- 64
n_head <- 8
n_head_kv <- 2 # GQA with 4:1 ratio
seq_len <- 32
q <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, head_dim, n_head, seq_len, 1)
k <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, head_dim, n_head_kv, seq_len, 1)
v <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, head_dim, n_head_kv, seq_len, 1)
ggml_set_f32(q, rnorm(head_dim * n_head * seq_len))
ggml_set_f32(k, rnorm(head_dim * n_head_kv * seq_len))
ggml_set_f32(v, rnorm(head_dim * n_head_kv * seq_len))
# Scale = 1/sqrt(head_dim)
scale <- 1.0 / sqrt(head_dim)
# Compute attention
out <- ggml_flash_attn_ext(ctx, q, k, v, NULL, scale, 0.0, 0.0)
graph <- ggml_build_forward_expand(ctx, out)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_floor	<i>Floor (Graph)</i>
------------	----------------------

Description

Creates a graph node for element-wise floor: floor(x)

Usage

```
ggml_floor(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the floor operation

ggml_floor_inplace	<i>Floor In-place (Graph)</i>
--------------------	-------------------------------

Description

Creates a graph node for in-place element-wise floor.

Usage

```
ggml_floor_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with floor values

ggml_free	<i>Free GGML context</i>
-----------	--------------------------

Description

Free GGML context

Usage

```
ggml_free(ctx)
```

Arguments

ctx	Context pointer
-----	-----------------

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_free(ctx)
```

ggml_freeze_weights *Freeze Layer Weights*

Description

Sets trainable = FALSE on layers, preventing their weights from being updated during training. Accepts optional from / to to freeze a range of layers by index, or layer_names to freeze by name. If none are provided, all layers are frozen.

Usage

```
ggml_freeze_weights(
  model,
  from = 1L,
  to = length(model$layers),
  layer_names = NULL,
  ...
)
```

Arguments

model	A model object (ggml_sequential_model or ggml_functional_model)
from	Integer index of the first layer to freeze (default: 1)
to	Integer index of the last layer to freeze (default: last layer)
layer_names	Character vector of layer names to freeze (overrides from/to)
...	Additional arguments passed to methods

Value

The model with selected layers frozen.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(64, activation = "relu") |>
  ggml_layer_dense(10, activation = "softmax")

# Freeze all layers
model <- ggml_freeze_weights(model)

# Freeze only the first layer
model <- ggml_freeze_weights(model, from = 1, to = 1)
```

ggml_ftype_to_ggml_type
Convert ftype to ggml_type

Description

Converts a file type (ftype) to the corresponding GGML type. Used when loading quantized models.

Usage

```
ggml_ftype_to_ggml_type(ftype)
```

Arguments

ftype	File type constant
-------	--------------------

Value

Integer GGML type

See Also

Other type_system: [ggml_blk_size\(\)](#), [ggml_is_quantized\(\)](#), [ggml_type_name\(\)](#), [ggml_type_sizef\(\)](#)

ggml_gallocr_alloc_graph
Allocate Memory for Graph

Description

Allocates memory for all tensors in the computation graph. This must be called before computing the graph.

Usage

```
ggml_gallocr_alloc_graph(galloc, graph)
```

Arguments

galloc	Graph allocator object
graph	Graph object

Value

TRUE on success, FALSE on failure

Examples

```

ctx <- ggml_init(16 * 1024 * 1024)
galloc <- ggml_gallocr_new()

# Create graph
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)

# Allocate and compute
ggml_gallocr_alloc_graph(galloc, graph)
ggml_graph_compute(ctx, graph)

ggml_gallocr_free(galloc)
ggml_free(ctx)

```

ggml_gallocr_free	<i>Free Graph Allocator</i>
-------------------	-----------------------------

Description

Frees a graph allocator and all associated buffers.

Usage

```
ggml_gallocr_free(galloc)
```

Arguments

galloc	Graph allocator object
--------	------------------------

Value

No return value, called for side effects

ggml_gallocr_get_buffer_size	<i>Get Graph Allocator Buffer Size</i>
------------------------------	--

Description

Returns the size of the buffer used by the graph allocator.

Usage

```
ggml_gallocr_get_buffer_size(galloc, buffer_id = 0L)
```

Arguments

- galloc Graph allocator object
- buffer_id Buffer ID (default: 0 for single-buffer allocator)

Value

Size in bytes

ggml_gallocr_new	<i>Create Graph Allocator</i>
------------------	-------------------------------

Description

Creates a new graph allocator for efficient memory management. The allocator can automatically allocate and reuse memory for graph tensors.

Usage

```
ggml_gallocr_new()
```

Value

Graph allocator object (external pointer)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
galloc <- ggml_gallocr_new()

a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)

# Allocate graph
ggml_gallocr_alloc_graph(galloc, graph)

ggml_gallocr_free(galloc)
ggml_free(ctx)
```

ggml_gallocr_reserve	<i>Reserve Memory for Graph</i>
----------------------	---------------------------------

Description

Pre-allocates memory for a graph. This is optional but recommended when running the same graph multiple times to avoid reallocation.

Usage

```
ggml_gallocr_reserve(galloc, graph)
```

Arguments

galloc	Graph allocator object
graph	Graph object

Value

TRUE on success, FALSE on failure

ggml_geglu	<i>GeGLU (GELU Gated Linear Unit) (Graph)</i>
------------	---

Description

Creates a graph node for GeGLU operation. GeGLU uses GELU as the activation function on the first half. CRITICAL for models like GPT-NeoX and Falcon.

Usage

```
ggml_geglu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (first dimension must be even)

Details

Formula: $output = GELU(x) * gate$

Value

Tensor with half the first dimension of input

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 8, 3)
ggml_set_f32(a, rnorm(24))
r <- ggml_geglu(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # Shape: 4x3
ggml_free(ctx)
```

<code>ggml_geglu_quick</code>	<i>GeGLU Quick (Fast GeGLU) (Graph)</i>
-------------------------------	---

Description

Creates a graph node for fast GeGLU approximation. Uses faster but less accurate GELU approximation for gating.

Usage

```
ggml_geglu_quick(ctx, a)
```

Arguments

<code>ctx</code>	GGML context
<code>a</code>	Input tensor (first dimension must be even)

Value

Tensor with half the first dimension of input

<code>ggml_geglu_split</code>	<i>GeGLU Split (Graph)</i>
-------------------------------	----------------------------

Description

Creates a graph node for GeGLU with separate input and gate tensors.

Usage

```
ggml_geglu_split(ctx, a, b)
```


Arguments

ctx	GGML context
a	Input tensor (the values to be gated)
b	Gate tensor (same shape as a)

Details

Formula: $\text{output} = \text{GELU}(a) * b$

Value

Tensor with same shape as input tensors

ggml_gelu	<i>GELU Activation (Graph)</i>
-----------	--------------------------------

Description

Creates a graph node for GELU (Gaussian Error Linear Unit) activation. CRITICAL for GPT models.

Usage

```
ggml_gelu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the GELU operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_gelu(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_gelu_erf	<i>Exact GELU Activation (Graph)</i>
---------------	--------------------------------------

Description

Creates a graph node for exact GELU using the error function (erf). $GELU(x) = x * 0.5 * (1 + erf(x / \sqrt{2}))$. More accurate than approximate GELU but potentially slower on some backends.

Usage

```
ggml_gelu_erf(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the exact GELU operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
r <- ggml_gelu_erf(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r)
ggml_free(ctx)
```

ggml_gelu_inplace	<i>GELU Activation In-place (Graph)</i>
-------------------	---

Description

Creates a graph node for in-place GELU (Gaussian Error Linear Unit) activation. CRITICAL for GPT models with memory efficiency.

Usage

```
ggml_gelu_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with GELU applied

ggml_gelu_quick	<i>GELU Quick Activation (Graph)</i>
-----------------	--------------------------------------

Description

Creates a graph node for fast approximation of GELU. Faster than standard GELU with minimal accuracy loss.

Usage

```
ggml_gelu_quick(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the GELU quick operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_gelu_quick(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result)
ggml_free(ctx)
```

ggml_get_f32	<i>Get F32 data</i>
--------------	---------------------

Description

Get F32 data

Get F32 Data

Usage

ggml_get_f32(tensor)

ggml_get_f32(tensor)

Arguments

tensor Tensor

Value

Numeric vector with tensor values

Numeric vector

Examples

```
ctx <- ggml_init(1024 * 1024)
tensor <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(tensor, c(1, 2, 3, 4, 5))
ggml_get_f32(tensor)
ggml_free(ctx)
```

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(t, c(1, 2, 3, 4, 5))
ggml_get_f32(t)
ggml_free(ctx)
```

ggml_get_f32_nd	<i>Get Single Float Value by N-D Index</i>
-----------------	--

Description

Gets a single f32 value from the tensor at position [i0, i1, i2, i3]. Works with any tensor type (auto-converts to float).

Usage

```
ggml_get_f32_nd(tensor, i0, i1 = 0, i2 = 0, i3 = 0)
```

Arguments

tensor	Tensor pointer
i0, i1, i2, i3	Indices (0-based)

Value

Float value

ggml_get_first_tensor	<i>Get First Tensor from Context</i>
-----------------------	--------------------------------------

Description

Get First Tensor from Context

Usage

```
ggml_get_first_tensor(ctx)
```

Arguments

ctx	GGML context
-----	--------------

Value

Tensor pointer or NULL

<code>ggml_get_i32</code>	<i>Get I32 Data</i>
---------------------------	---------------------

Description

Gets integer data from an I32 tensor (e.g., from `ggml_argmax`)

Usage

```
ggml_get_i32(tensor)
```

Arguments

tensor	Tensor of type GGML_TYPE_I32
--------	------------------------------

Value

Integer vector

Examples

```
ctx <- ggml_init(1024 * 1024)
pos <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 10)
ggml_set_i32(pos, 0:9)
ggml_get_i32(pos)
ggml_free(ctx)
```

<code>ggml_get_i32_nd</code>	<i>Get Single Int32 Value by N-D Index</i>
------------------------------	--

Description

Gets a single i32 value from the tensor at position `[i0, i1, i2, i3]`.

Usage

```
ggml_get_i32_nd(tensor, i0, i1 = 0, i2 = 0, i3 = 0)
```

Arguments

tensor	Tensor pointer
i0, i1, i2, i3	Indices (0-based)

Value

Integer value

ggml_get_layer	<i>Get a Layer from a Sequential Model</i>
----------------	--

Description

Retrieves a layer by name or by integer index (1-based).

Usage

```
ggml_get_layer(model, index = NULL, name = NULL)
```

Arguments

model	A ggml_sequential_model object
index	Integer index of the layer (1-based), or NULL
name	Character name of the layer, or NULL

Value

The layer list object

Examples

```
model <- ggml_model_sequential() |>  
  ggml_layer_dense(64, activation = "relu", name = "hidden") |>  
  ggml_layer_dense(10, activation = "softmax", name = "output")  
  
ggml_get_layer(model, index = 1)  
ggml_get_layer(model, name = "output")
```

ggml_get_max_tensor_size	<i>Get Maximum Tensor Size</i>
--------------------------	--------------------------------

Description

Returns the maximum tensor size that can be allocated in the context

Usage

```
ggml_get_max_tensor_size(ctx)
```

Arguments

ctx	GGML context
-----	--------------

Value

Maximum tensor size in bytes

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_get_max_tensor_size(ctx)
ggml_free(ctx)
```

<code>ggml_get_mem_size</code>	<i>Get Context Memory Size</i>
--------------------------------	--------------------------------

Description

Returns the total memory pool size of the context

Usage

```
ggml_get_mem_size(ctx)
```

Arguments

ctx GGML context

Value

Total memory size in bytes

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_get_mem_size(ctx)
ggml_free(ctx)
```

ggml_get_name	<i>Get Tensor Name</i>
---------------	------------------------

Description

Retrieves the name of a tensor.

Usage

ggml_get_name(tensor)

Arguments

tensor	Tensor pointer
--------	----------------

Value

Character string name or NULL if not set

ggml_get_next_tensor	<i>Get Next Tensor from Context</i>
----------------------	-------------------------------------

Description

Get Next Tensor from Context

Usage

ggml_get_next_tensor(ctx, tensor)

Arguments

ctx	GGML context
tensor	Current tensor

Value

Next tensor pointer or NULL

ggml_get_no_alloc	<i>Get No Allocation Mode</i>
-------------------	-------------------------------

Description

Check if no-allocation mode is enabled

Usage

```
ggml_get_no_alloc(ctx)
```

Arguments

ctx	GGML context
-----	--------------

Value

Logical indicating if no_alloc is enabled

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_get_no_alloc(ctx)
ggml_free(ctx)
```

ggml_get_n_threads	<i>Get Number of Threads</i>
--------------------	------------------------------

Description

Get the current number of threads for GGML operations

Usage

```
ggml_get_n_threads()
```

Value

Number of threads

Examples

```
ggml_get_n_threads()
```

ggml_get_op_params	<i>Get Tensor Operation Parameters</i>
--------------------	--

Description

Returns the raw op_params bytes from a tensor. These parameters control operation-specific behavior (e.g., precision, mode).

Usage

```
ggml_get_op_params(tensor)
```

Arguments

tensor	External pointer to tensor
--------	----------------------------

Value

Raw vector of op_params bytes

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_f32\(\)](#), [ggml_set_op_params_i32\(\)](#)

ggml_get_op_params_f32	<i>Get Float Op Parameter</i>
------------------------	-------------------------------

Description

Gets a single float value from tensor op_params at given index.

Usage

```
ggml_get_op_params_f32(tensor, index)
```

Arguments

tensor	External pointer to tensor
index	0-based index (0-15 for 64-byte op_params)

Value

Numeric value

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_f32\(\)](#), [ggml_set_op_params_i32\(\)](#)

ggml_get_op_params_i32	<i>Get Integer Op Parameter</i>
------------------------	---------------------------------

Description

Gets a single int32 value from tensor op_params at given index.

Usage

ggml_get_op_params_i32(tensor, index)

Arguments

tensor	External pointer to tensor
index	0-based index (0-15 for 64-byte op_params)

Value

Integer value

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_f32\(\)](#), [ggml_set_op_params_i32\(\)](#)

ggml_get_rel_pos	<i>Get Relative Position (Graph)</i>
------------------	--------------------------------------

Description

Gathers relative-position rows for relative-position attention bias.

Usage

ggml_get_rel_pos(ctx, a, qh, kh)

Arguments

ctx	GGML context
a	Input tensor
qh	Query height
kh	Key height

Value

Relative-position tensor

ggml_get_rows	<i>Get Rows by Indices (Graph)</i>
---------------	------------------------------------

Description

Creates a graph node that extracts rows from a tensor by index. This is commonly used for embedding lookup in LLMs.

Usage

ggml_get_rows(ctx, a, b)

Arguments

ctx	GGML context
a	Data tensor of shape [n_embd, n_rows, ...] - the embedding table
b	Index tensor (int32) of shape [n_indices] - which rows to extract

Details

This operation is fundamental for embedding lookup in transformers: given a vocabulary embedding matrix and token indices, it retrieves the corresponding embedding vectors.

Value

Tensor of shape [n_embd, n_indices, ...] containing the selected rows

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Create embedding matrix: 10 tokens, 4-dim embeddings
embeddings <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 10)
ggml_set_f32(embeddings, rnorm(40))
# Token indices to look up
indices <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 3)
ggml_set_i32(indices, c(0L, 5L, 2L))
# Get embeddings for tokens 0, 5, 2
```

```
result <- ggml_get_rows(ctx, embeddings, indices)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_get_rows_back	<i>Get Rows Backward (Graph)</i>
--------------------	----------------------------------

Description

Backward pass for ggml_get_rows operation. Accumulates gradients at the original row positions.

Usage

```
ggml_get_rows_back(ctx, a, b, c)
```

Arguments

ctx	GGML context
a	Gradient of get_rows output
b	Index tensor (same as forward pass)
c	Reference tensor defining output shape

Value

Gradient tensor for the embedding matrix

ggml_get_unary_op	<i>Get Unary Operation from Tensor</i>
-------------------	--

Description

Returns the unary operation type for a unary operation tensor.

Usage

```
ggml_get_unary_op(tensor)
```

Arguments

tensor	Tensor pointer (must be a unary operation result)
--------	---

Value

Integer unary operation type

See Also

Other op_info: [ggml_op_desc\(\)](#), [ggml_op_name\(\)](#), [ggml_op_symbol\(\)](#), [ggml_unary_op_name\(\)](#)

ggml_glu

*Generic GLU (Gated Linear Unit) (Graph)***Description**

Creates a graph node for GLU operation with specified gating type. GLU splits the input tensor in half along the first dimension, applies an activation to the first half (x), and multiplies it with the second half (gate).

Usage

```
ggml_glu(ctx, a, op, swapped = FALSE)
```

Arguments

ctx	GGML context
a	Input tensor (first dimension must be even)
op	GLU operation type (GGML_GLU_OP_REGLU, GGML_GLU_OP_GEGLU, etc.)
swapped	If TRUE, swap x and gate halves (default FALSE)

Details

Formula: $\text{output} = \text{activation}(x) * \text{gate}$ where x and gate are the two halves of the input tensor.

Value

Tensor with shape $[n/2, \dots]$ where n is the first dimension of input

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Create tensor with 10 columns (will be split into 5 + 5)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 4)
ggml_set_f32(a, rnorm(40))
# Apply SwiGLU
r <- ggml_glu(ctx, a, GGML_GLU_OP_SWIGLU, FALSE)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # Shape: 5x4
ggml_free(ctx)
```

GGML_GLU_OP_REGLU	<i>GLU Operation Types</i>
-------------------	----------------------------

Description

Constants for GLU (Gated Linear Unit) operation types. Used with `ggml_glu()` and `ggml_glu_split()`.

Usage

- GGML_GLU_OP_REGLU
- GGML_GLU_OP_GEGLU
- GGML_GLU_OP_SWIGLU
- GGML_GLU_OP_SWIGLU_OAI
- GGML_GLU_OP_GEGLU_ERF
- GGML_GLU_OP_GEGLU_QUICK

Format

- Integer constants
- An object of class integer of length 1.
- An object of class integer of length 1.
- An object of class integer of length 1.
- An object of class integer of length 1.
- An object of class integer of length 1.

Details

- GGML_GLU_OP_REGLU (0): ReGLU - ReLU gating
- GGML_GLU_OP_GEGLU (1): GeGLU - GELU gating (used in GPT-NeoX, Falcon)
- GGML_GLU_OP_SWIGLU (2): SwiGLU - SiLU/Swish gating (used in LLaMA, Mistral)
- GGML_GLU_OP_SWIGLU_OAI (3): SwiGLU OpenAI variant
- GGML_GLU_OP_GEGLU_ERF (4): GeGLU with exact erf implementation
- GGML_GLU_OP_GEGLU_QUICK (5): GeGLU with fast approximation

Value

An integer constant representing a GLU operation type

Examples

```
GGML_GLU_OP_REGLU      # 0 - ReLU gating
GGML_GLU_OP_GEGLU      # 1 - GELU gating
GGML_GLU_OP_SWIGLU      # 2 - SiLU/Swish gating
GGML_GLU_OP_SWIGLU_OAI  # 3 - SwiGLU OpenAI
GGML_GLU_OP_GEGLU_ERF   # 4 - GELU with erf
GGML_GLU_OP_GEGLU_QUICK # 5 - Fast GELU
```

ggml_glu_split	<i>Generic GLU Split (Graph)</i>
----------------	----------------------------------

Description

Creates a graph node for GLU with separate input and gate tensors. Unlike standard GLU which splits a single tensor, this takes two separate tensors.

Usage

```
ggml_glu_split(ctx, a, b, op)
```

Arguments

ctx	GGML context
a	Input tensor (the values to be gated)
b	Gate tensor (same shape as a)
op	GLU operation type (GGML_GLU_OP_REGLU, GGML_GLU_OP_GEGLU, etc.)

Value

Tensor with same shape as input tensors

ggml_graph_compute	<i>Compute graph</i>
--------------------	----------------------

Description

Executes all operations in the computation graph.
Executes the computation graph using CPU backend

Usage

```
ggml_graph_compute(ctx, graph)

ggml_graph_compute(ctx, graph)
```

Arguments

ctx	GGML context
graph	Graph object created by ggml_build_forward_expand

Value

NULL (invisible)
 No return value, called for side effects

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
result <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_set_f32(a, 1:10)
ggml_set_f32(b, 11:20)
c <- ggml_add(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, c)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(c)
ggml_free(ctx)
```

ggml_graph_compute_with_ctx

Compute Graph with Context (Alternative Method)

Description

Computes the computation graph using the context-based method. This is an alternative to ggml_graph_compute() that uses ggml_graph_plan() and ggml_graph_compute() internally.

Usage

```
ggml_graph_compute_with_ctx(ctx, graph, n_threads = 0L)
```

Arguments

ctx	GGML context
graph	Graph object created by ggml_build_forward_expand
n_threads	Number of threads to use (0 for auto-detect, default: 0)

Value

No return value, called for side effects

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_set_f32(a, 1:10)
c <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, c)
ggml_graph_compute_with_ctx(ctx, graph)
result <- ggml_get_f32(c)
ggml_free(ctx)
```

ggml_graph_dump_dot *Export Graph to DOT Format*

Description

Exports the computation graph to a DOT file for visualization. The DOT file can be converted to an image using Graphviz tools.

Usage

```
ggml_graph_dump_dot(graph, leafs = NULL, filename)
```

Arguments

graph	Graph object
leafs	Optional graph with leaf tensors (NULL for none)
filename	Output filename (should end with .dot)

Value

No return value, called for side effects

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)
ggml_graph_dump_dot(graph, NULL, tempfile(fileext = ".dot"))
ggml_free(ctx)
```

ggml_graph_get_tensor	<i>Get Tensor from Graph by Name</i>
-----------------------	--------------------------------------

Description

Finds a tensor in the computation graph by its name

Usage

```
ggml_graph_get_tensor(graph, name)
```

Arguments

graph	Graph object
name	Character string with tensor name

Value

Tensor pointer or NULL if not found

ggml_graph_node	<i>Get Graph Node</i>
-----------------	-----------------------

Description

Gets a specific node (tensor) from the computation graph by index

Usage

```
ggml_graph_node(graph, i)
```

Arguments

graph	Graph object
i	Node index (0-based, negative indices count from end)

Value

Tensor pointer

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_add(ctx, a, a)
graph <- ggml_build_forward_expand(ctx, b)
# Get the last node (output)
output <- ggml_graph_node(graph, -1)
ggml_free(ctx)
```

ggml_graph_n_nodes	<i>Get Number of Nodes in Graph</i>
--------------------	-------------------------------------

Description

Returns the number of computation nodes in the graph

Usage

```
ggml_graph_n_nodes(graph)
```

Arguments

graph	Graph object
-------	--------------

Value

Integer number of nodes

ggml_graph_overhead	<i>Get Graph Overhead</i>
---------------------	---------------------------

Description

Returns the memory overhead required for a computation graph

Usage

```
ggml_graph_overhead()
```

Value

Size in bytes

ggml_graph_print	<i>Print Graph Information</i>
------------------	--------------------------------

Description

Prints debug information about the computation graph

Usage

```
ggml_graph_print(graph)
```

Arguments

graph	Graph object
-------	--------------

Value

No return value, called for side effects

ggml_graph_reset	<i>Reset Graph (for backpropagation)</i>
------------------	--

Description

Resets the computation graph for a new backward pass. NOTE: This function requires the graph to have gradients allocated (used for training/backpropagation). For inference-only graphs, this function will cause an error.

Usage

```
ggml_graph_reset(graph)
```

Arguments

graph	Graph object with gradients allocated
-------	---------------------------------------

Value

No return value, called for side effects

ggml_graph_view	Create a View of a Subgraph
-----------------	-----------------------------

Description

Creates a view of a portion of a computation graph, containing nodes from index `i0` to `i1` (exclusive). The view shares the underlying nodes but does not include leaf tensors or gradients.

Usage

```
ggml_graph_view(graph, i0, i1)
```

Arguments

<code>graph</code>	External pointer to computation graph
<code>i0</code>	Start index (0-based, inclusive)
<code>i1</code>	End index (exclusive)

Value

External pointer to graph view

See Also

Other graph: [ggml_op_can_inplace\(\)](#)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
b <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, b)
n_nodes <- ggml_graph_n_nodes(graph)
view <- ggml_graph_view(graph, 0, n_nodes)
ggml_free(ctx)
```

ggml_group_norm	<i>Group Normalization (Graph)</i>
-----------------	------------------------------------

Description

Creates a graph node for group normalization. Normalizes along $ne0*ne1*n_groups$ dimensions. Used in Stable Diffusion and other image generation models.

Usage

```
ggml_group_norm(ctx, a, n_groups, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor
n_groups	Number of groups to divide channels into
eps	Epsilon for numerical stability (default 1e-5)

Value

Tensor representing the group norm operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# 4 channels, 2 groups (2 channels per group)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 8)
ggml_set_f32(a, rnorm(32))
result <- ggml_group_norm(ctx, a, n_groups = 2)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_group_norm_inplace	<i>Group Normalization In-place (Graph)</i>
-------------------------	---

Description

Creates a graph node for in-place group normalization.

Usage

```
ggml_group_norm_inplace(ctx, a, n_groups, eps = 1e-05)
```


Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
n_groups	Number of groups
eps	Epsilon for numerical stability (default 1e-5)

Value

View of input tensor with group norm applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 8)
ggml_set_f32(a, rnorm(32))
result <- ggml_group_norm_inplace(ctx, a, n_groups = 2)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_gru

Create a GRU Layer Object

Description

Create a GRU Layer Object

Usage

```
ggml_gru(
  units,
  return_sequences = FALSE,
  activation = "tanh",
  recurrent_activation = "sigmoid",
  name = NULL,
  trainable = TRUE
)
```

Arguments

units	Integer, number of hidden units.
return_sequences	Logical.
activation	Candidate activation (default "tanh").

recurrent_activation Gate activation (default "sigmoid").
name Optional character name.
trainable Logical.

Value

A ggml_layer object.

ggml_hardsigmoid	<i>Hard Sigmoid Activation (Graph)</i>
------------------	--

Description

Creates a graph node for Hard Sigmoid activation. $\text{HardSigmoid}(x) = \text{ReLU6}(x + 3) / 6 = \min(\max(0, x + 3), 6) / 6$. A computationally efficient approximation of the sigmoid function.

Usage

ggml_hardsigmoid(ctx, a)

Arguments

ctx GGML context
a Input tensor

Value

Tensor representing the Hard Sigmoid operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-4, -1, 0, 1, 4))
r <- ggml_hardsigmoid(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # [0, 0.333, 0.5, 0.667, 1]
ggml_free(ctx)
```

ggml_hardswish	<i>Hard Swish Activation (Graph)</i>
----------------	--------------------------------------

Description

Creates a graph node for Hard Swish activation. $\text{HardSwish}(x) = x * \text{ReLU6}(x + 3) / 6 = x * \min(\max(0, x + 3), 6) / 6$. Used in MobileNetV3 and other efficient architectures.

Usage

```
ggml_hardswish(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the Hard Swish operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-4, -1, 0, 1, 4))
r <- ggml_hardswish(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r)
ggml_free(ctx)
```

ggml_im2col	<i>Image to Column (Graph)</i>
-------------	--------------------------------

Description

Transforms image data into column format for efficient convolution. This is a low-level operation used internally by convolution implementations.

Usage

```
ggml_im2col(  
    ctx,  
    a,  
    b,  
    s0,  
    s1,  
    p0,  
    p1,  
    d0,  
    d1,  
    is_2D = TRUE,  
    dst_type = GGML_TYPE_F16  
)
```

Arguments

ctx	GGML context
a	Convolution kernel tensor
b	Input data tensor
s0	Stride dimension 0
s1	Stride dimension 1
p0	Padding dimension 0
p1	Padding dimension 1
d0	Dilation dimension 0
d1	Dilation dimension 1
is_2D	Whether this is a 2D operation (default TRUE)
dst_type	Output type (default GGML_TYPE_F16)

Value

Transformed tensor in column format

ggml_init	<i>Initialize GGML context</i>
-----------	--------------------------------

Description

Initialize GGML context

Usage

```
ggml_init(mem_size = 16 * 1024 * 1024, no_alloc = FALSE)
```

Arguments

mem_size	Memory size in bytes
no_alloc	If TRUE, don't allocate memory for tensors (default: FALSE)

Value

GGML context pointer

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_free(ctx)
```

ggml_init_auto	<i>Create Context with Auto-sizing</i>
----------------	--

Description

Creates a context with automatically calculated size based on planned tensors.

Usage

```
ggml_init_auto(..., extra_mb = 10, type = GGML_TYPE_F32, no_alloc = FALSE)
```

Arguments

...	Named arguments with tensor dimensions (integer vectors)
extra_mb	Extra megabytes to add (default: 10)
type	Tensor type (default: GGML_TYPE_F32)
no_alloc	If TRUE, don't allocate memory for tensors (default: FALSE)

Value

GGML context

Examples

```
ctx <- ggml_init_auto(mat1 = c(1000L, 1000L), mat2 = c(1000L, 1000L))
ggml_free(ctx)
```

ggml_inject

*Inject a single-cell result back into its container***Description**

Writes a `ggml_result` back into the appropriate slot of a Seurat or SingleCellExperiment object, returning the updated object. The destination depends on `result$metadata$kind`: a dimensionality reduction (the default) goes into a `DimReduc`; a "transform" (normalize / scale) overwrites an assay layer; a "graph" (neighbors) writes `<assay>_nn` and `<assay>_snn` Graph objects into `@graphs`, exactly where Seurat's `FindNeighbors` puts them so `FindClusters` can consume them. Component standard deviations and the backend used are recorded alongside so downstream tools and the user can see how the result was produced.

Usage

```
ggml_inject(
  x,
  result,
  reduction_name = "ggml",
  key = "GGML_",
  assay = NULL,
  ...
)

## S3 method for class 'Seurat'
ggml_inject(
  x,
  result,
  reduction_name = "ggml",
  key = "GGML_",
  assay = NULL,
  ...
)

## S3 method for class 'SingleCellExperiment'
ggml_inject(
  x,
  result,
  reduction_name = "ggml",
  key = "GGML_",
  assay = NULL,
  ...
)
```

Arguments

`x` A Seurat object (the one the data was extracted from).

result	A <code>ggml_result</code> .
reduction_name	Name of the reduction slot to create, e.g. "ggml" (default). For Seurat this becomes <code>x[["ggml"]]</code> .
key	Column-name prefix for the embedding, e.g. "GGM_L_".
assay	Assay to associate the reduction with (Seurat). Defaults to the object's default assay.
...	Passed to methods.

Value

The updated container.

ggml_input	<i>Declare a Functional API Input Tensor</i>
------------	--

Description

Creates a symbolic input node for the Functional API. The node records only the *shape* of one sample (without batch dimension); actual memory is allocated when `ggml_compile()` is called.

Usage

```
ggml_input(shape, name = NULL, dtype = "float32")
```

Arguments

shape	Integer vector describing the shape of a single sample. For flat feature vectors use a scalar, e.g. <code>shape = 64L</code> . For 2-D inputs (sequences) use <code>c(length, channels)</code> . For 3-D inputs (images) use <code>c(H, W, C)</code> .
name	Optional character name for the input tensor.
dtype	Data type of the input: "float32" (default) or "int32" (for embedding/token-index inputs).

Value

A `ggml_tensor_node` object.

Examples

```
x <- ggml_input(shape = 64L)
x <- ggml_input(shape = c(28L, 28L, 1L), name = "image")
x <- ggml_input(shape = 10L, dtype = "int32") # token indices
```

<code>ggml_is_available</code>	<i>Check if GGML is available</i>
--------------------------------	-----------------------------------

Description

Check if GGML is available

Usage

`ggml_is_available()`

Value

TRUE if GGML library is loaded

Examples

`ggml_is_available()`

<code>ggml_is_contiguous</code>	<i>Check if Tensor is Contiguous</i>
---------------------------------	--------------------------------------

Description

Returns TRUE if tensor data is stored contiguously in memory

Usage

`ggml_is_contiguous(tensor)`

Arguments

tensor Tensor pointer

Value

Logical

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_is_contiguous(t)
ggml_free(ctx)
```

ggml_is_contiguously_allocated
<i>Check If Tensor is Contiguously Allocated</i>

Description

Check if tensor data is contiguously allocated in memory. Different from contiguous layout - this checks the actual allocation.

Usage

ggml_is_contiguously_allocated(tensor)

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating if data is contiguously allocated

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#)

ggml_is_contiguous_0	<i>Check Tensor Contiguity (Dimension 0)</i>
----------------------	--

Description

Check if tensor is contiguous. Same as ggml_is_contiguous.

Usage

ggml_is_contiguous_0(tensor)

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating contiguity

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_is_contiguous_1 *Check Tensor Contiguity (Dimensions >= 1)*

Description

Check if tensor is contiguous for dimensions >= 1. Allows non-contiguous first dimension.

Usage

```
ggml_is_contiguous_1(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating contiguity for dims >= 1

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_is_contiguous_2 *Check Tensor Contiguity (Dimensions >= 2)*

Description

Check if tensor is contiguous for dimensions >= 2. Allows non-contiguous first two dimensions.

Usage

```
ggml_is_contiguous_2(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating contiguity for dims >= 2

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_is_contiguous_channels
<i>Check Channel-wise Contiguity</i>

Description

Check if tensor has contiguous channels (important for CNN operations). Data for each channel should be stored contiguously.

Usage

ggml_is_contiguous_channels(tensor)

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating channel-wise contiguity

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_rows\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_is_contiguous_rows	<i>Check Row-wise Contiguity</i>
-------------------------	----------------------------------

Description

Check if tensor has contiguous rows (important for matrix operations). Each row should be stored contiguously in memory.

Usage

```
ggml_is_contiguous_rows(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical indicating row-wise contiguity

See Also

Other tensor_layout: [ggml_are_same_stride\(\)](#), [ggml_can_repeat\(\)](#), [ggml_count_equal\(\)](#), [ggml_is_contiguous_0\(\)](#), [ggml_is_contiguous_1\(\)](#), [ggml_is_contiguous_2\(\)](#), [ggml_is_contiguous_channels\(\)](#), [ggml_is_contiguously_allocated\(\)](#)

ggml_is_permuted	<i>Check if Tensor is Permuted</i>
------------------	------------------------------------

Description

Returns TRUE if tensor dimensions have been permuted

Usage

```
ggml_is_permuted(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Logical

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_is_permuted(t)
ggml_free(ctx)
```

ggml_is_quantized	<i>Check If Type is Quantized</i>
-------------------	-----------------------------------

Description

Returns TRUE if the GGML type is a quantized format.

Usage

```
ggml_is_quantized(type)
```

Arguments

type	GGML type constant
------	--------------------

Value

Logical indicating if type is quantized

See Also

Other type_system: [ggml_blk_size\(\)](#), [ggml_ftype_to_ggml_type\(\)](#), [ggml_type_name\(\)](#), [ggml_type_sizef\(\)](#)

Examples

```
ggml_is_quantized(GGML_TYPE_F32) # FALSE
ggml_is_quantized(GGML_TYPE_Q4_0) # TRUE
```

ggml_is_transposed	<i>Check if Tensor is Transposed</i>
--------------------	--------------------------------------

Description

Returns TRUE if tensor has been transposed

Usage

```
ggml_is_transposed(tensor)
```

Arguments

tensor Tensor pointer

Value

Logical

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_is_transposed(t)
ggml_free(ctx)
```

ggml_l2_norm	<i>L2 Normalization (Graph)</i>
--------------	---------------------------------

Description

Creates a graph node for L2 normalization (unit norm). Normalizes vectors to unit length: $x / \|x\|_2$.
Used in RWKV v7 and embedding normalization.

Usage

```
ggml_l2_norm(ctx, a, eps = 1e-05)
```

Arguments

ctx GGML context
a Input tensor
eps Epsilon for numerical stability (default 1e-5)

Value

Tensor representing the L2 norm operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(3, 0, 0, 4)) # Length = 5
result <- ggml_l2_norm(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [0.6, 0, 0, 0.8] unit vector
ggml_free(ctx)
```

ggml_l2_norm_inplace *L2 Normalization In-place (Graph)*

Description

Creates a graph node for in-place L2 normalization.

Usage

```
ggml_l2_norm_inplace(ctx, a, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
eps	Epsilon for numerical stability (default 1e-5)

Value

View of input tensor with L2 norm applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(3, 0, 0, 4))
result <- ggml_l2_norm_inplace(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_layer_add	<i>Element-wise Addition of Two Tensor Nodes</i>
----------------	--

Description

Adds two (or more) tensor nodes element-wise. All tensors must have the same shape. This is the functional equivalent of a residual / skip connection.

Usage

```
ggml_layer_add(tensors, name = NULL)
```

Arguments

tensors	A list of ggml_tensor_node objects (length >= 2).
name	Optional character name for the layer.

Value

A new ggml_tensor_node representing the sum.

Examples

```
x  <- ggml_input(shape = 64L)
a  <- x |> ggml_layer_dense(64, activation = "relu")
b  <- x |> ggml_layer_dense(64)
out <- ggml_layer_add(list(a, b))
```

ggml_layer_batch_norm	<i>Add Batch Normalization Layer</i>
-----------------------	--------------------------------------

Description

Applies normalization: RMS-normalizes the input, then scales by gamma and shifts by beta (both learnable). Uses ggml_rms_norm which supports backward pass for training.

Usage

```
ggml_layer_batch_norm(model, eps = 1e-05, name = NULL, trainable = TRUE)
```

Arguments

model	A ggml_sequential_model object
eps	Small constant for numerical stability (default 1e-5)
name	Optional character name for the layer.
trainable	Logical; whether the layer weights are updated during training.

Value

The model object with the batch_norm layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(128, input_shape = 784) |>
  ggml_layer_batch_norm() |>
  ggml_layer_dense(10, activation = "softmax")
```

ggml_layer_concatenate

Concatenate Tensor Nodes Along an Axis

Description

Concatenates two or more tensor nodes along the specified axis.

Usage

```
ggml_layer_concatenate(tensors, axis = 0L, name = NULL)
```

Arguments

tensors	A list of ggml_tensor_node objects (length >= 2).
axis	Integer axis along which to concatenate (0-based, ggml convention). Default 0L concatenates along the first dimension (features for flat tensors).
name	Optional character name for the layer.

Value

A new ggml_tensor_node representing the concatenated tensor.

Examples

```
x <- ggml_input(shape = 32L)
y <- ggml_input(shape = 32L)
out <- ggml_layer_concatenate(list(x, y), axis = 0L)
```

ggml_layer_conv_1d *Create a Conv1D Layer Object*

Description

Create a Conv1D Layer Object

Add 1D Convolution Layer

Usage

```
ggml_layer_conv_1d(
    model,
    filters,
    kernel_size,
    activation = NULL,
    input_shape = NULL,
    strides = 1L,
    padding = "valid",
    name = NULL,
    trainable = TRUE
)
```

```
ggml_layer_conv_1d(
    model,
    filters,
    kernel_size,
    activation = NULL,
    input_shape = NULL,
    strides = 1L,
    padding = "valid",
    name = NULL,
    trainable = TRUE
)
```

Arguments

model	A ggml_sequential_model object
filters	Number of output filters
kernel_size	Integer kernel size
activation	Activation function name: "relu", "sigmoid", "tanh", "softmax", or NULL
input_shape	Input shape c(L, C) - required for first layer only (length, channels)
strides	Integer stride (default 1)
padding	"valid" (no padding) or "same" (preserve length)
name	Optional character name for the layer.
trainable	Logical; whether the layer weights are updated during training.

Value

A ggml_layer object.

The model object with the conv_1d layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_1d(32, 3, activation = "relu",
                    input_shape = c(100, 1))
```

ggml_layer_conv_2d	<i>Create a Conv2D Layer Object</i>
--------------------	-------------------------------------

Description

Create a Conv2D Layer Object

Add 2D Convolution Layer

Usage

```
ggml_layer_conv_2d(  
  model,  
  filters,  
  kernel_size,  
  activation = NULL,  
  input_shape = NULL,  
  strides = c(1L, 1L),  
  padding = "valid",  
  name = NULL,  
  trainable = TRUE  
)
```

```
ggml_layer_conv_2d(  
  model,  
  filters,  
  kernel_size,  
  activation = NULL,  
  input_shape = NULL,  
  strides = c(1L, 1L),  
  padding = "valid",  
  name = NULL,  
  trainable = TRUE  
)
```

Arguments

model	A ggml_sequential_model object
filters	Number of output filters
kernel_size	Integer or vector of 2 integers for kernel height and width
activation	Activation function name: "relu", "sigmoid", "tanh", "softmax", or NULL
input_shape	Input shape c(H, W, C) - required for first layer only
strides	Integer or vector of 2 integers for stride
padding	"valid" (no padding) or "same" (preserve spatial dims)
name	Optional character name for the layer.
trainable	Logical; whether the layer weights are updated during training.

Value

A ggml_layer object.
The model object with the conv_2d layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1))
```

ggml_layer_dense	<i>Add Dense (Fully Connected) Layer</i>
------------------	--

Description

Add Dense (Fully Connected) Layer

Usage

```
ggml_layer_dense(
  model,
  units,
  activation = NULL,
  input_shape = NULL,
  name = NULL,
  trainable = TRUE
)
```

Arguments

model	A ggml_sequential_model object
units	Number of output units
activation	Activation function name: "relu", "sigmoid", "tanh", "softmax", or NULL
input_shape	Integer or integer vector specifying the input shape (only needed for the first layer)
name	Optional character name for the layer.
trainable	Logical; whether the layer weights are updated during training.

Value

The model object with the dense layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_flatten() |>
  ggml_layer_dense(128, activation = "relu")
```

ggml_layer_dropout	<i>Add Dropout Layer</i>
--------------------	--------------------------

Description

Applies dropout regularization. During training, multiplies all activations by (1 - rate) (deterministic expected-value scaling). During inference (training = FALSE), the layer is an identity (no change).

Usage

```
ggml_layer_dropout(
  model,
  rate,
  stochastic = FALSE,
  name = NULL,
  trainable = FALSE
)
```

Arguments

model	A ggml_sequential_model or ggml_tensor_node.
rate	Dropout rate in $[0, 1)$. Fraction of units to "drop".
stochastic	Logical. If TRUE, use inverted dropout with a random Bernoulli mask regenerated each epoch (proper regularization). If FALSE (default), use deterministic scaling by $(1 - \text{rate})$ — cheaper but weaker regularization.
name	Optional layer name.
trainable	Ignored for dropout (no weights); kept for API consistency.

Value

The model with the dropout layer appended, or a new tensor node.

Difference from Keras / inverted dropout

Keras implements *inverted dropout*: during training it applies a random Bernoulli mask and scales surviving activations *up* by $1 / (1 - \text{rate})$, so the expected value of each unit is preserved and no scaling is needed at inference.

This implementation uses *deterministic scaling* (multiply by $(1 - \text{rate})$ at training, identity at inference) — equivalent in expected value but without stochastic noise. Consequences:

- No random mask → the regularization signal is weaker (no co-adaptation breaking).
- Activations at training are scaled *down*, not up — the magnitude seen by subsequent layers differs from Keras behaviour.
- Results are fully deterministic and reproducible without setting a seed.

Note

With `stochastic = TRUE` the Bernoulli mask is regenerated once per epoch (not per batch), because `ggml_opt_fit` processes all batches inside a single C call. This is weaker than per-batch dropout but stronger than the deterministic variant.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(128, activation = "relu", input_shape = 784L) |>
  ggml_layer_dropout(0.5, stochastic = TRUE) |>
  ggml_layer_dense(10, activation = "softmax")
```

ggml_layer_embedding *Add Embedding Layer*

Description

Looks up dense vectors for integer token indices. The input must be an integer matrix of 0-based indices in $[0, \text{vocab_size} - 1]$ (use `ggml_input(shape, dtype = "int32")` in Functional mode).

Usage

```
ggml_layer_embedding(model, vocab_size, dim, name = NULL, trainable = TRUE)
```

Arguments

<code>model</code>	A <code>ggml_sequential_model</code> or <code>ggml_tensor_node</code> .
<code>vocab_size</code>	Number of distinct tokens (vocabulary size).
<code>dim</code>	Embedding dimension (vector length per token).
<code>name</code>	Optional layer name.
<code>trainable</code>	Logical; whether embedding weights are updated during training.

Value

The model with the embedding layer appended, or a new tensor node.

Axis order (ggml vs Keras)

ggml stores tensors in column-major order, so the output shape is $[\text{dim}, \text{seq_len}]$ per sample (ggml convention) rather than $[\text{seq_len}, \text{dim}]$ as in Keras. When you call `ggml_layer_flatten()` after embedding the result is the same flattened vector regardless of order, but if you access raw output tensors be aware of this transposition.

Index validation

Indices must be in $[0, \text{vocab_size} - 1]$. Out-of-range values cause undefined behaviour inside the ggml kernel (no bounds check is performed at the R level).

Examples

```
inp <- ggml_input(shape = 10L, dtype = "int32")
out <- inp |>
  ggml_layer_embedding(vocab_size = 1000L, dim = 32L) |>
  ggml_layer_flatten() |>
  ggml_layer_dense(10L, activation = "softmax")
model <- ggml_model(inputs = inp, outputs = out)
```

ggml_layer_flatten *Add Flatten Layer*

Description

Flattens the spatial dimensions into a single vector per sample.

Usage

```
ggml_layer_flatten(model, name = NULL, trainable = TRUE)
```

Arguments

model	A ggml_sequential_model object
name	Optional character name for the layer.
trainable	Logical; reserved for API consistency (no weights).

Value

The model object with the flatten layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_flatten()
```

ggml_layer_global_average_pooling_2d
Global Average Pooling for 2D Feature Maps

Description

Reduces a [H, W, C] feature map to [C] by averaging all spatial positions per channel. Equivalent to Keras GlobalAveragePooling2D().

Usage

```
ggml_layer_global_average_pooling_2d(model, name = NULL, trainable = TRUE)
```

Arguments

model	A ggml_sequential_model or ggml_tensor_node.
name	Optional character name for the layer.
trainable	Logical; reserved for API consistency (no weights).

Value

Updated model or a new ggml_tensor_node.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_global_average_pooling_2d() |>
  ggml_layer_dense(10, activation = "softmax")
```

ggml_layer_global_max_pooling_2d

Global Max Pooling for 2D Feature Maps

Description

Reduces a [H, W, C] feature map to [C] by taking the maximum value per channel across all spatial positions. Equivalent to Keras GlobalMaxPooling2D().

Usage

```
ggml_layer_global_max_pooling_2d(model, name = NULL, trainable = TRUE)
```

Arguments

model	A ggml_sequential_model or ggml_tensor_node.
name	Optional character name for the layer.
trainable	Logical; reserved for API consistency (no weights).

Value

Updated model or a new ggml_tensor_node.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_global_max_pooling_2d() |>
  ggml_layer_dense(10, activation = "softmax")
```

ggml_layer_gru	<i>Add a GRU Layer</i>
----------------	------------------------

Description

Gated Recurrent Unit recurrent layer. Implemented as an unrolled computation graph (BPTT).

Usage

```
ggml_layer_gru(
    model,
    units,
    return_sequences = FALSE,
    activation = "tanh",
    recurrent_activation = "sigmoid",
    input_shape = NULL,
    name = NULL,
    trainable = TRUE
)
```

Arguments

model	A ggml_sequential_model or ggml_tensor_node.
units	Integer, number of hidden units.
return_sequences	Logical; return all hidden states or only the last.
activation	Activation for the candidate hidden state ("tanh").
recurrent_activation	Activation for z/r gates ("sigmoid").
input_shape	Input shape c(seq_len, input_size) – required for the first layer only.
name	Optional layer name.
trainable	Logical.

Value

Updated model or a new ggml_tensor_node.

Weight layout

- `W_zh [input_size, 2*units]` — input kernel for z and r gates.
- `U_zh [units, 2*units]` — recurrent kernel for z and r.
- `b_zh [2*units]` — bias for z and r.
- `W_n [input_size, units]` — input kernel for candidate.
- `U_n [units, units]` — recurrent kernel for candidate.
- `b_n [units]` — bias for candidate.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_gru(64L, input_shape = c(10L, 32L)) |>
  ggml_layer_dense(10L, activation = "softmax")
```

ggml_layer_lstm

*Add an LSTM Layer***Description**

Long Short-Term Memory recurrent layer. Implemented as an unrolled computation graph (BPTT) so that ggml's automatic differentiation works without any C extensions.

Usage

```
ggml_layer_lstm(
  model,
  units,
  return_sequences = FALSE,
  activation = "tanh",
  recurrent_activation = "sigmoid",
  input_shape = NULL,
  name = NULL,
  trainable = TRUE
)
```

Arguments

model	A ggml_sequential_model or ggml_tensor_node.
units	Integer, number of hidden units.
return_sequences	Logical; if TRUE return all hidden states, otherwise return only the last hidden state.
activation	Activation for the cell gate (default "tanh").
recurrent_activation	Activation for the recurrent step (default "sigmoid").
input_shape	Input shape c(seq_len, input_size) – required for the first layer only.
name	Optional layer name.
trainable	Logical.

Value

Updated model or a new ggml_tensor_node.

Weight layout

- `W_gates` [`input_size`, `4*units`] — input kernel for all four gates (i, f, g, o) concatenated.
- `U_gates` [`units`, `4*units`] — recurrent kernel.
- `b_gates` [`4*units`] — bias.

Input / output shapes

Input: [`seq_len`, `input_size`] per sample (R row-major), or a 3-D array [`N`, `seq_len`, `input_size`].
 In the Functional API the input node shape should be `c(seq_len, input_size)`.

Output (Sequential): [`units`] per sample when `return_sequences = FALSE` (default), or `c(seq_len, units)` when `return_sequences = TRUE`.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_lstm(64L, input_shape = c(10L, 32L)) |>
  ggml_layer_dense(10L, activation = "softmax")
```

```
ggml_layer_max_pooling_2d
```

Add 2D Max Pooling Layer

Description

Add 2D Max Pooling Layer

Usage

```
ggml_layer_max_pooling_2d(
  model,
  pool_size = c(2L, 2L),
  strides = NULL,
  name = NULL,
  trainable = TRUE
)
```

Arguments

<code>model</code>	A <code>ggml_sequential_model</code> object
<code>pool_size</code>	Integer or vector of 2 integers for pool height and width
<code>strides</code>	Integer or vector of 2 integers (defaults to <code>pool_size</code>)
<code>name</code>	Optional character name for the layer.
<code>trainable</code>	Logical; reserved for API consistency (no weights).

Value

The model object with the max pooling layer appended (invisibly).

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_max_pooling_2d(c(2, 2))
```

ggml_leaky_relu	<i>Leaky ReLU Activation (Graph)</i>
-----------------	--------------------------------------

Description

Creates a graph node for Leaky ReLU activation. $\text{LeakyReLU}(x) = x$ if $x > 0$, else $\text{negative_slope} * x$. Unlike standard ReLU, Leaky ReLU allows a small gradient for negative values.

Usage

```
ggml_leaky_relu(ctx, a, negative_slope = 0.01, inplace = FALSE)
```

Arguments

ctx	GGML context
a	Input tensor
negative_slope	Slope for negative values (default: 0.01)
inplace	If TRUE, operation is performed in-place (default: FALSE)

Value

Tensor representing the Leaky ReLU operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
r <- ggml_leaky_relu(ctx, a, negative_slope = 0.1)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # [-0.2, -0.1, 0, 1, 2]
ggml_free(ctx)
```

ggml_load_model	<i>Load a Full Model (Architecture + Weights)</i>
-----------------	---

Description

Restores a model previously saved with `ggml_save_model()`. The returned model is compiled and ready for `ggml_predict()` / `ggml_evaluate()`. Call `ggml_fit()` again to continue training.

Usage

```
ggml_load_model(path, backend = "auto")
```

Arguments

path	File path to an RDS file written by <code>ggml_save_model()</code> .
backend	Backend selection: "auto", "cpu", or "vulkan".

Value

A compiled model object.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(16L, activation = "relu", input_shape = 4L) |>
  ggml_layer_dense(2L, activation = "softmax")
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")
x <- matrix(runif(64 * 4), 64, 4)
y <- matrix(c(rep(c(1,0), 32), rep(c(0,1), 32)), 64, 2)
model <- ggml_fit(model, x, y, epochs = 1L, batch_size = 32L, verbose = 0L)
tmp <- tempfile(fileext = ".rds")
ggml_save_model(model, tmp)
model2 <- ggml_load_model(tmp)
```

ggml_load_weights	<i>Load Model Weights from File</i>
-------------------	-------------------------------------

Description

Loads previously saved weights into a compiled model. The model architecture must match the saved weights (same layer types, sizes, and shapes).

Usage

```
ggml_load_weights(model, path)
```

Arguments

- model A compiled ggml_sequential_model (same architecture as saved)
- path File path to load weights from

Value

The model with loaded weights.

ggml_log	<i>Natural Logarithm (Graph)</i>
----------	----------------------------------

Description

Creates a graph node for element-wise natural logarithm: $\log(x)$

Usage

```
ggml_log(ctx, a)
```

Arguments

- ctx GGML context
- a Input tensor

Value

Tensor representing the log operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 3)
ggml_set_f32(a, c(1, exp(1), exp(2)))
result <- ggml_log(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [0, 1, 2]
ggml_free(ctx)
```

<code>ggml_log_inplace</code>	<i>Natural Logarithm In-place (Graph)</i>
-------------------------------	---

Description

Creates a graph node for in-place element-wise natural logarithm.

Usage

```
ggml_log_inplace(ctx, a)
```

Arguments

<code>ctx</code>	GGML context
<code>a</code>	Input tensor (will be modified in-place)

Value

View of tensor `a` with log values

<code>ggml_log_is_r_enabled</code>	<i>Check if R Logging is Enabled</i>
------------------------------------	--------------------------------------

Description

Check if R Logging is Enabled

Usage

```
ggml_log_is_r_enabled()
```

Value

Logical indicating if R-compatible logging is active

See Also

Other logging: [ggml_abort_is_r_enabled\(\)](#), [ggml_log_set_default\(\)](#), [ggml_log_set_r\(\)](#), [ggml_set_abort_callback_default\(\)](#), [ggml_set_abort_callback_r\(\)](#)

ggml_log_set_default	<i>Restore Default GGML Logging</i>
----------------------	-------------------------------------

Description

Restores GGML to default logging behavior (stderr output).

Usage

```
ggml_log_set_default()
```

Value

NULL invisibly

See Also

Other logging: [ggml_abort_is_r_enabled\(\)](#), [ggml_log_is_r_enabled\(\)](#), [ggml_log_set_r\(\)](#), [ggml_set_abort_callback_default\(\)](#), [ggml_set_abort_callback_r\(\)](#)

ggml_log_set_r	<i>Enable R-compatible GGML Logging</i>
----------------	---

Description

Redirects GGML log messages to R's message system: - INFO/DEBUG messages go to stdout (via Rprintf) - WARN/ERROR messages go to stderr (via REprintf)

Usage

```
ggml_log_set_r()
```

Value

NULL invisibly

See Also

Other logging: [ggml_abort_is_r_enabled\(\)](#), [ggml_log_is_r_enabled\(\)](#), [ggml_log_set_default\(\)](#), [ggml_set_abort_callback_default\(\)](#), [ggml_set_abort_callback_r\(\)](#)

Examples

```
ggml_log_set_r()
# Now GGML messages will appear in R console
```

ggml_lstm	<i>Create an LSTM Layer Object</i>
-----------	------------------------------------

Description

Create an LSTM Layer Object

Usage

```
ggml_lstm(
  units,
  return_sequences = FALSE,
  activation = "tanh",
  recurrent_activation = "sigmoid",
  name = NULL,
  trainable = TRUE
)
```

Arguments

units	Integer, number of hidden units.
return_sequences	Logical.
activation	Cell gate activation (default "tanh").
recurrent_activation	Recurrent gate activation (default "sigmoid").
name	Optional character name.
trainable	Logical.

Value

A ggml_layer object.

ggml_marshal_model	<i>Marshal a ggmlR model to an in-memory container</i>
--------------------	--

Description

Serializes a trained sequential or functional ggmlR model into a self-describing raw container suitable for transport between R sessions or parallel workers (e.g. for **mlr3** parallel resampling and tuning).

Usage

```
ggml_marshal_model(model)
```

Arguments

model A compiled ggml_sequential_model or ggml_functional_model.

Details

The container wraps the bytes produced by [ggml_save_model](#) together with a format tag, schema version, package/R versions, a SHA-256 integrity checksum, and a timestamp. Autograd modules are **not** supported in this version and cause the function to signal an error; the mlr3 learners catch this and fall back to `marshaled = FALSE`.

Value

A named list with class "ggmlR_marshaled" containing the serialized payload and metadata. Pass it to [ggml_unmarshal_model](#) to reconstruct the model.

See Also

[ggml_unmarshal_model](#), [ggml_save_model](#)

ggml_matmul	<i>GPU matrix multiply (drop-in for %*%)</i>
-------------	--

Description

Computes $A \%*\% B$ on the Vulkan GPU and returns an ordinary R matrix, with a transparent CPU fallback. A drop-in accelerator for a large matrix multiply: no autograd, no wrapper types required — plain matrices in and out.

Usage

```
ggml_matmul(A, B, device = "auto", prec = "f32")
```

Arguments

A, B	Numeric matrices with <code>ncol(A) == nrow(B)</code> .
device	"auto" (default: GPU when present and the multiply is large enough to amortise the transfer, else CPU), "gpu" (force GPU, still falls back to CPU if none), or "cpu".
prec	"f32" (default; requests f32 accumulation, ~1e-3 relative error or better depending on the driver) or "f16" (faster, never more precise). Only affects the GPU path.

Details

Precision: R multiplies in double precision (f64); a GPU offers f32 at best, so the GPU result never matches R to full double precision. `prec = "f32"` (the default) requests the f32 accumulation kernel, but how close the result actually lands depends on the Vulkan driver: some accumulate `mul_mat` in f16 regardless (e.g. RADV / Mesa), giving a relative error around $1e^{-3}$ either way. Treat the GPU path as a fast, approximate multiply — typically $1e^{-3}$, better on drivers with true f32 accumulation — not a bit-for-bit replacement for `%%`. `prec = "f16"` only ever lowers precision; use it when speed matters more than the last digits.

Value

The product `A %% B` as a numeric matrix.

See Also

[ggml_crossprod](#), [ggml_tcrossprod](#), [as_gpu_matrix](#)

Examples

```
A <- matrix(rnorm(4), 2); B <- matrix(rnorm(4), 2)
ggml_matmul(A, B, device = "cpu")
```

<code>ggml_matmul_f64</code>	<i>GPU double-precision matrix multiply</i>
------------------------------	---

Description

Computes `A %% B` on the Vulkan GPU in **full double precision** (fp64), returning an ordinary R matrix, with a transparent CPU fallback. Unlike [ggml_matmul](#) (which is f32/f16 and approximate), this matches R's `%%` to machine precision ($\sim 1e^{-15}$), because the kernel accumulates in double throughout.

Usage

```
ggml_matmul_f64(A, B, device = "auto")
```

Arguments

- A, B Numeric matrices with `ncol(A) == nrow(B)`.
- device "auto" (default: GPU when present, fp64-capable and the multiply is large enough to amortise the transfer, else CPU), "gpu" (force GPU, still falls back to CPU if fp64 is unavailable), or "cpu".

Details

fp64 on the GPU is only worthwhile on hardware with fast double throughput: data-centre cards (NVIDIA Tesla P100/V100, AMD Instinct) run fp64 at 1:2 of fp32, whereas consumer GPUs cripple it (RDNA ~1:16, GeForce ~1:64), where the CPU (multithreaded double BLAS) is usually faster. Use this when you need bit-accurate double results on a capable GPU, or for numerically sensitive work where the f32 path's error is unacceptable; otherwise prefer `ggml_matmul`. Requires the `shaderFloat64` device feature — without it (or without a GPU), the multiply silently falls back to the CPU.

Value

The product `A %*% B` as a numeric matrix, accurate to double precision.

See Also

[ggml_matmul](#) (faster f32/f16 approximate path).

Examples

```
A <- matrix(rnorm(4), 2); B <- matrix(rnorm(4), 2)
ggml_matmul_f64(A, B, device = "cpu")
```

ggml_matrix-class	<i>A GPU-backed matrix wrapper</i>
-------------------	------------------------------------

Description

An S4 class holding an ordinary numeric matrix, tagged so that `%*%`, `crossprod()` and `tcrossprod()` run on the GPU (see [ggml_matmul](#)). Wrap a matrix with [as_gpu_matrix](#).

Usage

```
## S4 method for signature 'ggml_matrix,ANY'
x %*% y

## S4 method for signature 'ANY,ggml_matrix'
x %*% y

## S4 method for signature 'ggml_matrix,ggml_matrix'
x %*% y

## S4 method for signature 'ggml_matrix'
crossprod(x, y)

## S4 method for signature 'ggml_matrix'
tcrossprod(x, y)
```

```
## S4 method for signature 'ggml_matrix'
dim(x)

## S4 method for signature 'ggml_matrix'
show(object)
```

Arguments

- x, y Operands; at least one is a ggml_matrix (the other is coerced with as.matrix). For crossprod/tcrossprod, y may be missing.
- object A ggml_matrix (for the show method).

Value

%%, crossprod and tcrossprod return a plain numeric matrix (the GPU result); dim an integer vector.

Slots

- data The underlying numeric matrix.
- device Dispatch preference: "auto", "gpu" or "cpu".
- prec Precision of the GPU path: "f32" or "f16".

ggml_mean	<i>Mean (Graph)</i>
-----------	---------------------

Description

Creates a graph node that computes the mean of all elements.

Usage

```
ggml_mean(ctx, a)
```

Arguments

- ctx GGML context
- a Input tensor

Value

Scalar tensor with the mean

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(2, 4, 6, 8, 10))
result <- ggml_mean(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # 6
ggml_free(ctx)
```

ggml_model

*Create a Functional Model***Description**

Assembles a `ggml_functional_model` from symbolic input and output nodes produced by `ggml_input()` and `ggml_layer_*` calls.

Usage

```
ggml_model(inputs, outputs)
```

Arguments

`inputs` A `ggml_tensor_node` or a list of them (model inputs).
`outputs` A `ggml_tensor_node` or a list of them (model outputs).

Value

A `ggml_functional_model` object.

Examples

```
x <- ggml_input(shape = 64L)
out <- x |> ggml_layer_dense(10, activation = "softmax")
model <- ggml_model(inputs = x, outputs = out)
```

ggml_model_backend	<i>Backend a fitted ggml model actually ran on</i>
--------------------	--

Description

Reports the backend the model was *actually* compiled onto, making a silent backend = "auto" fallback to CPU (when no GPU is available) inspectable. Works on a raw sequential/functional model or a fitted parsnip engine object.

Usage

```
ggml_model_backend(object, verbose = FALSE)
```

Arguments

object	A compiled/fitted ggml_sequential_model, ggml_functional_model or ggmlr_parsnip_model.
verbose	If FALSE (default) returns a single string ("vulkan" or "cpu"). If TRUE returns a list with details (see Value).

Value

If verbose = FALSE, a length-1 character: the backend in use ("vulkan" or "cpu"). If verbose = TRUE, a list with: requested (what was asked: "auto"/"cpu"/"vulkan"), used ("vulkan"/"cpu"), device (GPU device description, or "cpu") and fallback (logical: TRUE when a non-CPU backend was requested but CPU was used instead).

See Also

[ggml_training_history](#)

ggml_model_sequential	<i>Create a Sequential Neural Network Model</i>
-----------------------	---

Description

Creates an empty sequential model that layers can be added to using pipe (|>) operators.

Usage

```
ggml_model_sequential()
```

Value

A ggml_sequential_model object

Examples

```
## Not run:
model <- ggml_model_sequential() |>
  ggml_layer_conv_2d(32, c(3,3), activation = "relu",
    input_shape = c(28, 28, 1)) |>
  ggml_layer_max_pooling_2d(c(2,2)) |>
  ggml_layer_flatten() |>
  ggml_layer_dense(128, activation = "relu") |>
  ggml_layer_dense(10, activation = "softmax")

## End(Not run)
```

ggml_mul

*Multiply tensors***Description**

Creates a graph node for element-wise multiplication.

Usage

```
ggml_mul(ctx, a, b)
```

```
ggml_mul(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor
b	Second tensor (same shape as a)

Value

Tensor representing the multiplication operation

Tensor representing the multiplication operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(2, 2, 2, 2, 2))
result <- ggml_mul(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
ggml_set_f32(b, c(2, 2, 2, 2, 2))
result <- ggml_mul(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_mul_inplace	<i>Element-wise Multiplication In-place (Graph)</i>
------------------	---

Description

Creates a graph node for in-place element-wise multiplication. Result is stored in tensor a, saving memory allocation.

Usage

```
ggml_mul_inplace(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor (will be modified in-place)
b	Second tensor (same shape as a)

Value

View of tensor a with the multiplication result

ggml_mul_mat	<i>Matrix Multiplication (Graph)</i>
--------------	--------------------------------------

Description

Creates a graph node for matrix multiplication. CRITICAL for LLM operations. For matrices A (m x n) and B (n x p), computes C = A * B (m x p).

Usage

```
ggml_mul_mat(ctx, a, b)
```

Arguments

ctx	GGML context
a	First matrix tensor
b	Second matrix tensor

Value

Tensor representing the matrix multiplication

Examples

```
ctx <- ggml_init(1024 * 1024)
A <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 3)
B <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 4, 2)
ggml_set_f32(A, 1:12)
ggml_set_f32(B, 1:8)
C <- ggml_mul_mat(ctx, A, B)
graph <- ggml_build_forward_expand(ctx, C)
ggml_graph_compute(ctx, graph)
ggml_get_f32(C)
ggml_free(ctx)
```

ggml_mul_mat_id	<i>Matrix Multiplication with Expert Selection (Graph)</i>
-----------------	--

Description

Indirect matrix multiplication for Mixture of Experts architectures. Selects expert weights based on indices and performs batched matmul.

Usage

```
ggml_mul_mat_id(ctx, as, b, ids)
```

Arguments

ctx	GGML context
as	Stacked expert weight matrices [n_embd, n_ff, n_experts]
b	Input tensor
ids	Expert selection indices tensor (I32)

Value

Output tensor after expert-selected matrix multiplication

Examples

```
ctx <- ggml_init(64 * 1024 * 1024)
# 4 experts, each with 8x16 weights (small for example)
experts <- ggml_new_tensor_3d(ctx, GGML_TYPE_F32, 8, 16, 4)
ggml_set_f32(experts, rnorm(8 * 16 * 4))
input <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 8, 2)
ggml_set_f32(input, rnorm(16))
# Select expert 0 for token 0, expert 2 for token 1
ids <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 2)
ggml_set_i32(ids, c(0L, 2L))
output <- ggml_mul_mat_id(ctx, experts, input, ids)
graph <- ggml_build_forward_expand(ctx, output)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_nbytes	<i>Get number of bytes</i>
-------------	----------------------------

Description

Get number of bytes
Get Number of Bytes

Usage

```
ggml_nbytes(tensor)

ggml_nbytes(tensor)
```

Arguments

tensor Tensor

Value

Integer number of bytes
Integer number of bytes

Examples

```
ctx <- ggml_init(1024 * 1024)
tensor <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_nbytes(tensor)
ggml_free(ctx)

ctx <- ggml_init(1024 * 1024)
```

```
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_nbytes(t)
ggml_free(ctx)
```

ggml_neg	<i>Negation (Graph)</i>
----------	-------------------------

Description

Creates a graph node for element-wise negation: -x

Usage

```
ggml_neg(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the negation operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, -2, 3, -4))
result <- ggml_neg(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [-1, 2, -3, 4]
ggml_free(ctx)
```

ggml_neg_inplace	<i>Negation In-place (Graph)</i>
------------------	----------------------------------

Description

Creates a graph node for in-place element-wise negation: -x

Usage

```
ggml_neg_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with negated values

ggml_nelements	<i>Get number of elements</i>
----------------	-------------------------------

Description

Get number of elements
Get Number of Elements

Usage

```
ggml_nelements(tensor)  
  
ggml_nelements(tensor)
```

Arguments

tensor	Tensor
--------	--------

Value

Integer number of elements
Integer number of elements

Examples

```
ctx <- ggml_init(1024 * 1024)
tensor <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_nelements(tensor)
ggml_free(ctx)
```

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_nelements(t)
ggml_free(ctx)
```

`ggml_new_f32`*Create Scalar F32 Tensor*

Description

Creates a 1-element tensor containing a single float value. Useful for scalar operations, learning rates, and other scalar floats.

Usage

```
ggml_new_f32(ctx, value)
```

Arguments

<code>ctx</code>	GGML context
<code>value</code>	Numeric value

Value

Tensor pointer (1-element F32 tensor)

Examples

```
ctx <- ggml_init(1024 * 1024)
scalar <- ggml_new_f32(ctx, 3.14)
ggml_get_f32(scalar)
ggml_free(ctx)
```

ggml_new_i32	Create Scalar I32 Tensor
--------------	--------------------------

Description

Creates a 1-element tensor containing a single integer value. Useful for indices, counters, and other scalar integer operations.

Usage

```
ggml_new_i32(ctx, value)
```

Arguments

ctx	GGML context
value	Integer value

Value

Tensor pointer (1-element I32 tensor)

Examples

```
ctx <- ggml_init(1024 * 1024)
scalar <- ggml_new_i32(ctx, 42)
ggml_get_i32(scalar)
ggml_free(ctx)
```

ggml_new_tensor	Create Tensor with Arbitrary Dimensions
-----------------	---

Description

Generic tensor constructor for creating tensors with 1-4 dimensions. This is more flexible than the ggml_new_tensor_Nd functions.

Usage

```
ggml_new_tensor(ctx, type = GGML_TYPE_F32, n_dims, ne)
```

Arguments

ctx	GGML context
type	Data type (GGML_TYPE_F32, etc.)
n_dims	Number of dimensions (1-4)
ne	Numeric vector of dimension sizes

Value

Tensor pointer

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor(ctx, GGML_TYPE_F32, 3, c(10, 20, 30))
ggml_nelements(t)
ggml_free(ctx)
```

ggml_new_tensor_1d	Create 1D tensor
--------------------	------------------

Description

Create 1D tensor
Create 1D Tensor

Usage

```
ggml_new_tensor_1d(ctx, type = GGML_TYPE_F32, ne0)

ggml_new_tensor_1d(ctx, type = GGML_TYPE_F32, ne0)
```

Arguments

ctx	GGML context
type	Data type
ne0	Size

Value

Tensor pointer
Tensor pointer

Examples

```
ctx <- ggml_init(1024 * 1024)
tensor <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_nelements(tensor)
ggml_free(ctx)
```

ggml_new_tensor_2d	Create 2D tensor
--------------------	------------------

Description

Create 2D tensor
Create 2D Tensor

Usage

```
ggml_new_tensor_2d(ctx, type = GGML_TYPE_F32, ne0, ne1)  
  
ggml_new_tensor_2d(ctx, type = GGML_TYPE_F32, ne0, ne1)
```

Arguments

ctx	GGML context
type	Data type
ne0	Rows
ne1	Columns

Value

Tensor pointer
Tensor pointer

Examples

```
ctx <- ggml_init(1024 * 1024)  
tensor <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)  
ggml_nelements(tensor)  
ggml_free(ctx)
```

ggml_new_tensor_3d	Create 3D Tensor
--------------------	------------------

Description

Create 3D Tensor

Usage

```
ggml_new_tensor_3d(ctx, type = GGML_TYPE_F32, ne0, ne1, ne2)
```

Arguments

ctx	GGML context
type	Data type (default GGML_TYPE_F32)
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2

Value

Tensor pointer

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_3d(ctx, GGML_TYPE_F32, 10, 20, 30)
ggml_nelements(t)
ggml_free(ctx)
```

ggml_new_tensor_4d	Create 4D Tensor
--------------------	------------------

Description

Create 4D Tensor

Usage

```
ggml_new_tensor_4d(ctx, type = GGML_TYPE_F32, ne0, ne1, ne2, ne3)
```

Arguments

ctx	GGML context
type	Data type (default GGML_TYPE_F32)
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2
ne3	Size of dimension 3

Value

Tensor pointer

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, 8, 8, 3, 2)
ggml_nelements(t)
ggml_free(ctx)
```

ggml_norm	<i>Layer Normalization (Graph)</i>
-----------	------------------------------------

Description

Creates a graph node for layer normalization. Normalizes input to zero mean and unit variance.

Usage

```
ggml_norm(ctx, a, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor
eps	Epsilon value for numerical stability (default: 1e-5)

Value

Tensor representing the layer normalization operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_norm(ctx, a, eps = 1e-5)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result) # Normalized values
ggml_free(ctx)
```

ggml_norm_inplace	<i>Layer Normalization In-place (Graph)</i>
-------------------	---

Description

Creates a graph node for in-place layer normalization. Returns a view of the input tensor.

Usage

```
ggml_norm_inplace(ctx, a, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
eps	Epsilon value for numerical stability (default: 1e-5)

Value

View of input tensor with layer normalization applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_norm_inplace(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_nrows	<i>Get Number of Rows</i>
------------	---------------------------

Description

Returns the number of rows in a tensor (product of all dimensions except ne[0]).

Usage

```
ggml_nrows(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Number of rows

ggml_n_dims	<i>Get Number of Dimensions</i>
-------------	---------------------------------

Description

Returns the number of dimensions of a tensor

Usage

```
ggml_n_dims(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Integer number of dimensions (1-4)

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_n_dims(t)
ggml_free(ctx)
```

ggml_ops_registry	<i>Supported single-cell operations</i>
-------------------	---

Description

Returns the registry of operations the single-cell adapter can dispatch. Use this to check capabilities (and required parameters) before building a [ggml_task](#) — capability is declared, never discovered at runtime.

Usage

```
ggml_ops_registry(op = NULL)
```

Arguments

op	Optional operation name. If supplied, returns that single entry (or NULL if unknown); otherwise a named list of all entries.
----	--

ggml_opt_context_optimizer_type
<i>Get optimizer type from context</i>

Description

Get optimizer type from context

Usage

ggml_opt_context_optimizer_type(opt_ctx)

Arguments

opt_ctx External pointer to optimizer context

Value

Integer optimizer type constant

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adam\(\)](#), [ggml_opt_optimizer_type_nesterov\(\)](#), [ggml_opt_optimizer_type_rmsprop\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_data	<i>Get data tensor from dataset</i>
-----------------------	-------------------------------------

Description

Returns the underlying data tensor with shape [ne_datapoint, ndata].

Usage

ggml_opt_dataset_data(dataset)

Arguments

dataset External pointer to dataset

Value

External pointer to data tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_free *Free optimization dataset*

Description

Releases memory associated with a dataset.

Usage

```
ggml_opt_dataset_free(dataset)
```

Arguments

dataset External pointer to dataset

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

```
ggml_opt_labels(), ggml_opt_loss(), ggml_opt_loss_type_cross_entropy(), ggml_opt_loss_type_mean(),
ggml_opt_loss_type_mse(), ggml_opt_loss_type_sum(), ggml_opt_loss_type_weighted_mse(),
ggml_opt_ncorrect(), ggml_opt_optimizer_name(), ggml_opt_optimizer_type_adamw(), ggml_opt_optimizer_type_adagrad(),
ggml_opt_outputs(), ggml_opt_pred(), ggml_opt_prepare_alloc(), ggml_opt_reset(), ggml_opt_result_accuracy(),
ggml_opt_result_free(), ggml_opt_result_init(), ggml_opt_result_loss(), ggml_opt_result_ndata(),
ggml_opt_result_pred(), ggml_opt_result_reset(), ggml_opt_set_lr(), ggml_opt_static_graphs()
```

ggml_opt_dataset_get_batch

Get batch from dataset

Description

Copies a batch of data and labels to the provided tensors.

Usage

```
ggml_opt_dataset_get_batch(dataset, data_batch, labels_batch = NULL, ibatch)
```

Arguments

dataset	External pointer to dataset
data_batch	Tensor to receive data batch
labels_batch	Tensor to receive labels batch (can be NULL)
ibatch	Batch index

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_init *Create a new optimization dataset*

Description

Creates a dataset for training with specified data and label types.

Usage

```
ggml_opt_dataset_init(
    type_data,
    type_label,
    ne_datapoint,
    ne_label,
    ndata,
    ndata_shard = 1
)
```

Arguments

type_data	GGML type for data tensor (e.g., GGML_TYPE_F32)
type_label	GGML type for label tensor (e.g., GGML_TYPE_F32)
ne_datapoint	Number of elements per datapoint
ne_label	Number of elements per label (0 if no labels)
ndata	Total number of datapoints
ndata_shard	Shard size for shuffling (1 is fine for most cases)

Value

External pointer to dataset

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_labels

Get labels tensor from dataset

Description

Returns the underlying labels tensor with shape [ne_label, ndata].

Usage

```
ggml_opt_dataset_labels(dataset)
```

Arguments

dataset External pointer to dataset

Value

External pointer to labels tensor, or NULL if no labels

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_ndata

Get number of datapoints in dataset

Description

Get number of datapoints in dataset

Usage

```
ggml_opt_dataset_ndata(dataset)
```

Arguments

dataset External pointer to dataset

Value

Number of datapoints

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_shuffle
<i>Shuffle dataset</i>

Description

Shuffles the dataset using the RNG from the optimizer context.

Usage

ggml_opt_dataset_shuffle(opt_ctx, dataset, idata = -1)

Arguments

opt_ctx External pointer to optimizer context
dataset External pointer to dataset
idata Number of datapoints to shuffle (-1 for all)

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_parameters\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_dataset_weights

Get dataset per-datapoint weights tensor

Description

Returns the (lazily allocated) per-datapoint weights tensor with shape [1, ndata]. The first call allocates it; fill it via [ggml_backend_tensor_set_data\(\)](#). Used together with [ggml_opt_loss_type_weighted_mse](#).

Usage

```
ggml_opt_dataset_weights(dataset)
```

Arguments

dataset	External pointer to dataset
---------	-----------------------------

Value

External pointer to weights tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_default_parameters\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_default_params	<i>Get default optimizer parameters</i>
-------------------------	---

Description

Returns a list with default optimization parameters.

Usage

ggml_opt_default_params(sched, loss_type)

Arguments

- | | |
|-----------|--------------------|
| sched | Backend scheduler |
| loss_type | Loss type constant |

Value

List with loss_type, build_type, opt_period, optimizer

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_epoch	<i>Run one training epoch</i>
----------------	-------------------------------

Description

Performs training on the front portion of the dataset and evaluation on the back portion. This gives more control than ggml_opt_fit.

Usage

```
ggml_opt_epoch(
  opt_ctx,
  dataset,
  result_train = NULL,
  result_eval = NULL,
  idata_split,
  callback_train = TRUE,
  callback_eval = TRUE
)
```

Arguments

<code>opt_ctx</code>	External pointer to optimizer context
<code>dataset</code>	External pointer to dataset
<code>result_train</code>	Result object to accumulate training stats (or NULL)
<code>result_eval</code>	Result object to accumulate evaluation stats (or NULL)
<code>idata_split</code>	Data index at which to split training and evaluation
<code>callback_train</code>	Callback for training: TRUE for progress bar, FALSE for none, or a function(train, ibatch, ibatch_max, t_start_us, result)
<code>callback_eval</code>	Callback for evaluation: TRUE for progress bar, FALSE for none, or a function(train, ibatch, ibatch_max, t_start_us, result)

Value

NULL invisibly

See Also

Other optimization: `ggml_fit_opt()`, `ggml_opt_alloc()`, `ggml_opt_context_optimizer_type()`, `ggml_opt_dataset_data()`, `ggml_opt_dataset_free()`, `ggml_opt_dataset_get_batch()`, `ggml_opt_dataset_init()`, `ggml_opt_dataset_labels()`, `ggml_opt_dataset_ndata()`, `ggml_opt_dataset_shuffle()`, `ggml_opt_dataset_weight()`, `ggml_opt_default_params()`, `ggml_opt_eval()`, `ggml_opt_fit()`, `ggml_opt_free()`, `ggml_opt_get_lr()`, `ggml_opt_grad_acc()`, `ggml_opt_init()`, `ggml_opt_init_for_fit()`, `ggml_opt_inputs()`, `ggml_opt_labels()`, `ggml_opt_loss()`, `ggml_opt_loss_type_cross_entropy()`, `ggml_opt_loss_type_mean()`, `ggml_opt_loss_type_mse()`, `ggml_opt_loss_type_sum()`, `ggml_opt_loss_type_weighted_mse()`, `ggml_opt_ncorrect()`, `ggml_opt_optimizer_name()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_sgd()`, `ggml_opt_outputs()`, `ggml_opt_pred()`, `ggml_opt_prepare_alloc()`, `ggml_opt_reset()`, `ggml_opt_result_accuracy()`, `ggml_opt_result_free()`, `ggml_opt_result_init()`, `ggml_opt_result_loss()`, `ggml_opt_result_ndata()`, `ggml_opt_result_pred()`, `ggml_opt_result_reset()`, `ggml_opt_set_lr()`, `ggml_opt_static_graphs()`

Examples

```
# Requires full optimizer setup - see ggml_opt_fit() for simpler API
if (FALSE) {
  result_train <- ggml_opt_result_init()
  result_eval <- ggml_opt_result_init()
```



```
ggml_opt_epoch(opt_ctx, dataset, result_train, result_eval,
               idata_split = 900, callback_train = TRUE)
ggml_opt_result_free(result_train)
ggml_opt_result_free(result_eval)
}
```

ggml_opt_eval	<i>Evaluate model</i>
---------------	-----------------------

Description

Performs forward pass, optionally increments result, and does backward pass if allocated.

Usage

```
ggml_opt_eval(opt_ctx, result = NULL)
```

Arguments

- opt_ctx External pointer to optimizer context
- result External pointer to result object (optional)

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_fit

*Fit model to dataset***Description**

High-level function to train a model on a dataset. This is the recommended way to train models.

Usage

```
ggml_opt_fit(
  sched,
  ctx_compute,
  inputs,
  outputs,
  dataset,
  loss_type = ggml_opt_loss_type_mse(),
  optimizer = ggml_opt_optimizer_type_adamw(),
  nepoch = 1,
  nbatch_logical = 32,
  val_split = 0,
  silent = FALSE
)
```

Arguments

sched	Backend scheduler
ctx_compute	Compute context (for temporary tensors)
inputs	Input tensor with shape [ne_datapoint, batch_size]
outputs	Output tensor with shape [ne_label, batch_size]
dataset	Dataset created with ggml_opt_dataset_init
loss_type	Loss type (default: MSE)
optimizer	Optimizer type (default: AdamW)
nepoch	Number of epochs
nbatch_logical	Logical batch size (for gradient accumulation)
val_split	Fraction of data for validation (0.0 to 1.0)
silent	Whether to suppress progress output

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

Examples

```
# Full training requires building a computation graph
# See package vignettes for complete examples
if (FALSE) {
  cpu <- ggml_backend_cpu_init()
  sched <- ggml_backend_sched_new(list(cpu))
  dataset <- ggml_opt_dataset_init(GGML_TYPE_F32, GGML_TYPE_F32, 10, 1, 1000)
  # ... build model graph with ctx_compute, inputs, outputs ...
  ggml_opt_fit(sched, ctx_compute, inputs, outputs, dataset,
              nepoch = 10, val_split = 0.1)
  ggml_opt_dataset_free(dataset)
  ggml_backend_sched_free(sched)
  ggml_backend_free(cpu)
}
```

ggml_opt_free	<i>Free optimizer context</i>
---------------	-------------------------------

Description

Releases memory associated with an optimizer context.

Usage

```
ggml_opt_free(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_get_lr	<i>Get current learning rate from optimizer context</i>
-----------------	---

Description

Get current learning rate from optimizer context

Usage

`ggml_opt_get_lr(lr_ud)`

Arguments

`lr_ud` LR userdata pointer (from `'ggml_opt_init_for_fit()$lr_ud'`)

Value

Named numeric vector with 'adamw' and 'sgd' LR values

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_grad_acc	<i>Get gradient accumulator for a tensor</i>
-------------------	--

Description

Returns the gradient accumulator tensor for a node from the forward graph.

Usage

```
ggml_opt_grad_acc(opt_ctx, node)
```

Arguments

opt_ctx	External pointer to optimizer context
node	External pointer to tensor node

Value

External pointer to gradient accumulator tensor, or NULL if not found

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_init	<i>Initialize optimizer context</i>
---------------	-------------------------------------

Description

Creates a new optimizer context for training.

ggml_opt_init_for_fit *Initialize optimizer context for R-side epoch loop*

Description

Returns a list with ‘opt_ctx’ and ‘lr_ud’ (learning rate userdata pointer). Use ‘ggml_opt_set_lr()’ to update LR between epochs. The optimizer state (momentum) is preserved across epochs.

Usage

```
ggml_opt_init_for_fit(
    sched,
    loss_type,
    optimizer = ggml_opt_optimizer_type_adamw(),
    opt_period = 1L,
    ctx_compute = NULL,
    inputs = NULL,
    outputs = NULL
)
```

Arguments

sched	Backend scheduler
loss_type	Loss type constant
optimizer	Optimizer type constant
opt_period	Gradient accumulation period
ctx_compute	Compute context (for static graphs)
inputs	Input tensor (for static graphs)
outputs	Output tensor (for static graphs)

Value

List with elements ‘opt_ctx’ and ‘lr_ud’

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_inputs	<i>Get inputs tensor from optimizer context</i>
-----------------	---

Description

Get inputs tensor from optimizer context

Usage

```
ggml_opt_inputs(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

External pointer to inputs tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_labels	<i>Get labels tensor from optimizer context</i>
-----------------	---

Description

Get labels tensor from optimizer context

Usage

```
ggml_opt_labels(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

External pointer to labels tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_loss

Get loss tensor from optimizer context

Description

Get loss tensor from optimizer context

Usage

```
ggml_opt_loss(opt_ctx)
```

Arguments

opt_ctx	External pointer to optimizer context
---------	---------------------------------------

Value

External pointer to loss tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#)

```
ggml_opt_result_free(), ggml_opt_result_init(), ggml_opt_result_loss(), ggml_opt_result_ndata(),
ggml_opt_result_pred(), ggml_opt_result_reset(), ggml_opt_set_lr(), ggml_opt_static_graphs()
```

```
ggml_opt_loss_type_cross_entropy
```

Loss type: Cross Entropy

Description

Returns the constant for cross entropy loss type. Use for classification tasks.

Usage

```
ggml_opt_loss_type_cross_entropy()
```

Value

Integer constant for cross entropy loss

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

```
ggml_opt_loss_type_mean
```

Loss type: Mean

Description

Returns the constant for mean loss type. Custom loss - reduces outputs to mean value.

Usage

```
ggml_opt_loss_type_mean()
```

Value

Integer constant for mean loss

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_loss_type_mse

Loss type: Mean Squared Error

Description

Returns the constant for MSE loss type. Use for regression tasks.

Usage

`ggml_opt_loss_type_mse()`

Value

Integer constant for MSE loss

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_loss_type_sum

Loss type: Sum

Description

Returns the constant for sum loss type. Custom loss - reduces outputs to sum value.

Usage

```
ggml_opt_loss_type_sum()
```

Value

Integer constant for sum loss

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weights\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_loss_type_weighted_mse

Loss type: Weighted Mean Squared Error

Description

Returns the constant for per-datapoint weighted MSE loss type. Computes $\sum(w * (\text{pred} - y)^2) / \text{nelements}$, where w is a per-sample weight supplied via [ggml_opt_dataset_weights](#).

Usage

```
ggml_opt_loss_type_weighted_mse()
```

Value

Integer constant for weighted MSE loss

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_ncorrect	<i>Get number of correct predictions tensor</i>
-------------------	---

Description

Get number of correct predictions tensor

Usage

```
ggml_opt_ncorrect(opt_ctx)
```

Arguments

opt_ctx	External pointer to optimizer context
---------	---------------------------------------

Value

External pointer to ncorrect tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_optimizer_name
<i>Get optimizer name</i>

Description

Get optimizer name

Usage

ggml_opt_optimizer_name(optimizer_type)

Arguments

optimizer_type Integer optimizer type constant

Value

Character string with optimizer name

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_optimizer_type_adamw
<i>Optimizer type: AdamW</i>

Description

Returns the constant for AdamW optimizer. Adam with weight decay - recommended for most tasks.

Usage

ggml_opt_optimizer_type_adamw()

Value

Integer constant for AdamW optimizer

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_optimizer_type_sgd

Optimizer type: SGD

Description

Returns the constant for SGD optimizer. Stochastic gradient descent - simpler but may require tuning.

Usage

```
ggml_opt_optimizer_type_sgd()
```

Value

Integer constant for SGD optimizer

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_outputs	<i>Get outputs tensor from optimizer context</i>
------------------	--

Description

Get outputs tensor from optimizer context

Usage

```
ggml_opt_outputs(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

External pointer to outputs tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_pred	<i>Get predictions tensor from optimizer context</i>
---------------	--

Description

Get predictions tensor from optimizer context

Usage

```
ggml_opt_pred(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

External pointer to predictions tensor

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_prepare_alloc	<i>Prepare allocation for non-static graphs</i>
------------------------	---

Description

Must be called before `ggml_opt_alloc` when not using static graphs. Sets up the optimizer context with the computation graph and input/output tensors.

Usage

`ggml_opt_prepare_alloc(opt_ctx, ctx_compute, graph, inputs, outputs)`

Arguments

<code>opt_ctx</code>	External pointer to optimizer context
<code>ctx_compute</code>	Compute context for temporary tensors
<code>graph</code>	Computation graph (from <code>ggml_build_forward_expand</code>)
<code>inputs</code>	Input tensor
<code>outputs</code>	Output tensor

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_optimizer_type_adam\(\)](#), [ggml_opt_optimizer_type_rmsprop\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_reset

*Reset optimizer context***Description**

Resets gradients to zero, initializes loss, and optionally resets optimizer state.

Usage

```
ggml_opt_reset(opt_ctx, optimizer = FALSE)
```

Arguments

opt_ctx	External pointer to optimizer context
optimizer	Whether to also reset optimizer state (momentum, etc.)

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_optimizer_type_adam\(\)](#), [ggml_opt_optimizer_type_rmsprop\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_result_accuracy *Get accuracy from result*

Description

Get accuracy from result

Usage

```
ggml_opt_result_accuracy(result)
```

Arguments

result	External pointer to result object
--------	-----------------------------------

Value

Named numeric vector with 'accuracy' and 'uncertainty'

See Also

Other optimization: `ggml_fit_opt()`, `ggml_opt_alloc()`, `ggml_opt_context_optimizer_type()`,
`ggml_opt_dataset_data()`, `ggml_opt_dataset_free()`, `ggml_opt_dataset_get_batch()`, `ggml_opt_dataset_init()`,
`ggml_opt_dataset_labels()`, `ggml_opt_dataset_ndata()`, `ggml_opt_dataset_shuffle()`, `ggml_opt_dataset_weight`,
`ggml_opt_default_params()`, `ggml_opt_epoch()`, `ggml_opt_eval()`, `ggml_opt_fit()`, `ggml_opt_free()`,
`ggml_opt_get_lr()`, `ggml_opt_grad_acc()`, `ggml_opt_init()`, `ggml_opt_init_for_fit()`, `ggml_opt_inputs()`,
`ggml_opt_labels()`, `ggml_opt_loss()`, `ggml_opt_loss_type_cross_entropy()`, `ggml_opt_loss_type_mean()`,
`ggml_opt_loss_type_mse()`, `ggml_opt_loss_type_sum()`, `ggml_opt_loss_type_weighted_mse()`,
`ggml_opt_ncorrect()`, `ggml_opt_optimizer_name()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type`,
`ggml_opt_outputs()`, `ggml_opt_pred()`, `ggml_opt_prepare_alloc()`, `ggml_opt_reset()`, `ggml_opt_result_free()`,
`ggml_opt_result_init()`, `ggml_opt_result_loss()`, `ggml_opt_result_ndata()`, `ggml_opt_result_pred()`,
`ggml_opt_result_reset()`, `ggml_opt_set_lr()`, `ggml_opt_static_graphs()`

ggml_opt_result_free *Free optimization result*

Description

Free optimization result

Usage

```
ggml_opt_result_free(result)
```

Arguments

result External pointer to result object

Value

NULL invisibly

See Also

Other optimization: `ggml_fit_opt()`, `ggml_opt_alloc()`, `ggml_opt_context_optimizer_type()`, `ggml_opt_dataset_data()`, `ggml_opt_dataset_free()`, `ggml_opt_dataset_get_batch()`, `ggml_opt_dataset_init()`, `ggml_opt_dataset_labels()`, `ggml_opt_dataset_ndata()`, `ggml_opt_dataset_shuffle()`, `ggml_opt_dataset_weight()`, `ggml_opt_default_params()`, `ggml_opt_epoch()`, `ggml_opt_eval()`, `ggml_opt_fit()`, `ggml_opt_free()`, `ggml_opt_get_lr()`, `ggml_opt_grad_acc()`, `ggml_opt_init()`, `ggml_opt_init_for_fit()`, `ggml_opt_inputs()`, `ggml_opt_labels()`, `ggml_opt_loss()`, `ggml_opt_loss_type_cross_entropy()`, `ggml_opt_loss_type_mean()`, `ggml_opt_loss_type_mse()`, `ggml_opt_loss_type_sum()`, `ggml_opt_loss_type_weighted_mse()`, `ggml_opt_ncorrect()`, `ggml_opt_optimizer_name()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_outputs()`, `ggml_opt_pred()`, `ggml_opt_prepare_alloc()`, `ggml_opt_reset()`, `ggml_opt_result_accuracy()`, `ggml_opt_result_init()`, `ggml_opt_result_loss()`, `ggml_opt_result_ndata()`, `ggml_opt_result_pred()`, `ggml_opt_result_reset()`, `ggml_opt_set_lr()`, `ggml_opt_static_graphs()`

`ggml_opt_result_init` *Initialize optimization result*

Description

Creates a new result object to accumulate training statistics.

Usage

```
ggml_opt_result_init()
```

Value

External pointer to result object

See Also

Other optimization: `ggml_fit_opt()`, `ggml_opt_alloc()`, `ggml_opt_context_optimizer_type()`, `ggml_opt_dataset_data()`, `ggml_opt_dataset_free()`, `ggml_opt_dataset_get_batch()`, `ggml_opt_dataset_init()`, `ggml_opt_dataset_labels()`, `ggml_opt_dataset_ndata()`, `ggml_opt_dataset_shuffle()`, `ggml_opt_dataset_weight()`, `ggml_opt_default_params()`, `ggml_opt_epoch()`, `ggml_opt_eval()`, `ggml_opt_fit()`, `ggml_opt_free()`, `ggml_opt_get_lr()`, `ggml_opt_grad_acc()`, `ggml_opt_init()`, `ggml_opt_init_for_fit()`, `ggml_opt_inputs()`, `ggml_opt_labels()`, `ggml_opt_loss()`, `ggml_opt_loss_type_cross_entropy()`, `ggml_opt_loss_type_mean()`, `ggml_opt_loss_type_mse()`, `ggml_opt_loss_type_sum()`, `ggml_opt_loss_type_weighted_mse()`, `ggml_opt_ncorrect()`, `ggml_opt_optimizer_name()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_outputs()`, `ggml_opt_pred()`, `ggml_opt_prepare_alloc()`, `ggml_opt_reset()`, `ggml_opt_result_accuracy()`

```
ggml_opt_result_free(), ggml_opt_result_loss(), ggml_opt_result_ndata(), ggml_opt_result_pred(),
ggml_opt_result_reset(), ggml_opt_set_lr(), ggml_opt_static_graphs()
```

ggml_opt_result_loss *Get loss from result*

Description

Get loss from result

Usage

```
ggml_opt_result_loss(result)
```

Arguments

result External pointer to result object

Value

Named numeric vector with 'loss' and 'uncertainty'

See Also

Other optimization: `ggml_fit_opt()`, `ggml_opt_alloc()`, `ggml_opt_context_optimizer_type()`, `ggml_opt_dataset_data()`, `ggml_opt_dataset_free()`, `ggml_opt_dataset_get_batch()`, `ggml_opt_dataset_init()`, `ggml_opt_dataset_labels()`, `ggml_opt_dataset_ndata()`, `ggml_opt_dataset_shuffle()`, `ggml_opt_dataset_weighted_loss()`, `ggml_opt_default_params()`, `ggml_opt_epoch()`, `ggml_opt_eval()`, `ggml_opt_fit()`, `ggml_opt_free()`, `ggml_opt_get_lr()`, `ggml_opt_grad_acc()`, `ggml_opt_init()`, `ggml_opt_init_for_fit()`, `ggml_opt_inputs()`, `ggml_opt_labels()`, `ggml_opt_loss()`, `ggml_opt_loss_type_cross_entropy()`, `ggml_opt_loss_type_mean()`, `ggml_opt_loss_type_mse()`, `ggml_opt_loss_type_sum()`, `ggml_opt_loss_type_weighted_mse()`, `ggml_opt_ncorrect()`, `ggml_opt_optimizer_name()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_optimizer_type_adamw()`, `ggml_opt_optimizer_type_adagrad()`, `ggml_opt_outputs()`, `ggml_opt_pred()`, `ggml_opt_prepare_alloc()`, `ggml_opt_reset()`, `ggml_opt_result_accuracy()`, `ggml_opt_result_free()`, `ggml_opt_result_init()`, `ggml_opt_result_ndata()`, `ggml_opt_result_pred()`, `ggml_opt_result_reset()`, `ggml_opt_set_lr()`, `ggml_opt_static_graphs()`

ggml_opt_result_ndata *Get number of datapoints from result*

Description

Get number of datapoints from result

Usage

```
ggml_opt_result_ndata(result)
```

Arguments

result External pointer to result object

Value

Number of datapoints processed

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adam\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

`ggml_opt_result_pred` *Get predictions from result*

Description

Returns the predictions as an integer vector. The length equals the number of datapoints processed.

Usage

```
ggml_opt_result_pred(result)
```

Arguments

result External pointer to result object

Value

Integer vector of predictions

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#),

```
ggml_opt_labels(), ggml_opt_loss(), ggml_opt_loss_type_cross_entropy(), ggml_opt_loss_type_mean(),
ggml_opt_loss_type_mse(), ggml_opt_loss_type_sum(), ggml_opt_loss_type_weighted_mse(),
ggml_opt_ncorrect(), ggml_opt_optimizer_name(), ggml_opt_optimizer_type_adamw(), ggml_opt_optimizer_type_adamw(),
ggml_opt_outputs(), ggml_opt_pred(), ggml_opt_prepare_alloc(), ggml_opt_reset(), ggml_opt_result_accuracy(),
ggml_opt_result_free(), ggml_opt_result_init(), ggml_opt_result_loss(), ggml_opt_result_ndata(),
ggml_opt_result_reset(), ggml_opt_set_lr(), ggml_opt_static_graphs()
```

ggml_opt_result_reset *Reset optimization result*

Description

Reset optimization result

Usage

```
ggml_opt_result_reset(result)
```

Arguments

result	External pointer to result object
--------	-----------------------------------

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weighted_loss\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_set_lr\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_set_lr	<i>Set learning rate in optimizer context</i>
-----------------	---

Description

Updates the LR used for subsequent backward passes. Can be called between epochs to implement LR scheduling.

Usage

```
ggml_opt_set_lr(lr_ud, adamw_lr = NA, sgd_lr = NA)
```

Arguments

lr_ud	LR userdata pointer (from 'ggml_opt_init_for_fit()\$lr_ud')
adamw_lr	New AdamW learning rate (NA to keep current)
sgd_lr	New SGD learning rate (NA to keep current)

Value

NULL invisibly

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_sgd\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_static_graphs\(\)](#)

ggml_opt_static_graphs	<i>Check if using static graphs</i>
------------------------	-------------------------------------

Description

Check if using static graphs

Usage

```
ggml_opt_static_graphs(opt_ctx)
```

Arguments

opt_ctx External pointer to optimizer context

Value

Logical indicating if graphs are statically allocated

See Also

Other optimization: [ggml_fit_opt\(\)](#), [ggml_opt_alloc\(\)](#), [ggml_opt_context_optimizer_type\(\)](#), [ggml_opt_dataset_data\(\)](#), [ggml_opt_dataset_free\(\)](#), [ggml_opt_dataset_get_batch\(\)](#), [ggml_opt_dataset_init\(\)](#), [ggml_opt_dataset_labels\(\)](#), [ggml_opt_dataset_ndata\(\)](#), [ggml_opt_dataset_shuffle\(\)](#), [ggml_opt_dataset_weight\(\)](#), [ggml_opt_default_params\(\)](#), [ggml_opt_epoch\(\)](#), [ggml_opt_eval\(\)](#), [ggml_opt_fit\(\)](#), [ggml_opt_free\(\)](#), [ggml_opt_get_lr\(\)](#), [ggml_opt_grad_acc\(\)](#), [ggml_opt_init\(\)](#), [ggml_opt_init_for_fit\(\)](#), [ggml_opt_inputs\(\)](#), [ggml_opt_labels\(\)](#), [ggml_opt_loss\(\)](#), [ggml_opt_loss_type_cross_entropy\(\)](#), [ggml_opt_loss_type_mean\(\)](#), [ggml_opt_loss_type_mse\(\)](#), [ggml_opt_loss_type_sum\(\)](#), [ggml_opt_loss_type_weighted_mse\(\)](#), [ggml_opt_ncorrect\(\)](#), [ggml_opt_optimizer_name\(\)](#), [ggml_opt_optimizer_type_adamw\(\)](#), [ggml_opt_optimizer_type_adagrad\(\)](#), [ggml_opt_outputs\(\)](#), [ggml_opt_pred\(\)](#), [ggml_opt_prepare_alloc\(\)](#), [ggml_opt_reset\(\)](#), [ggml_opt_result_accuracy\(\)](#), [ggml_opt_result_free\(\)](#), [ggml_opt_result_init\(\)](#), [ggml_opt_result_loss\(\)](#), [ggml_opt_result_ndata\(\)](#), [ggml_opt_result_pred\(\)](#), [ggml_opt_result_reset\(\)](#), [ggml_opt_set_lr\(\)](#)

ggml_op_can_inplace	<i>Check if Operation Can Be Done In-place</i>
---------------------	--

Description

Returns whether a GGML operation can reuse memory from its source tensors. This is useful for memory optimization.

Usage

```
ggml_op_can_inplace(op)
```

Arguments

op Operation code (integer)

Value

Logical indicating if operation supports in-place execution

See Also

Other graph: [ggml_graph_view\(\)](#)

Examples

```
# Check if operation code 1 (ADD) can be in-place
can_inplace <- ggml_op_can_inplace(1L)
```

ggml_op_desc	<i>Get Operation Description from Tensor</i>
--------------	--

Description

Returns a description of the operation that produces a tensor.

Usage

```
ggml_op_desc(tensor)
```

Arguments

tensor Tensor pointer

Value

Character string describing the operation

See Also

Other op_info: [ggml_get_unary_op\(\)](#), [ggml_op_name\(\)](#), [ggml_op_symbol\(\)](#), [ggml_unary_op_name\(\)](#)

ggml_op_name	<i>Get Operation Name</i>
--------------	---------------------------

Description

Returns the string name of a GGML operation.

Usage

```
ggml_op_name(op)
```

Arguments

op GGML operation constant

Value

Character string with operation name

See Also

Other op_info: [ggml_get_unary_op\(\)](#), [ggml_op_desc\(\)](#), [ggml_op_symbol\(\)](#), [ggml_unary_op_name\(\)](#)

ggml_op_symbol	<i>Get Operation Symbol</i>
----------------	-----------------------------

Description

Returns the mathematical symbol for a GGML operation.

Usage

ggml_op_symbol(op)

Arguments

op GGML operation constant

Value

Character string with operation symbol

See Also

Other op_info: [ggml_get_unary_op\(\)](#), [ggml_op_desc\(\)](#), [ggml_op_name\(\)](#), [ggml_unary_op_name\(\)](#)

ggml_out_prod	<i>Outer Product (Graph)</i>
---------------	------------------------------

Description

Computes the outer product of two vectors: $C = a * b^T$ For vectors $a[m]$ and $b[n]$, produces matrix $C[m, n]$.

Usage

ggml_out_prod(ctx, a, b)

Arguments

ctx GGML context
a First vector tensor
b Second vector tensor

Value

Matrix tensor representing the outer product

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 3)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3))
ggml_set_f32(b, c(1, 2, 3, 4))
c <- ggml_out_prod(ctx, a, b) # Result: 3x4 matrix
graph <- ggml_build_forward_expand(ctx, c)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_pad	<i>Pad Tensor with Zeros (Graph)</i>
----------	--------------------------------------

Description

Pads tensor dimensions with zeros on the right side. Useful for aligning tensor sizes in attention operations.

Usage

```
ggml_pad(ctx, a, p0 = 0L, p1 = 0L, p2 = 0L, p3 = 0L)
```

Arguments

ctx	GGML context
a	Input tensor to pad
p0	Padding for dimension 0 (default 0)
p1	Padding for dimension 1 (default 0)
p2	Padding for dimension 2 (default 0)
p3	Padding for dimension 3 (default 0)

Value

Padded tensor with shape [ne0+p0, ne1+p1, ne2+p2, ne3+p3]

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 5, 3)
ggml_set_f32(a, 1:15)
# Pad to 8x4
b <- ggml_pad(ctx, a, 3, 1) # Add 3 zeros to dim0, 1 to dim1
graph <- ggml_build_forward_expand(ctx, b)
ggml_graph_compute(ctx, graph)
# Result shape: [8, 4]
ggml_free(ctx)
```

ggml_pad_reflect_1d	<i>Reflective 1D Padding (Graph)</i>
---------------------	--------------------------------------

Description

Pads the first dimension of a tensor using reflection of its values.

Usage

```
ggml_pad_reflect_1d(ctx, a, p0, p1)
```

Arguments

ctx	GGML context
a	Input tensor
p0	Left padding
p1	Right padding

Value

Padded tensor

ggml_permute	<i>Permute Tensor Dimensions (Graph)</i>
--------------	--

Description

Permutes the tensor dimensions according to specified axes. CRITICAL for attention mechanisms in transformers.

Usage

```
ggml_permute(ctx, a, axis0, axis1, axis2, axis3)
```

Arguments

ctx	GGML context
a	Input tensor
axis0	New position for axis 0
axis1	New position for axis 1
axis2	New position for axis 2
axis3	New position for axis 3

Value

Permuted tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Create 4D tensor: (2, 3, 4, 5)
t <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, 2, 3, 4, 5)
# Swap axes 0 and 1: result shape (3, 2, 4, 5)
t_perm <- ggml_permute(ctx, t, 1, 0, 2, 3)
ggml_free(ctx)
```

ggml_pool_1d	<i>1D Pooling (Graph)</i>
--------------	---------------------------

Description

Applies 1D pooling operation for downsampling.

Usage

```
ggml_pool_1d(ctx, a, op, k0, s0 = k0, p0 = 0L)
```

GGML_OP_POOL_MAX

GGML_OP_POOL_AVG

Arguments

ctx	GGML context
a	Input tensor
op	Pool operation constant (see details)
k0	Kernel size (window size)
s0	Stride (default = k0 for non-overlapping windows)
p0	Padding (default 0)

Format

An object of class integer of length 1.

An object of class integer of length 1.

Details

Pool operation constants:

- GGML_OP_POOL_MAX (0): Max pooling - takes maximum value in each window
- GGML_OP_POOL_AVG (1): Average pooling - takes mean of values in each window

Value

Pooled tensor with reduced dimensions

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 8)
ggml_set_f32(a, c(1, 3, 2, 4, 5, 2, 8, 1))

# Max pooling with kernel 2, stride 2
max_pool <- ggml_pool_1d(ctx, a, GGML_OP_POOL_MAX, k0 = 2)
# Result: [3, 4, 5, 8] (max of each pair)

# Average pooling with kernel 2, stride 2
avg_pool <- ggml_pool_1d(ctx, a, GGML_OP_POOL_AVG, k0 = 2)
# Result: [2, 3, 3.5, 4.5] (mean of each pair)

ggml_free(ctx)
```

ggml_pool_2d

2D Pooling (Graph)

Description

Applies 2D pooling operation.

Usage

```
ggml_pool_2d(ctx, a, op, k0, k1, s0 = k0, s1 = k1, p0 = 0, p1 = 0)
```

Arguments

ctx	GGML context
a	Input tensor
op	Pool operation: GGML_OP_POOL_MAX (0) or GGML_OP_POOL_AVG (1)
k0	Kernel size dimension 0
k1	Kernel size dimension 1
s0	Stride dimension 0 (default = k0)
s1	Stride dimension 1 (default = k1)
p0	Padding dimension 0 (default 0)
p1	Padding dimension 1 (default 0)

Value

Pooled tensor

ggml_pop_layer	<i>Remove the Last Layer from a Sequential Model</i>
----------------	--

Description

Removes the last layer from the model. The model must not be compiled.

Usage

```
ggml_pop_layer(model)
```

Arguments

model	A ggml_sequential_model object
-------	--------------------------------

Value

The model with the last layer removed.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(64, activation = "relu") |>
  ggml_layer_dense(10, activation = "softmax")

model <- ggml_pop_layer(model)
length(model$layers) # 1
```

ggml_pp_dp_forward *Pipeline + data parallel forward pass (PP x DP)*

Description

Runs data parallelism over the batch across replicas of a pipeline-parallel model: the batch is split into contiguous shards, one per replica, and each replica runs the whole model as a device-by-layer **pipeline** on its own set of GPUs (via [ggml_pp_forward](#)). This is the pipeline counterpart of [ggml_tp_dp_forward](#) — use it when the model is too large for one card (split it by layers) and you also want to raise throughput by replicating the pipeline across independent GPU sets.

Usage

```
ggml_pp_dp_forward(
    make_stages,
    x,
    replicas,
    out_ncol,
    mem_per_stage = 16L * 1024L * 1024L
)
```

Arguments

make_stages	A function function(devices, m_shard) returning the list of pipeline stage descriptors (see ggml_pp_forward) for a replica whose pipeline runs on devices (an integer GPU-index vector) and whose batch shard has m_shard samples. The stages' in_shape / weight setup must be built for that m_shard.
x	Input activation matrix, M x K (M samples, K features). Its rows (the batch) are split across the replicas.
replicas	A list of integer vectors, each the GPU-index pipeline for one replica, e.g. list(c(0, 1), c(2, 3)) for PP=2 x DP=2.
out_ncol	Number of columns N of the per-sample output (the final stage produces an N x m_shard result per replica; the gathered result is M x N).
mem_per_stage	Bytes of ggml context metadata per stage (see ggml_pp_forward).

Details

On a 4-GPU box the natural layout is **PP=2 x DP=2**: replica A pipelines its layers across GPUs c(0, 1), replica B across c(2, 3), and the two replicas each handle half the batch with no cross-replica traffic.

Value

The M x N result matrix, the batch-concatenation of each replica's pipeline output.

See Also

[ggml_pp_forward](#), [ggml_tp_dp_forward](#).

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 4) {
  K <- 64L
  W1 <- matrix(rnorm(K * K), K); W2 <- matrix(rnorm(K * K), K)
  make_stages <- function(devices, m) {
    mk <- function(dev, Wt, relu) list(
      device = dev, in_shape = c(K, m),
      build = function(ctx, input) {
        w <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, K, K)
        z <- ggml_mul_mat(ctx, w, input)
        list(output = if (relu) ggml_relu(ctx, z) else z,
              set_weights = function() ggml_backend_tensor_set_data(w, as.numeric(Wt)))
      })
    list(mk(devices[1], W1, TRUE), mk(devices[2], W2, FALSE))
  }
  X <- matrix(rnorm(8 * K), nrow = 8) # batch of 8
  Y <- ggml_pp_dp_forward(make_stages, X, replicas = list(c(0, 1), c(2, 3)),
                          out_ncol = K)
}
```

ggml_pp_forward

Pipeline-parallel forward pass across per-device layer stages

Description

Runs a forward pass split BY LAYERS across devices (pipeline parallelism), the complement of [ggml_vulkan_split_mul_mat](#)'s split-by-matrix tensor parallelism. Each stage owns a contiguous block of the model's layers on one GPU; the activation tensor is handed from one stage to the next exactly once per pass (a single cross-device copy), versus TP's per-layer gather. This suits models too large for one card's VRAM: at ~1 GB/s host staging the single handoff costs only ~10-20 ms per pass.

Usage

```
ggml_pp_forward(stages, x, out_shape, mem_per_stage = 16L * 1024L * 1024L)
```

Arguments

stages	A list of stage descriptors (see Details), in pipeline order.
x	A numeric vector/array of input activations for the first stage, laid out in ggml (column-major) order matching stages[[1]]\$in_shape.
out_shape	Integer vector: the ggml ne shape of the final stage's output, used to size the returned vector.

`mem_per_stage` Bytes of ggml context metadata to reserve per stage (default 16 MiB) — raise it for stages with very many ops.

Details

Each stage is a list with:

`device` Physical GPU index (0-based) the stage runs on.

`build` A function `function(ctx, input)` that builds this stage's sub-graph from the input tensor using the `ggml_*` ops and returns either the output tensor, or a list `list(output = <tensor>, set_weights = function() ...)`. The `set_weights` closure (if given) is called AFTER the stage's tensors are allocated on the device, and is where weight tensors created inside `build` get their values (via `ggml_backend_tensor_set_data`).

`in_shape` Integer vector: the ggml ne shape of this stage's input tensor (fastest dim first). For stage 1 this must match the supplied `x`; for later stages it must match the previous stage's output shape.

Value

A numeric vector of the final stage's output activations (ggml order).

See Also

[ggml_tp_dp_forward](#) for the tensor-parallel counterpart.

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  K <- 32L; M <- 4L
  W1 <- matrix(rnorm(K * K), K); W2 <- matrix(rnorm(K * K), K)
  stage <- function(dev, Wt) list(
    device = dev, in_shape = c(K, M),
    build = function(ctx, input) {
      w <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, K, K)
      ggml_mul_mat(ctx, w, input) # (weights set inside build via attr, see vignette)
    })
  # See inst/examples/pp_pipeline.R for a complete runnable stage definition.
}
```

ggml_predict.ggml_functional_model

Get Predictions from a Trained Model

Description

Runs forward pass on input data and returns prediction probabilities (or raw output values for regression). Unlike `ggml_evaluate()`, this does not require labels.

Usage

```
## S3 method for class 'ggml_functional_model'
ggml_predict(model, x, batch_size = 32L, ...)

ggml_predict(model, ...)

## S3 method for class 'ggml_sequential_model'
ggml_predict(model, x, batch_size = 32L, ...)
```

Arguments

model	A trained ggml_sequential_model
x	Input data (matrix or array)
batch_size	Batch size for inference
...	Additional arguments (ignored).

Value

Matrix of predictions with shape [N, output_units]

ggml_predict_classes *Predict Classes from a Trained Model*

Description

Returns predicted class indices (1-based) by applying argmax to the output of ggml_predict().

Usage

```
ggml_predict_classes(model, x, batch_size = 32L)
```

Arguments

model	A trained ggml_sequential_model
x	Input data (matrix or array)
batch_size	Batch size for inference

Value

Integer vector of predicted class indices (1-based)

ggml_print_mem_status *Print Context Memory Status*

Description

Helper to print memory usage information

Usage

```
ggml_print_mem_status(ctx)
```

Arguments

ctx	GGML context
-----	--------------

Value

List with total, used, free memory (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_print_mem_status(ctx)
ggml_free(ctx)
```

ggml_print_objects *Print Objects in Context*

Description

Debug function to print all objects (tensors) in the context

Usage

```
ggml_print_objects(ctx)
```

Arguments

ctx	GGML context
-----	--------------

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_print_objects(ctx)
ggml_free(ctx)
```

ggml_quantize_chunk	<i>Quantize Data Chunk</i>
---------------------	----------------------------

Description

Quantizes a chunk of floating-point data to a lower precision format.

Usage

```
ggml_quantize_chunk(type, src, nrows, n_per_row)
```

Arguments

type	Target GGML type (e.g., GGML_TYPE_Q4_0)
src	Source numeric vector (F32 data)
nrows	Number of rows
n_per_row	Number of elements per row

Value

Raw vector containing quantized data

Examples

```
# Quantize 256 floats to Q8_0 (block size 32)
data <- rnorm(256)
quantized <- ggml_quantize_chunk(GGML_TYPE_Q8_0, data, 1, 256)
ggml_quantize_free() # Clean up
```

ggml_quantize_free	<i>Free Quantization Resources</i>
--------------------	------------------------------------

Description

Frees any memory allocated by quantization. Call at end of program to avoid memory leaks.

Usage

```
ggml_quantize_free()
```

Value

NULL invisibly

ggml_quantize_init	<i>Initialize Quantization Tables</i>
--------------------	---------------------------------------

Description

Initializes quantization tables for a given type. Called automatically by ggml_quantize_chunk, but can be called manually.

Usage

```
ggml_quantize_init(type)
```

Arguments

type	GGML type (e.g., GGML_TYPE_Q4_0)
------	----------------------------------

Value

NULL invisibly

ggml_quantize_requires_imatrix
<i>Check if Quantization Requires Importance Matrix</i>

Description

Some quantization types require an importance matrix for optimal quality.

Usage

ggml_quantize_requires_imatrix(type)

Arguments

type GGML type

Value

TRUE if importance matrix is required

ggml_quant_block_info	<i>Get Quantization Block Info</i>
-----------------------	------------------------------------

Description

Returns information about a quantization type including name, type size, block size, and whether it's quantized.

Usage

ggml_quant_block_info(type)

Arguments

type GGML type constant

Value

List with type_name, type_size, block_size, is_quantized

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

ggml_reglu	<i>ReLU (ReLU Gated Linear Unit) (Graph)</i>
------------	--

Description

Creates a graph node for ReGLU operation. ReGLU uses ReLU as the activation function on the first half.

Usage

```
ggml_reglu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (first dimension must be even)

Details

Formula: $\text{output} = \text{ReLU}(x) * \text{gate}$

Value

Tensor with half the first dimension of input

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 8, 3)
ggml_set_f32(a, rnorm(24))
r <- ggml_reglu(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # Shape: 4x3
ggml_free(ctx)
```

ggml_reglu_split	<i>ReLU Split (Graph)</i>
------------------	---------------------------

Description

Creates a graph node for ReGLU with separate input and gate tensors.

Usage

```
ggml_reglu_split(ctx, a, b)
```

Arguments

ctx	GGML context
a	Input tensor (the values to be gated)
b	Gate tensor (same shape as a)

Details

Formula: $\text{output} = \text{ReLU}(a) * b$

Value

Tensor with same shape as input tensors

ggml_relu	<i>ReLU Activation (Graph)</i>
-----------	--------------------------------

Description

Creates a graph node for ReLU (Rectified Linear Unit) activation: $\max(0, x)$

Usage

```
ggml_relu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the ReLU operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_relu(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_relu_inplace	<i>ReLU Activation In-place (Graph)</i>
-------------------	---

Description

Creates a graph node for in-place ReLU activation: $\max(0, x)$

Usage

```
ggml_relu_inplace(ctx, a)
```

Arguments

- | | |
|-----|--|
| ctx | GGML context |
| a | Input tensor (will be modified in-place) |

Value

View of tensor a with ReLU applied

ggml_repeat	<i>Repeat (Graph)</i>
-------------	-----------------------

Description

Creates a graph node that repeats tensor 'a' to match shape of tensor 'b'.

Usage

```
ggml_repeat(ctx, a, b)
```

Arguments

- | | |
|-----|--------------------------------------|
| ctx | GGML context |
| a | Tensor to repeat |
| b | Target tensor (defines output shape) |

Value

Tensor with repeated values

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 1, 2)
ggml_set_f32(a, c(1, 2))
b <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 3, 2)
result <- ggml_repeat(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [1, 1, 1, 2, 2, 2]
ggml_free(ctx)
```

<code>ggml_repeat_back</code>	<i>Repeat Backward (Graph)</i>
-------------------------------	--------------------------------

Description

Backward pass for repeat operation - sums repetitions back to original shape. Used for gradient computation during training.

Usage

```
ggml_repeat_back(ctx, a, b)
```

Arguments

<code>ctx</code>	GGML context
<code>a</code>	Input tensor (gradients from repeated tensor)
<code>b</code>	Target shape tensor (original tensor before repeat)

Value

Tensor with summed gradients matching shape of b

<code>ggml_reset</code>	<i>Reset GGML Context</i>
-------------------------	---------------------------

Description

Clears all tensor allocations in the context memory pool. The context can be reused without recreating it. This is more efficient than free + init for temporary operations.

Usage

```
ggml_reset(ctx)
```

Arguments

ctx GGML context pointer

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
ggml_reset(ctx)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 200)
ggml_free(ctx)
```

ggml_reshape_1d	<i>Reshape to 1D (Graph)</i>
-----------------	------------------------------

Description

Reshapes tensor to 1D with ne0 elements

Usage

```
ggml_reshape_1d(ctx, a, ne0)
```

Arguments

ctx GGML context
a Input tensor
ne0 Size of dimension 0

Value

Reshaped tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 3, 4)
ggml_set_f32(a, 1:12)
result <- ggml_reshape_1d(ctx, a, 12)
ggml_free(ctx)
```

ggml_reshape_2d	<i>Reshape to 2D (Graph)</i>
-----------------	------------------------------

Description

Reshapes tensor to 2D with shape (ne0, ne1)

Usage

ggml_reshape_2d(ctx, a, ne0, ne1)

Arguments

ctx	GGML context
a	Input tensor
ne0	Size of dimension 0
ne1	Size of dimension 1

Value

Reshaped tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 12)
ggml_set_f32(a, 1:12)
result <- ggml_reshape_2d(ctx, a, 3, 4)
ggml_free(ctx)
```

ggml_reshape_3d	<i>Reshape to 3D (Graph)</i>
-----------------	------------------------------

Description

Reshapes tensor to 3D with shape (ne0, ne1, ne2)

Usage

ggml_reshape_3d(ctx, a, ne0, ne1, ne2)

Arguments

ctx	GGML context
a	Input tensor
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2

Value

Reshaped tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 24)
ggml_set_f32(a, 1:24)
result <- ggml_reshape_3d(ctx, a, 2, 3, 4)
ggml_free(ctx)
```

ggml_reshape_4d	<i>Reshape to 4D (Graph)</i>
-----------------	------------------------------

Description

Reshapes tensor to 4D with shape (ne0, ne1, ne2, ne3)

Usage

```
ggml_reshape_4d(ctx, a, ne0, ne1, ne2, ne3)
```

Arguments

ctx	GGML context
a	Input tensor
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2
ne3	Size of dimension 3

Value

Reshaped tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 120)
ggml_set_f32(a, 1:120)
result <- ggml_reshape_4d(ctx, a, 2, 3, 4, 5)
ggml_free(ctx)
```

ggml_result	Construct a single-cell result
-------------	--------------------------------

Description

A ggml_result is the contract object returned by the dispatch layer and consumed by the injection layer. The embedding is stored cell-by-component (cells in rows), ready to drop into reducedDim() / a Seurat reduction.

Usage

```
ggml_result(embedding, metadata = list(), timings = numeric(0))
```

Arguments

- embedding A numeric matrix, cells in rows, components in columns.
- metadata Named list (e.g. stdev, loadings, backend used).
- timings Named numeric vector of elapsed seconds per stage.

Value

An object of class ggml_result.

ggml_rms_norm	RMS Normalization (Graph)
---------------	---------------------------

Description

Creates a graph node for RMS (Root Mean Square) normalization. Normalizes by $x / \sqrt{\text{mean}(x^2) + \text{eps}}$. CRITICAL for LLaMA models.

Usage

```
ggml_rms_norm(ctx, a, eps = 1e-05)
```


Arguments

ctx	GGML context
a	Input tensor
eps	Epsilon value for numerical stability (default: 1e-5)

Value

Tensor representing the RMS normalization operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_rms_norm(ctx, a, eps = 1e-5)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result)
# sqrt(mean(output^2)) should be ~1
ggml_free(ctx)
```

ggml_rms_norm_back	<i>RMS Norm Backward (Graph)</i>
--------------------	----------------------------------

Description

Creates a graph node for backward pass of RMS normalization. Used in training for computing gradients.

Usage

```
ggml_rms_norm_back(ctx, a, b, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor (x from forward pass)
b	Gradient tensor (dy)
eps	Epsilon for numerical stability (default 1e-5)

Value

Tensor representing the gradient with respect to input

ggml_rms_norm_inplace	<i>RMS Normalization In-place (Graph)</i>
-----------------------	---

Description

Creates a graph node for in-place RMS normalization. Returns a view of the input tensor. CRITICAL for LLaMA models when memory efficiency is important.

Usage

```
ggml_rms_norm_inplace(ctx, a, eps = 1e-05)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
eps	Epsilon value for numerical stability (default: 1e-5)

Value

View of input tensor with RMS normalization applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_rms_norm_inplace(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_roll	<i>Roll (Graph)</i>
-----------	---------------------

Description

Circularly shifts tensor elements along dimensions 0..3.

Usage

```
ggml_roll(ctx, a, shift0 = 0L, shift1 = 0L, shift2 = 0L, shift3 = 0L)
```

Arguments

ctx	GGML context
a	Input tensor
shift0, shift1, shift2, shift3	Shift amount along each dimension

Value

Rolled tensor

ggml_rope	<i>Rotary Position Embedding (Graph)</i>
-----------	--

Description

Creates a graph node for RoPE (Rotary Position Embedding). RoPE is the dominant position encoding method in modern LLMs like LLaMA, Mistral, and many others.

Usage

```
ggml_rope(ctx, a, b, n_dims, mode = 0L)
```

Arguments

ctx	GGML context
a	Input tensor of shape [head_dim, n_head, seq_len, batch]
b	Position tensor (int32) of shape [seq_len] containing position indices
n_dims	Number of dimensions to apply rotation to (usually head_dim)
mode	RoPE mode: GGML_ROPE_TYPE_NORM (0), GGML_ROPE_TYPE_NEOX (2), etc.

Details

RoPE encodes position information by rotating pairs of dimensions in the embedding space. The rotation angle depends on position and dimension index.

Key benefits of RoPE: - Relative position information emerges naturally from rotation - Better extrapolation to longer sequences than absolute embeddings - No additional parameters needed

Value

Tensor with same shape as input, with rotary embeddings applied

Examples

```

ctx <- ggml_init(16 * 1024 * 1024)
# Query tensor: head_dim=8, n_head=4, seq_len=16, batch=1
q <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, 8, 4, 16, 1)
ggml_set_f32(q, rnorm(8 * 4 * 16))
# Position indices
pos <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 16)
ggml_set_i32(pos, 0:15)
# Apply RoPE
q_rope <- ggml_rope(ctx, q, pos, 8, GGML_ROPE_TYPE_NORM)
graph <- ggml_build_forward_expand(ctx, q_rope)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)

```

ggml_rope_ext

*Extended RoPE with Frequency Scaling (Graph)***Description**

Creates a graph node for extended RoPE with frequency scaling parameters. Supports context extension techniques like YaRN, Linear Scaling, etc.

Usage

```

ggml_rope_ext(
  ctx,
  a,
  b,
  c = NULL,
  n_dims,
  mode = 0L,
  n_ctx_orig = 0L,
  freq_base = 10000,
  freq_scale = 1,
  ext_factor = 0,
  attn_factor = 1,
  beta_fast = 32,
  beta_slow = 1
)

```

Arguments

ctx	GGML context
a	Input tensor
b	Position tensor (int32)
c	Optional frequency factors tensor (NULL for default)

n_dims	Number of dimensions to apply rotation to
mode	RoPE mode
n_ctx_orig	Original context length the model was trained on
freq_base	Base frequency for RoPE (default 10000 for most models)
freq_scale	Frequency scale factor (1.0 = no scaling)
ext_factor	YaRN extension factor (0.0 to disable)
attn_factor	Attention scale factor (typically 1.0)
beta_fast	YaRN parameter for fast dimensions
beta_slow	YaRN parameter for slow dimensions

Details

This extended version supports various context extension techniques:

- **Linear Scaling**: Set `freq_scale = original_ctx / new_ctx` - **YaRN**: Set `ext_factor > 0` with appropriate `beta_fast/beta_slow` - **NTK-aware**: Adjust `freq_base` for NTK-style scaling

Common `freq_base` values: - LLaMA 1/2: 10000 - LLaMA 3: 500000 - Mistral: 10000 - Phi-3: 10000

Value

Tensor with extended RoPE applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
q <- ggml_new_tensor_4d(ctx, GGML_TYPE_F32, 64, 8, 32, 1)
ggml_set_f32(q, rnorm(64 * 8 * 32))
pos <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 32)
ggml_set_i32(pos, 0:31)
# Standard RoPE with default freq_base
q_rope <- ggml_rope_ext(ctx, q, pos, NULL,
                        n_dims = 64, mode = 0L,
                        n_ctx_orig = 4096,
                        freq_base = 10000, freq_scale = 1.0,
                        ext_factor = 0.0, attn_factor = 1.0,
                        beta_fast = 32, beta_slow = 1)
graph <- ggml_build_forward_expand(ctx, q_rope)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_rope_ext_back	<i>RoPE Extended Backward (Graph)</i>
--------------------	---------------------------------------

Description

Backward pass for extended RoPE (Rotary Position Embedding). Used during training to compute gradients through RoPE.

Usage

```
ggml_rope_ext_back(  
    ctx,  
    a,  
    b,  
    c = NULL,  
    n_dims,  
    mode = 0L,  
    n_ctx_orig = 0L,  
    freq_base = 10000,  
    freq_scale = 1,  
    ext_factor = 0,  
    attn_factor = 1,  
    beta_fast = 32,  
    beta_slow = 1  
)
```

Arguments

ctx	GGML context
a	Gradient tensor from upstream (gradients of ggml_rope_ext result)
b	Position tensor (same as forward pass)
c	Optional frequency factors tensor (NULL for default)
n_dims	Number of dimensions for rotation
mode	RoPE mode
n_ctx_orig	Original context length
freq_base	Base frequency
freq_scale	Frequency scale factor
ext_factor	Extension factor (YaRN)
attn_factor	Attention factor
beta_fast	YaRN fast beta
beta_slow	YaRN slow beta

Value

Gradient tensor for the input

ggml_rope_ext_inplace *Extended RoPE Inplace (Graph)*

Description

Creates a graph node for extended RoPE, modifying input tensor in place. Returns a view of the input tensor.

Usage

```
ggml_rope_ext_inplace(
    ctx,
    a,
    b,
    c = NULL,
    n_dims,
    mode = 0L,
    n_ctx_orig = 0L,
    freq_base = 10000,
    freq_scale = 1,
    ext_factor = 0,
    attn_factor = 1,
    beta_fast = 32,
    beta_slow = 1
)
```

Arguments

ctx	GGML context
a	Input tensor
b	Position tensor (int32)
c	Optional frequency factors tensor (NULL for default)
n_dims	Number of dimensions to apply rotation to
mode	RoPE mode
n_ctx_orig	Original context length the model was trained on
freq_base	Base frequency for RoPE (default 10000 for most models)
freq_scale	Frequency scale factor (1.0 = no scaling)
ext_factor	YaRN extension factor (0.0 to disable)
attn_factor	Attention scale factor (typically 1.0)
beta_fast	YaRN parameter for fast dimensions
beta_slow	YaRN parameter for slow dimensions

Value

View of input tensor with RoPE applied in place

See Also

Other rope: [ggml_rope_multi\(\)](#), [ggml_rope_multi_inplace\(\)](#)

ggml_rope_inplace	<i>Rotary Position Embedding In-place (Graph)</i>
-------------------	---

Description

In-place version of `ggml_rope`. Returns a view of the input tensor.

Usage

```
ggml_rope_inplace(ctx, a, b, n_dims, mode = 0L)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
b	Position tensor (int32)
n_dims	Number of dimensions to apply rotation to
mode	RoPE mode

Value

View of input tensor with RoPE applied

ggml_rope_multi	<i>Multi-RoPE for Vision Models (Graph)</i>
-----------------	---

Description

Creates a graph node for multi-dimensional RoPE (MRoPE) used in vision transformers. Supports separate rotation for different positional dimensions (e.g., height, width, time).

Usage

```

ggml_rope_multi(
    ctx,
    a,
    b,
    c = NULL,
    n_dims,
    sections = c(0L, 0L, 0L, 0L),
    mode = 0L,
    n_ctx_orig = 0L,
    freq_base = 10000,
    freq_scale = 1,
    ext_factor = 0,
    attn_factor = 1,
    beta_fast = 32,
    beta_slow = 1
)

```

Arguments

ctx	GGML context
a	Input tensor
b	Position tensor (int32)
c	Optional frequency factors tensor (NULL for default)
n_dims	Number of dimensions to apply rotation to
sections	Integer vector of length 4 specifying dimension sections for MRoPE
mode	RoPE mode
n_ctx_orig	Original context length the model was trained on
freq_base	Base frequency for RoPE (default 10000 for most models)
freq_scale	Frequency scale factor (1.0 = no scaling)
ext_factor	YaRN extension factor (0.0 to disable)
attn_factor	Attention scale factor (typically 1.0)
beta_fast	YaRN parameter for fast dimensions
beta_slow	YaRN parameter for slow dimensions

Value

Tensor with multi-dimensional RoPE applied

See Also

Other rope: [ggml_rope_ext_inplace\(\)](#), [ggml_rope_multi_inplace\(\)](#)

ggml_rope_multi_inplace

Multi-RoPE Inplace (Graph)

Description

Creates a graph node for multi-dimensional RoPE, modifying input in place.

Usage

```
ggml_rope_multi_inplace(
    ctx,
    a,
    b,
    c = NULL,
    n_dims,
    sections = c(0L, 0L, 0L, 0L),
    mode = 0L,
    n_ctx_orig = 0L,
    freq_base = 10000,
    freq_scale = 1,
    ext_factor = 0,
    attn_factor = 1,
    beta_fast = 32,
    beta_slow = 1
)
```

Arguments

ctx	GGML context
a	Input tensor
b	Position tensor (int32)
c	Optional frequency factors tensor (NULL for default)
n_dims	Number of dimensions to apply rotation to
sections	Integer vector of length 4 specifying dimension sections for MRoPE
mode	RoPE mode
n_ctx_orig	Original context length the model was trained on
freq_base	Base frequency for RoPE (default 10000 for most models)
freq_scale	Frequency scale factor (1.0 = no scaling)
ext_factor	YaRN extension factor (0.0 to disable)
attn_factor	Attention scale factor (typically 1.0)
beta_fast	YaRN parameter for fast dimensions
beta_slow	YaRN parameter for slow dimensions

Value

View of input tensor with MRoPE applied in place

See Also

Other rope: [ggml_rope_ext_inplace\(\)](#), [ggml_rope_multi\(\)](#)

ggml_round	<i>Round (Graph)</i>
------------	----------------------

Description

Creates a graph node for element-wise rounding: round(x)

Usage

```
ggml_round(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the round operation

ggml_round_inplace	<i>Round In-place (Graph)</i>
--------------------	-------------------------------

Description

Creates a graph node for in-place element-wise rounding.

Usage

```
ggml_round_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with rounded values

<code>ggml_run</code>	<i>Run a single-cell task on the GGML backend</i>
-----------------------	---

Description

Validates a `ggml_task` against `ggml_ops_registry`, resolves the compute device, invokes the registered engine and returns a `ggml_result`. This is the dispatch layer: the only place that decides GPU vs CPU. `device = "auto"` (the task default) uses Vulkan when a GPU is available and CPU otherwise; `device = "vulkan"` on a machine with no GPU degrades to CPU with a message rather than failing.

Usage

```
ggml_run(task, backend = NULL, ...)
```

Arguments

- `task` A `ggml_task`.
- `backend` Optional override of `task$device`: "auto" (default), "vulkan" or "cpu".
- `...` Additional parameters merged over `task$params` and passed to the engine.

Value

A `ggml_result`.

See Also

`ggml_task`, `ggml_ops_registry`

<code>ggml_save_model</code>	<i>Save a Full Model (Architecture + Weights)</i>
------------------------------	---

Description

Saves both the architecture and trained weights of a model to an RDS file. Unlike `ggml_save_weights()`, which requires the model to be manually reconstructed before loading, `ggml_save_model()` saves everything needed to restore the model with a single call to `ggml_load_model()`.

Usage

```
ggml_save_model(model, path)
```

Arguments

- `model` A trained `ggml_sequential_model` or `ggml_functional_model`.
- `path` File path (typically `.rds`).

Value

The model (invisibly).

Supported model types

- `ggml_sequential_model` — input shape, layer configs, trained weights, and compilation settings are all saved.
- `ggml_functional_model` — input/output node graphs (pure R lists, no ggml pointers) and trained `node_weights` are saved.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(16L, activation = "relu", input_shape = 4L) |>
  ggml_layer_dense(2L, activation = "softmax")
model <- ggml_compile(model, optimizer = "adam",
  loss = "categorical_crossentropy")
x <- matrix(runif(64 * 4), 64, 4)
y <- matrix(c(rep(c(1,0), 32), rep(c(0,1), 32)), 64, 2)
model <- ggml_fit(model, x, y, epochs = 1L, batch_size = 32L, verbose = 0L)
tmp <- tempfile(fileext = ".rds")
ggml_save_model(model, tmp)
model2 <- ggml_load_model(tmp)
```

<code>ggml_save_weights</code>	<i>Save Model Weights to File</i>
--------------------------------	-----------------------------------

Description

Saves the trained weights of a sequential model to an RDS file. The file includes both weights and architecture metadata for validation when loading.

Usage

```
ggml_save_weights(model, path)
```

Arguments

<code>model</code>	A trained <code>ggml_sequential_model</code>
<code>path</code>	File path to save weights (typically with <code>.rds</code> extension)

Value

The model (invisibly).

ggml_scale	Scale (Graph)
------------	---------------

Description

Creates a graph node for scaling tensor by a scalar: $x * s$

Usage

```
ggml_scale(ctx, a, s)
```

Arguments

ctx	GGML context
a	Input tensor
s	Scalar value to multiply by

Value

Tensor representing the scaled values

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_scale(ctx, a, 2.0)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [2, 4, 6, 8]
ggml_free(ctx)
```

ggml_scale_inplace	Scale Tensor In-place (Graph)
--------------------	-------------------------------

Description

Creates a graph node for in-place scaling: $a * s$

Usage

```
ggml_scale_inplace(ctx, a, s)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)
s	Scalar value to multiply by

Value

View of tensor a with scaled values

ggml_schedule_cosine_decay
Cosine annealing LR scheduler

Description

Anneals LR from initial value to 'eta_min' following a cosine curve.

Usage

```
ggml_schedule_cosine_decay(eta_min = 0, T_max = NULL)
```

Arguments

eta_min	Minimum LR at end of schedule
T_max	Total number of epochs (defaults to nepoch from fit state)

Value

List with on_epoch_begin function

See Also

Other callbacks: [ggml_callback_early_stopping\(\)](#), [ggml_schedule_reduce_on_plateau\(\)](#), [ggml_schedule_step_decay\(\)](#)

ggml_schedule_reduce_on_plateau
<i>Reduce on plateau LR scheduler</i>

Description

Reduces LR when a metric stops improving.

Usage

```
ggml_schedule_reduce_on_plateau(  
    monitor = "val_loss",  
    factor = 0.5,  
    patience = 5,  
    min_lr = 1e-07,  
    min_delta = 1e-04,  
    mode = "auto"  
)
```

Arguments

monitor	Metric to monitor: "val_loss", "train_loss", etc.
factor	Factor to reduce LR by
patience	Epochs with no improvement before reducing
min_lr	Minimum LR
min_delta	Minimum change to qualify as improvement
mode	"min" or "max". "auto" infers from monitor name.

Value

List with on_epoch_end function

See Also

Other callbacks: [ggml_callback_early_stopping\(\)](#), [ggml_schedule_cosine_decay\(\)](#), [ggml_schedule_step_decay\(\)](#)

ggml_schedule_step_decay	<i>Step decay LR scheduler</i>
--------------------------	--------------------------------

Description

Reduces LR by a factor every ‘step_size’ epochs.

Usage

ggml_schedule_step_decay(step_size = 10, gamma = 0.1)

Arguments

step_size	Reduce LR every this many epochs
gamma	Multiplicative factor of LR reduction

Value

List with on_epoch_begin function

See Also

Other callbacks: [ggml_callback_early_stopping\(\)](#), [ggml_schedule_cosine_decay\(\)](#), [ggml_schedule_reduce_on_pl](#)

ggml_set	<i>Set Tensor Region (Graph)</i>
----------	----------------------------------

Description

Copies tensor b into tensor a at a specified offset. This allows writing to a portion of a tensor.

Usage

ggml_set(ctx, a, b, nb1, nb2, nb3, offset)

Arguments

ctx	GGML context
a	Destination tensor
b	Source tensor (data to copy)
nb1	Stride for dimension 1 (in bytes)
nb2	Stride for dimension 2 (in bytes)
nb3	Stride for dimension 3 (in bytes)
offset	Byte offset in destination tensor

Value

Tensor representing the set operation

ggml_set_1d	<i>Set 1D Tensor Region (Graph)</i>
-------------	-------------------------------------

Description

Simplified 1D version of ggml_set. Copies tensor b into tensor a starting at offset.

Usage

ggml_set_1d(ctx, a, b, offset)

Arguments

ctx	GGML context
a	Destination tensor
b	Source tensor
offset	Byte offset in destination tensor

Value

Tensor representing the set operation

ggml_set_2d	<i>Set 2D Tensor Region (Graph)</i>
-------------	-------------------------------------

Description

Simplified 2D version of ggml_set.

Usage

ggml_set_2d(ctx, a, b, nb1, offset)

Arguments

ctx	GGML context
a	Destination tensor
b	Source tensor
nb1	Stride for dimension 1 (in bytes)
offset	Byte offset in destination tensor

Value

Tensor representing the set operation

`ggml_set_abort_callback_default`*Restore Default Abort Behavior*

Description

Restores GGML to default abort behavior (prints to stderr and aborts).

Usage

```
ggml_set_abort_callback_default()
```

Value

NULL invisibly

See Also

Other logging: [ggml_abort_is_r_enabled\(\)](#), [ggml_log_is_r_enabled\(\)](#), [ggml_log_set_default\(\)](#), [ggml_log_set_r\(\)](#), [ggml_set_abort_callback_r\(\)](#)

`ggml_set_abort_callback_r`*Enable R-compatible Abort Handling*

Description

Converts GGML abort calls into R errors (via `Rf_error`). This allows R to catch GGML failures with `tryCatch`.

Usage

```
ggml_set_abort_callback_r()
```

Value

NULL invisibly

See Also

Other logging: [ggml_abort_is_r_enabled\(\)](#), [ggml_log_is_r_enabled\(\)](#), [ggml_log_set_default\(\)](#), [ggml_log_set_r\(\)](#), [ggml_set_abort_callback_default\(\)](#)

Examples

```
ggml_set_abort_callback_r()
# Now GGML aborts will become R errors
result <- tryCatch({
  # ... ggml operations that might fail ...
}, error = function(e) {
  message("GGML error caught: ", e$message)
})
```

ggml_set_f32

*Set F32 data***Description**

Set F32 data

Set F32 Data

Usage

```
ggml_set_f32(tensor, data)
```

```
ggml_set_f32(tensor, data)
```

Arguments

tensor	Tensor
data	Numeric vector

Value

NULL (invisible)

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
tensor <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(tensor, c(1, 2, 3, 4, 5))
ggml_get_f32(tensor)
ggml_free(ctx)
```

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(t, c(1, 2, 3, 4, 5))
ggml_get_f32(t)
ggml_free(ctx)
```

ggml_set_f32_nd	<i>Set Single Float Value by N-D Index</i>
-----------------	--

Description

Sets a single f32 value in the tensor at position [i0, i1, i2, i3]. This is a direct data write, not a graph operation.

Usage

```
ggml_set_f32_nd(tensor, i0, i1 = 0, i2 = 0, i3 = 0, value)
```

Arguments

tensor	Tensor pointer
i0, i1, i2, i3	Indices (0-based)
value	Float value to set

Value

NULL (invisible)

ggml_set_i32	<i>Set I32 Data</i>
--------------	---------------------

Description

Sets integer data in an I32 tensor. Used for indices (ggml_get_rows) and position tensors (ggml_rope).

Usage

```
ggml_set_i32(tensor, data)
```

Arguments

tensor	Tensor of type GGML_TYPE_I32
data	Integer vector

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
pos <- ggml_new_tensor_1d(ctx, GGML_TYPE_I32, 10)
ggml_set_i32(pos, 0:9)
ggml_get_i32(pos)
ggml_free(ctx)
```

ggml_set_i32_nd	<i>Set Single Int32 Value by N-D Index</i>
-----------------	--

Description

Sets a single i32 value in the tensor at position [i0, i1, i2, i3].

Usage

```
ggml_set_i32_nd(tensor, i0, i1 = 0, i2 = 0, i3 = 0, value)
```

Arguments

tensor	Tensor pointer
i0, i1, i2, i3	Indices (0-based)
value	Integer value to set

Value

NULL (invisible)

ggml_set_input	<i>Mark Tensor as Input</i>
----------------	-----------------------------

Description

Mark Tensor as Input

Usage

```
ggml_set_input(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

The tensor (for chaining)

ggml_set_name	<i>Set Tensor Name</i>
---------------	------------------------

Description

Assigns a name to a tensor. Useful for debugging and graph visualization.

Usage

```
ggml_set_name(tensor, name)
```

Arguments

tensor	Tensor pointer
name	Character string name

Value

The tensor (for chaining)

ggml_set_no_alloc	<i>Set No Allocation Mode</i>
-------------------	-------------------------------

Description

When enabled, tensor creation will not allocate memory for data. Useful for creating computation graphs without allocating storage.

Usage

```
ggml_set_no_alloc(ctx, no_alloc)
```

Arguments

ctx	GGML context
no_alloc	Logical, TRUE to disable allocation

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
ggml_set_no_alloc(ctx, TRUE)
ggml_get_no_alloc(ctx)
ggml_set_no_alloc(ctx, FALSE)
ggml_free(ctx)
```

ggml_set_n_threads	<i>Set Number of Threads</i>
--------------------	------------------------------

Description

Set the number of threads for GGML operations

Usage

```
ggml_set_n_threads(n_threads)
```

Arguments

n_threads	Number of threads to use
-----------	--------------------------

Value

Number of threads set

Examples

```
# Use 4 threads
ggml_set_n_threads(4)

# Use all available cores
ggml_set_n_threads(parallel::detectCores())
```

ggml_set_omp_threads	<i>Set OpenMP Thread Count</i>
----------------------	--------------------------------

Description

Directly calls `omp_set_num_threads()` to limit OpenMP parallelism. Useful in tests to comply with CRAN policy on core usage.

Usage

```
ggml_set_omp_threads(n)
```


Arguments

n Number of threads

Value

NULL invisibly

ggml_set_op_params	<i>Set Tensor Operation Parameters</i>
--------------------	--

Description

Sets the raw op_params bytes for a tensor.

Usage

ggml_set_op_params(tensor, params)

Arguments

tensor External pointer to tensor
params Raw vector of parameters (max 64 bytes)

Value

NULL invisibly

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params_f32\(\)](#), [ggml_set_op_params_i32\(\)](#)

ggml_set_op_params_f32	<i>Set Float Op Parameter</i>
------------------------	-------------------------------

Description

Sets a single float value in tensor op_params at given index.

Usage

ggml_set_op_params_f32(tensor, index, value)

Arguments

tensor	External pointer to tensor
index	0-based index (0-15 for 64-byte op_params)
value	Numeric value to set

Value

NULL invisibly

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_i32\(\)](#)

ggml_set_op_params_i32
<i>Set Integer Op Parameter</i>

Description

Sets a single int32 value in tensor op_params at given index.

Usage

ggml_set_op_params_i32(tensor, index, value)

Arguments

tensor	External pointer to tensor
index	0-based index (0-15 for 64-byte op_params)
value	Integer value to set

Value

NULL invisibly

See Also

Other tensor: [ggml_are_same_layout\(\)](#), [ggml_get_op_params\(\)](#), [ggml_get_op_params_f32\(\)](#), [ggml_get_op_params_i32\(\)](#), [ggml_set_op_params\(\)](#), [ggml_set_op_params_f32\(\)](#)

ggml_set_output	<i>Mark Tensor as Output</i>
-----------------	------------------------------

Description

Mark Tensor as Output

Usage

ggml_set_output(tensor)

Arguments

tensor Tensor pointer

Value

The tensor (for chaining)

ggml_set_param	<i>Set Tensor as Trainable Parameter</i>
----------------	--

Description

Marks a tensor as a trainable parameter for backpropagation. The optimizer will compute gradients for this tensor during training.

Usage

ggml_set_param(tensor)

Arguments

tensor Tensor pointer

Value

The tensor (for chaining)

`ggml_set_seed`*Set the random seed for reproducible ggmlR runs*

Description

Fixes the random seed used by ggmlR for everything that is stochastic:

- weight initialisation (sequential, functional and autograd layers),
- dropout masks (training-time),
- data shuffling in the autograd dataloader / training loops.

Usage

```
ggml_set_seed(seed)
```

Arguments

seed	A single integer (or value coercible to integer) used as the RNG seed. NULL is a no-op, which lets callers thread an optional seed argument through without special-casing it.
------	--

Details

This is a thin wrapper around [set.seed](#): all randomness in ggmlR is produced by the base R RNG, so a fixed seed gives identical starting weights, dropout masks and batch ordering across runs. It is the single point of control used by the **mlr3** learners (seed hyperparameter) and the **parsnip** "ggml" engine (seed engine argument).

GPU note: this controls the random *inputs* to the computation, not the floating-point arithmetic itself. GPU (Vulkan) kernels are run-to-run stable on a given device/driver for the standard forward/backward paths, but ggmlR does not guarantee *bit-for-bit* identical results across different devices, drivers or backends (CPU vs Vulkan). Reproducibility is at the level of training dynamics, not exact bits.

Value

Invisibly returns seed.

See Also

[set.seed](#)

Examples

```
ggml_set_seed(42)
a <- runif(3)
ggml_set_seed(42)
b <- runif(3)
identical(a, b) # TRUE
```

ggml_set_zero	<i>Set Tensor to Zero</i>
---------------	---------------------------

Description

Sets all elements of a tensor to zero. This is more efficient than manually setting all elements.

Usage

```
ggml_set_zero(tensor)
```

Arguments

tensor	Tensor to zero out
--------	--------------------

Value

NULL (invisible)

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_set_f32(t, 1:10)
ggml_set_zero(t)
ggml_get_f32(t)
ggml_free(ctx)
```

ggml_sgn	<i>Sign Function (Graph)</i>
----------	------------------------------

Description

Creates a graph node for element-wise sign function. $\text{sgn}(x) = -1$ if $x < 0$, 0 if $x == 0$, 1 if $x > 0$

Usage

```
ggml_sgn(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the sign operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -0.5, 0, 0.5, 2))
r <- ggml_sgn(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # c(-1, -1, 0, 1, 1)
ggml_free(ctx)
```

ggml_sigmoid	<i>Sigmoid Activation (Graph)</i>
--------------	-----------------------------------

Description

Creates a graph node for sigmoid activation: $1 / (1 + \exp(-x))$

Usage

```
ggml_sigmoid(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the sigmoid operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_sigmoid(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_sigmoid_inplace	<i>Sigmoid Activation In-place (Graph)</i>
----------------------	--

Description

Creates a graph node for in-place sigmoid activation: $1 / (1 + e^{(-x)})$

Usage

```
ggml_sigmoid_inplace(ctx, a)
```

Arguments

- | | |
|-----|--|
| ctx | GGML context |
| a | Input tensor (will be modified in-place) |

Value

View of tensor a with sigmoid applied

ggml_silu	<i>SiLU Activation (Graph)</i>
-----------	--------------------------------

Description

Creates a graph node for SiLU (Sigmoid Linear Unit) activation, also known as Swish. **CRITICAL** for LLaMA models.

Usage

```
ggml_silu(ctx, a)
```

Arguments

- | | |
|-----|--------------|
| ctx | GGML context |
| a | Input tensor |

Value

Tensor representing the SiLU operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_silu(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_silu_back	<i>SiLU Backward (Graph)</i>
----------------	------------------------------

Description

Computes the backward pass for SiLU (Swish) activation. Used during training for gradient computation.

Usage

```
ggml_silu_back(ctx, a, b)
```

Arguments

ctx	GGML context
a	Forward input tensor
b	Gradient tensor from upstream

Value

Gradient tensor for the input

ggml_silu_inplace	<i>SiLU Activation In-place (Graph)</i>
-------------------	---

Description

Creates a graph node for in-place SiLU (Sigmoid Linear Unit) activation. CRITICAL for LLaMA models with memory efficiency.

Usage

```
ggml_silu_inplace(ctx, a)
```


Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with SiLU applied

ggml_sin	<i>Sine (Graph)</i>
----------	---------------------

Description

Creates a graph node for element-wise sine: $\sin(x)$

Usage

```
ggml_sin(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the sin operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(0, pi/6, pi/2, pi))
result <- ggml_sin(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [0, 0.5, 1, 0]
ggml_free(ctx)
```

ggml_softplus	<i>Softplus Activation (Graph)</i>
---------------	------------------------------------

Description

Creates a graph node for Softplus activation. $\text{Softplus}(x) = \log(1 + \exp(x))$. A smooth approximation of ReLU.

Usage

ggml_softplus(ctx, a)

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the Softplus operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
r <- ggml_softplus(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r)
ggml_free(ctx)
```

ggml_softplus_inplace	<i>Softplus Activation In-place (Graph)</i>
-----------------------	---

Description

Creates a graph node for in-place softplus activation: $\log(1 + e^x)$

Usage

ggml_softplus_inplace(ctx, a)

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with softplus applied

ggml_soft_max	<i>Softmax (Graph)</i>
---------------	------------------------

Description

Creates a graph node for softmax operation. CRITICAL for attention mechanisms.

Usage

```
ggml_soft_max(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the softmax operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_soft_max(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result)
# Output sums to 1.0
ggml_free(ctx)
```

ggml_soft_max_ext	<i>Extended Softmax with Masking and Scaling (Graph)</i>
-------------------	--

Description

Creates a graph node for fused softmax operation with optional masking and ALiBi (Attention with Linear Biases) support. Computes: $\text{softmax}(a * \text{scale} + \text{mask} * (\text{ALiBi slope}))$ CRITICAL for efficient attention computation in transformers.

Usage

```
ggml_soft_max_ext(ctx, a, mask = NULL, scale = 1, max_bias = 0)
```

Arguments

ctx	GGML context
a	Input tensor (typically attention scores)
mask	Optional attention mask tensor (F16 or F32). NULL for no mask. Shape must be broadcastable to input tensor.
scale	Scaling factor, typically $1/\text{sqrt}(\text{head_dim})$
max_bias	Maximum ALiBi bias (0.0 to disable ALiBi)

Details

This extended softmax is commonly used in transformer attention: 1. Scale attention scores by $1/\text{sqrt}(d_k)$ for numerical stability 2. Apply attention mask (e.g., causal mask, padding mask) 3. Optionally apply ALiBi position bias 4. Compute softmax
All these operations are fused for efficiency.

Value

Tensor representing the scaled and masked softmax

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
scores <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 10)
ggml_set_f32(scores, rnorm(100))
attn <- ggml_soft_max_ext(ctx, scores, NULL, 1.0, max_bias = 0.0)
graph <- ggml_build_forward_expand(ctx, attn)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_soft_max_ext_back
<i>Softmax Backward Extended (Graph)</i>

Description

Backward pass for extended softmax operation.

Usage

ggml_soft_max_ext_back(ctx, a, b, scale = 1, max_bias = 0)

Arguments

ctx	GGML context
a	Softmax output tensor (from forward pass)
b	Gradient tensor from upstream
scale	Scale factor (same as forward pass)
max_bias	Maximum ALiBi bias (same as forward pass)

Value

Gradient tensor for the input

ggml_soft_max_ext_back_inplace
<i>Extended Softmax Backward Inplace (Graph)</i>

Description

Creates a graph node for the backward pass of extended softmax, modifying in place.

Usage

ggml_soft_max_ext_back_inplace(ctx, a, b, scale = 1, max_bias = 0)

Arguments

ctx	GGML context
a	Gradient tensor from upstream
b	Softmax output from forward pass
scale	Scaling factor used in forward pass
max_bias	Maximum ALiBi bias used in forward pass

Value

View of input tensor with gradient computed in place

See Also

Other softmax: [ggml_soft_max_ext_inplace\(\)](#)

ggml_soft_max_ext_inplace
<i>Extended Softmax Inplace (Graph)</i>

Description

Creates a graph node for extended softmax, modifying input tensor in place. Returns a view of the input tensor.

Usage

```
ggml_soft_max_ext_inplace(ctx, a, mask = NULL, scale = 1, max_bias = 0)
```

Arguments

ctx	GGML context
a	Input tensor (typically attention scores)
mask	Optional attention mask tensor (F16 or F32). NULL for no mask. Shape must be broadcastable to input tensor.
scale	Scaling factor, typically 1/sqrt(head_dim)
max_bias	Maximum ALiBi bias (0.0 to disable ALiBi)

Value

View of input tensor with softmax applied in place

See Also

Other softmax: [ggml_soft_max_ext_back_inplace\(\)](#)

ggml_soft_max_inplace *Softmax In-place (Graph)*

Description

Creates a graph node for in-place softmax operation. Returns a view of the input tensor.

Usage

```
ggml_soft_max_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of input tensor with softmax applied

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_soft_max_inplace(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

GGML_SORT_ORDER_ASC *Sort Order Constants*

Description

Sort Order Constants

Usage

```
GGML_SORT_ORDER_ASC

GGML_SORT_ORDER_DESC
```

Format

Integer constants
An object of class integer of length 1.

Details

Constants for specifying sort order in argsort operations.

- GGML_SORT_ORDER_ASC (0): Ascending order (smallest first)
- GGML_SORT_ORDER_DESC (1): Descending order (largest first)

Value

An integer constant representing a sort order

Examples

```
GGML_SORT_ORDER_ASC    # 0 - Ascending order
GGML_SORT_ORDER_DESC    # 1 - Descending order

# Usage with ggml_argsort
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(3, 1, 4, 1, 5))
# Get ascending sort indices
idx_asc <- ggml_argsort(ctx, a, GGML_SORT_ORDER_ASC)
# Get descending sort indices
idx_desc <- ggml_argsort(ctx, a, GGML_SORT_ORDER_DESC)
ggml_free(ctx)
```

ggml_sqr	<i>Square (Graph)</i>
----------	-----------------------

Description

Creates a graph node for element-wise squaring: x^2

Usage

```
ggml_sqr(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the square operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 2, 3, 4))
result <- ggml_sqr(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [1, 4, 9, 16]
ggml_free(ctx)
```

ggml_sqrt	<i>Square Root (Graph)</i>
-----------	----------------------------

Description

Creates a graph node for element-wise square root: sqrt(x)

Usage

```
ggml_sqrt(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the sqrt operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 4)
ggml_set_f32(a, c(1, 4, 9, 16))
result <- ggml_sqrt(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [1, 2, 3, 4]
ggml_free(ctx)
```

ggml_sqrt_inplace	<i>Square Root In-place (Graph)</i>
-------------------	-------------------------------------

Description

Creates a graph node for in-place element-wise square root.

Usage

```
ggml_sqrt_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with square root values

ggml_sqr_inplace	<i>Square In-place (Graph)</i>
------------------	--------------------------------

Description

Creates a graph node for in-place element-wise square: x^2

Usage

```
ggml_sqr_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with squared values

ggml_step	<i>Step Function (Graph)</i>
-----------	------------------------------

Description

Creates a graph node for element-wise step function. $\text{step}(x) = 0$ if $x \leq 0$, 1 if $x > 0$ Also known as the Heaviside step function.

Usage

```
ggml_step(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the step operation

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -0.5, 0, 0.5, 2))
r <- ggml_step(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
result <- ggml_get_f32(r) # c(0, 0, 0, 1, 1)
ggml_free(ctx)
```

ggml_sub	<i>Element-wise Subtraction (Graph)</i>
----------	---

Description

Creates a graph node for element-wise subtraction.

Usage

```
ggml_sub(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor
b	Second tensor (same shape as a)

Value

Tensor representing the subtraction operation (a - b)

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
b <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(5, 4, 3, 2, 1))
ggml_set_f32(b, c(1, 1, 1, 1, 1))
result <- ggml_sub(ctx, a, b)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_sub_inplace	<i>Element-wise Subtraction In-place (Graph)</i>
------------------	--

Description

Creates a graph node for in-place element-wise subtraction. Result is stored in tensor a, saving memory allocation.

Usage

```
ggml_sub_inplace(ctx, a, b)
```

Arguments

ctx	GGML context
a	First tensor (will be modified in-place)
b	Second tensor (same shape as a)

Value

View of tensor a with the subtraction result

ggml_sum	<i>Sum (Graph)</i>
----------	--------------------

Description

Creates a graph node that computes the sum of all elements.

Usage

```
ggml_sum(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Scalar tensor with the sum

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(1, 2, 3, 4, 5))
result <- ggml_sum(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # 15
ggml_free(ctx)
```

ggml_sum_rows	<i>Sum Rows (Graph)</i>
---------------	-------------------------

Description

Creates a graph node that computes the sum along rows.

Usage

```
ggml_sum_rows(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor with row sums

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 3, 2)
ggml_set_f32(a, c(1, 2, 3, 4, 5, 6))
result <- ggml_sum_rows(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
output <- ggml_get_f32(result) # [6, 15]
ggml_free(ctx)
```

ggml_swiglu	<i>SwiGLU (Swish/SiLU Gated Linear Unit) (Graph)</i>
-------------	--

Description

Creates a graph node for SwiGLU operation. SwiGLU uses SiLU (Swish) as the activation function on the first half. CRITICAL for LLaMA, Mistral, and many modern LLMs.

Usage

```
ggml_swiglu(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (first dimension must be even)

Details

Formula: $output = SiLU(x) * gate$

Value

Tensor with half the first dimension of input

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 8, 3)
ggml_set_f32(a, rnorm(24))
r <- ggml_swiglu(ctx, a)
graph <- ggml_build_forward_expand(ctx, r)
ggml_graph_compute(ctx, graph)
```

```
result <- ggml_get_f32(r) # Shape: 4x3
ggml_free(ctx)
```

ggml_swiglu_split	<i>SwiGLU Split (Graph)</i>
-------------------	-----------------------------

Description

Creates a graph node for SwiGLU with separate input and gate tensors.

Usage

```
ggml_swiglu_split(ctx, a, b)
```

Arguments

ctx	GGML context
a	Input tensor (the values to be gated)
b	Gate tensor (same shape as a)

Details

Formula: $output = SiLU(a) * b$

Value

Tensor with same shape as input tensors

ggml_tanh	<i>Tanh Activation (Graph)</i>
-----------	--------------------------------

Description

Creates a graph node for hyperbolic tangent activation.

Usage

```
ggml_tanh(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor

Value

Tensor representing the tanh operation

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 5)
ggml_set_f32(a, c(-2, -1, 0, 1, 2))
result <- ggml_tanh(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
ggml_get_f32(result)
ggml_free(ctx)
```

ggml_tanh_inplace	<i>Tanh Activation In-place (Graph)</i>
-------------------	---

Description

Creates a graph node for in-place hyperbolic tangent activation.

Usage

```
ggml_tanh_inplace(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (will be modified in-place)

Value

View of tensor a with tanh applied

ggml_task	<i>Construct a single-cell compute task</i>
-----------	---

Description

A ggml_task is the contract object passed from the extraction layer to the dispatch layer. It bundles the operation name, the dense feature matrix (features in rows, cells in columns — the single-cell convention), the operation parameters and the requested device. It performs no computation.

Usage

```
ggml_task(op, matrix, params = list(), device = c("auto", "vulkan", "cpu"))
```


Arguments

op	Operation name; must be registered in ggml_ops_registry (e.g. "embed").
matrix	A numeric matrix (dense) or dgCMatrix (sparse). Rows are features (genes), columns are cells.
params	Named list of operation parameters (e.g. n_components).
device	"vulkan", "cpu" or "auto" (default).

Value

An object of class `ggml_task`.

See Also

[ggml_run](#), [ggml_ops_registry](#)

ggml_tcrossprod	<i>GPU transposed cross product (drop-in for tcrossprod)</i>
-----------------	--

Description

Computes $A \%*\% t(B)$ (or $A \%*\% t(A)$ when `B` is `NULL`) on the Vulkan GPU, returning an R matrix, with a CPU fallback. See [ggml_matmul](#) for the precision and dispatch notes.

Usage

```
ggml_tcrossprod(A, B = NULL, device = "auto", prec = "f32")
```

Arguments

A	A numeric matrix.
B	A numeric matrix with <code>ncol(B) == ncol(A)</code> , or <code>NULL</code> (default) for $A \%*\% t(A)$.
device, prec	See ggml_matmul .

Value

$A \%*\% t(B)$ as a numeric matrix.

ggml_tensor_copy	<i>Copy Tensor Data</i>
------------------	-------------------------

Description

Copies raw data from src tensor to dst tensor (must be same size).

Usage

```
ggml_tensor_copy(dst, src)
```

Arguments

dst	Destination tensor
src	Source tensor

Value

NULL (invisible)

ggml_tensor_nb	<i>Get Tensor Strides (nb)</i>
----------------	--------------------------------

Description

Returns the byte strides for each dimension of the tensor.

Usage

```
ggml_tensor_nb(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Numeric vector of 4 stride values (nb0, nb1, nb2, nb3)

ggml_tensor_num	<i>Count Tensors in Context</i>
-----------------	---------------------------------

Description

Counts the number of tensors allocated in a context.

Usage

ggml_tensor_num(ctx)

Arguments

ctx GGML context

Value

Number of tensors

ggml_tensor_overhead	<i>Get Tensor Overhead</i>
----------------------	----------------------------

Description

Returns the memory overhead (metadata) for each tensor in bytes

Usage

ggml_tensor_overhead()

Value

Size in bytes

Examples

ggml_tensor_overhead()

ggml_tensor_set_f32_scalar	<i>Fill Tensor with Scalar</i>
----------------------------	--------------------------------

Description

Sets all elements of a f32 tensor to a single value.

Usage

```
ggml_tensor_set_f32_scalar(tensor, value)
```

Arguments

tensor	Tensor pointer
value	Float value to fill with

Value

NULL (invisible)

ggml_tensor_shape	<i>Get Tensor Shape</i>
-------------------	-------------------------

Description

Returns the shape of a tensor as a numeric vector of 4 elements (ne0, ne1, ne2, ne3)

Usage

```
ggml_tensor_shape(tensor)
```

Arguments

tensor	Tensor pointer
--------	----------------

Value

Numeric vector of length 4 with dimensions

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 10, 20)
ggml_tensor_shape(t)
ggml_free(ctx)
```

ggml_tensor_type	<i>Get Tensor Type</i>
------------------	------------------------

Description

Returns the data type of a tensor as an integer code

Usage

ggml_tensor_type(tensor)

Arguments

tensor Tensor pointer

Value

Integer type code (0 = F32, 1 = F16, etc.)

Examples

```
ctx <- ggml_init(1024 * 1024)
t <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
ggml_tensor_type(t)
ggml_free(ctx)
```

ggml_test	<i>Test GGML</i>
-----------	------------------

Description

Runs GGML library self-test and prints version info.

Usage

ggml_test()

Value

TRUE if test passed

Examples

ggml_test()

ggml_timestep_embedding	<i>Timestep Embedding (Graph Operation)</i>
-------------------------	---

Description

Creates sinusoidal timestep embeddings as used in diffusion models. Reference: CompVis/stable-diffusion util.py timestep_embedding

Usage

```
ggml_timestep_embedding(ctx, timesteps, dim, max_period = 10000L)
```

Arguments

ctx	GGML context
timesteps	Input tensor of timestep values [N]
dim	Embedding dimension
max_period	Maximum period for sinusoidal embedding (default 10000)

Value

Tensor of shape [N, dim] with timestep embeddings

ggml_time_init	<i>Initialize GGML Timer</i>
----------------	------------------------------

Description

Initializes the GGML timing system. Call this once at the beginning of the program before using ggml_time_ms() or ggml_time_us().

Usage

```
ggml_time_init()
```

Value

NULL (invisible)

Examples

```
ggml_time_init()
start <- ggml_time_ms()
Sys.sleep(0.01)
elapsed <- ggml_time_ms() - start
```

ggml_time_ms	<i>Get Time in Milliseconds</i>
--------------	---------------------------------

Description

Returns the current time in milliseconds since the timer was initialized.

Usage

```
ggml_time_ms()
```

Value

Numeric value representing milliseconds

Examples

```
ggml_time_init()
start <- ggml_time_ms()
Sys.sleep(0.01)
elapsed <- ggml_time_ms() - start
```

ggml_time_us	<i>Get Time in Microseconds</i>
--------------	---------------------------------

Description

Returns the current time in microseconds since the timer was initialized. More precise than ggml_time_ms() for micro-benchmarking.

Usage

```
ggml_time_us()
```

Value

Numeric value representing microseconds

Examples

```
ggml_time_init()
start <- ggml_time_us()
Sys.sleep(0.001)
elapsed <- ggml_time_us() - start
```

ggml_top_k	Top-K Indices (Graph)
------------	-----------------------

Description

Returns the indices of top K elements per row. Useful for sampling strategies in language models (top-k sampling). Note: the resulting indices are in no particular order within top-k.

Usage

```
ggml_top_k(ctx, a, k)
```

Arguments

ctx	GGML context
a	Input tensor (F32)
k	Number of top elements to return per row

Value

Tensor containing I32 indices of top-k elements (not values)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
# Logits from model output
logits <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
ggml_set_f32(logits, rnorm(100))
# Get top 5 logits for sampling
top5 <- ggml_top_k(ctx, logits, 5)
graph <- ggml_build_forward_expand(ctx, top5)
ggml_graph_compute(ctx, graph)
ggml_free(ctx)
```

ggml_tp_dp_forward	TPxDP hybrid matrix multiply across replicas of Vulkan device groups
--------------------	--

Description

Computes $Y = X \times W$ as a hybrid of tensor parallelism and data parallelism: the weight matrix W is replicated across replicas device groups (data parallelism over the batch X), and within each group the weight rows are split across that group's GPUs (tensor parallelism, via [ggml_vulkan_split_mul_mat](#)). Each replica computes a contiguous shard of the batch rows; the shards are concatenated back into the full result.

Usage

```
ggml_tp_dp_forward(
  W,
  X,
  replicas,
  weights = NULL,
  transport = c("host-staging", "opaque-fd", "device-group")
)
```

Arguments

W	Weight matrix, N x K (replicated to every group).
X	Activation matrix, M x K; its rows (the batch) are split across the replicas.
replicas	A list of integer vectors, each a group of physical GPU indices (0-based) forming one replica, e.g. <code>list(c(0, 1), c(2, 3))</code> . Within a group the weights are tensor-split; across groups the batch is data-split.
weights	Optional numeric vector (length = group size) giving the per-device TP row weighting inside each group (applied to every group). NULL (default) splits evenly.
transport	Cross-device TP gather transport (see ggml_vulkan_split_mul_mat).

Details

This mirrors the inference layout for the 4x P100 target (2 replicas x TP=2): there is no cross-replica communication, so throughput scales with the replica count while each replica's cross-device TP gather stays local to its group.

Value

The M x N result matrix, equal to `X %*% t(W)` up to the GPU's floating-point accumulation.

See Also

[ggml_vulkan_split_mul_mat](#) for the single-group TP path.

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 4) {
  W <- matrix(rnorm(2048 * 64), nrow = 2048)
  X <- matrix(rnorm(8 * 64), nrow = 8)      # batch of 8
  # 2 replicas x TP=2: batch split 4/4, weights split within {0,1} and {2,3}
  Y <- ggml_tp_dp_forward(W, X, replicas = list(c(0, 1), c(2, 3)))
  max(abs(Y - X %*% t(W)))
}
```

ggml_training_history	<i>Training history of a fitted ggml model</i>
-----------------------	--

Description

Returns the per-epoch loss / accuracy curve recorded during [ggml_fit](#), in a tidy data frame. This is the standard accessor for the loss curve; it works on a raw sequential/functional model or on a fitted parsnip engine object (e.g. from `extract_fit_engine()`).

Usage

```
ggml_training_history(object, format = c("wide", "long"), ...)
```

Arguments

object	A fitted <code>ggml_sequential_model</code> , <code>ggml_functional_model</code> or <code>ggmlr_parsnip_model</code> .
format	"wide" (default) returns one row per epoch with one column per metric. "long" returns one row per (epoch, metric) with columns epoch, metric (e.g. "loss"/"accuracy"), split ("train"/"val") and value — convenient for ggplot2 faceting.
...	Unused; for extensibility.

Value

A data frame (tibble if **tibble** is installed). Wide columns: epoch, train_loss, train_accuracy, and val_loss / val_accuracy when a validation split was used. Returns NULL with a warning if the model has no recorded history (e.g. not yet fitted).

See Also

[ggml_model_backend](#)

ggml_transpose	<i>Transpose (Graph)</i>
----------------	--------------------------

Description

Creates a graph node for matrix transpose operation.

Usage

```
ggml_transpose(ctx, a)
```

Arguments

ctx	GGML context
a	Input tensor (2D matrix)

Value

Tensor representing the transposed matrix

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 3, 2)
ggml_set_f32(a, 1:6)
result <- ggml_transpose(ctx, a)
graph <- ggml_build_forward_expand(ctx, result)
ggml_graph_compute(ctx, graph)
shape <- ggml_tensor_shape(result) # [2, 3]
ggml_free(ctx)
```

GGML_TYPE_F32	<i>GGML Data Types</i>
---------------	------------------------

Description

Constants representing different data types supported by GGML.

Usage

- GGML_TYPE_F32
- GGML_TYPE_F16
- GGML_TYPE_Q4_0
- GGML_TYPE_Q4_1
- GGML_TYPE_Q8_0
- GGML_TYPE_Q2_K
- GGML_TYPE_Q3_K
- GGML_TYPE_Q4_K
- GGML_TYPE_Q5_K
- GGML_TYPE_Q6_K
- GGML_TYPE_I32
- GGML_TYPE_BF16

Format

Integer constants

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

Details

- GGML_TYPE_F32: 32-bit floating point (default)
- GGML_TYPE_F16: 16-bit floating point (half precision)
- GGML_TYPE_Q4_0: 4-bit quantization type 0
- GGML_TYPE_Q4_1: 4-bit quantization type 1
- GGML_TYPE_Q8_0: 8-bit quantization type 0
- GGML_TYPE_I32: 32-bit integer
- GGML_TYPE_BF16: 16-bit brain float (bfloat16)

Value

An integer constant representing a GGML data type

Examples

GGML_TYPE_F32

GGML_TYPE_F16

GGML_TYPE_I32

ggml_type_name	<i>Get Type Name</i>
----------------	----------------------

Description

Returns the string name of a GGML type.

Usage

```
ggml_type_name(type)
```

Arguments

type GGML type constant (e.g., GGML_TYPE_F32)

Value

Character string with type name

See Also

Other type_system: [ggml_blk_size\(\)](#), [ggml_ftype_to_ggml_type\(\)](#), [ggml_is_quantized\(\)](#), [ggml_type_sizef\(\)](#)

Examples

```
ggml_type_name(GGML_TYPE_F32) # "f32"
ggml_type_name(GGML_TYPE_Q4_0) # "q4_0"
```

ggml_type_size	<i>Get Type Size in Bytes</i>
----------------	-------------------------------

Description

Returns the size in bytes for all elements in a block for a given type.

Usage

```
ggml_type_size(type)
```

Arguments

type GGML type constant (e.g., GGML_TYPE_F32)

Value

Size in bytes

ggml_type_sizef	<i>Get Type Size as Float</i>
-----------------	-------------------------------

Description

Returns the size in bytes of a GGML type as a floating-point number. For quantized types, this is the average bytes per element.

Usage

```
ggml_type_sizef(type)
```

Arguments

type GGML type constant

Value

Numeric size in bytes (can be fractional for quantized types)

See Also

Other type_system: [ggml_blk_size\(\)](#), [ggml_ftype_to_ggml_type\(\)](#), [ggml_is_quantized\(\)](#), [ggml_type_name\(\)](#)

Examples

```
ggml_type_sizef(GGML_TYPE_F32)  # 4.0
ggml_type_sizef(GGML_TYPE_F16)  # 2.0
```

ggml_unary_op_name	<i>Get Unary Operation Name</i>
--------------------	---------------------------------

Description

Returns the string name of a GGML unary operation.

Usage

```
ggml_unary_op_name(op)
```

Arguments

op GGML unary operation constant

Value

Character string with operation name

See Also

Other op_info: [ggml_get_unary_op\(\)](#), [ggml_op_desc\(\)](#), [ggml_op_name\(\)](#), [ggml_op_symbol\(\)](#)

ggml_unfreeze_weights *Unfreeze Layer Weights*

Description

Sets trainable = TRUE on layers. Accepts optional from / to to unfreeze a range of layers, or layer_names to unfreeze by name. If none are provided, all layers are unfrozen.

Usage

```
ggml_unfreeze_weights(
  model,
  from = 1L,
  to = length(model$layers),
  layer_names = NULL,
  ...
)
```

Arguments

model	A model object (ggml_sequential_model or ggml_functional_model)
from	Integer index of the first layer to unfreeze (default: 1)
to	Integer index of the last layer to unfreeze (default: last layer)
layer_names	Character vector of layer names to unfreeze (overrides from/to)
...	Additional arguments passed to methods

Value

The model with selected layers unfrozen.

Examples

```
model <- ggml_model_sequential() |>
  ggml_layer_dense(64, activation = "relu") |>
  ggml_layer_dense(10, activation = "softmax")

model <- ggml_freeze_weights(model)
model <- ggml_unfreeze_weights(model, from = 2) # unfreeze last layer only
```

ggml_unmarshal_model	<i>Unmarshal a ggmlR model from an in-memory container</i>
----------------------	--

Description

Reconstructs a ggmlR model previously produced by [ggml_marshal_model](#). Validates the container’s format tag, schema version, and (if **digest** is installed) the SHA-256 checksum of the payload before deserializing.

Usage

```
ggml_unmarshal_model(x, backend = NULL)
```

Arguments

- x A "ggmlR_marshaled" container.
- backend Backend selection passed through to [ggml_load_model](#). Default "auto".

Value

A compiled ggmlR model object (sequential or functional).

See Also

[ggml_marshal_model](#), [ggml_load_model](#)

ggml_upscale	<i>Upscale Tensor (Graph)</i>
--------------	-------------------------------

Description

Upscales tensor by multiplying ne0 and ne1 by scale factor. Supports different interpolation modes for image upscaling.

Usage

```
ggml_upscale(ctx, a, scale_factor, mode = 0L)  
  
GGML_SCALE_MODE_NEAREST  
  
GGML_SCALE_MODE_BILINEAR  
  
GGML_SCALE_MODE_BICUBIC
```


Arguments

ctx	GGML context
a	Input tensor (typically 2D or 4D for images)
scale_factor	Integer scale factor (e.g., 2 = double size)
mode	Scale mode constant (see details)

Format

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

Details

Scale mode constants:

- GGML_SCALE_MODE_NEAREST (0): Nearest neighbor interpolation - fastest, pixelated
- GGML_SCALE_MODE_BILINEAR (1): Bilinear interpolation - smooth, good balance
- GGML_SCALE_MODE_BICUBIC (2): Bicubic interpolation - smoothest, most compute

Value

Upscaled tensor with dimensions multiplied by scale_factor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
img <- ggml_new_tensor_2d(ctx, GGML_TYPE_F32, 8, 8)
ggml_set_f32(img, rnorm(64))

# Nearest neighbor (fastest, pixelated)
up_nearest <- ggml_upscale(ctx, img, 2, GGML_SCALE_MODE_NEAREST)

# Bilinear (smooth)
up_bilinear <- ggml_upscale(ctx, img, 2, GGML_SCALE_MODE_BILINEAR)

# Bicubic (smoothest)
up_bicubic <- ggml_upscale(ctx, img, 2, GGML_SCALE_MODE_BICUBIC)

graph <- ggml_build_forward_expand(ctx, up_nearest)
ggml_graph_compute(ctx, graph)
# Result is 16x16
ggml_free(ctx)
```

ggml_used_mem	<i>Get Used Memory</i>
---------------	------------------------

Description

Returns the amount of memory currently used in the context

Usage

```
ggml_used_mem(ctx)
```

Arguments

ctx GGML context

Value

Used memory in bytes

Examples

```
ctx <- ggml_init(1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
ggml_used_mem(ctx)
ggml_free(ctx)
```

ggml_version	<i>Get GGML version</i>
--------------	-------------------------

Description

Get GGML version

Usage

```
ggml_version()
```

Value

Character string with GGML version

Examples

```
ggml_version()
```

ggml_view_1d	<i>1D View with Byte Offset (Graph)</i>
--------------	---

Description

Creates a 1D view of a tensor starting at a byte offset. The view shares memory with the source tensor.

Usage

```
ggml_view_1d(ctx, a, ne0, offset = 0)
```

Arguments

ctx	GGML context
a	Source tensor
ne0	Number of elements in the view
offset	Byte offset from the start of tensor data

Value

View tensor

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 100)
# View elements 10-19 (offset = 10 * 4 bytes = 40)
v <- ggml_view_1d(ctx, a, 10, 40)
ggml_free(ctx)
```

ggml_view_2d	<i>2D View with Byte Offset (Graph)</i>
--------------	---

Description

Creates a 2D view of a tensor starting at a byte offset. The view shares memory with the source tensor.

Usage

```
ggml_view_2d(ctx, a, ne0, ne1, nb1, offset = 0)
```

Arguments

ctx	GGML context
a	Source tensor
ne0	Size of dimension 0
ne1	Size of dimension 1
nb1	Stride for dimension 1 (in bytes)
offset	Byte offset from the start of tensor data

Value

View tensor

ggml_view_3d	<i>3D View with Byte Offset (Graph)</i>
--------------	---

Description

Creates a 3D view of a tensor starting at a byte offset. The view shares memory with the source tensor.

Usage

```
ggml_view_3d(ctx, a, ne0, ne1, ne2, nb1, nb2, offset = 0)
```

Arguments

ctx	GGML context
a	Source tensor
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2
nb1	Stride for dimension 1 (in bytes)
nb2	Stride for dimension 2 (in bytes)
offset	Byte offset from the start of tensor data

Value

View tensor

ggml_view_4d	<i>4D View with Byte Offset (Graph)</i>
--------------	---

Description

Creates a 4D view of a tensor starting at a byte offset. The view shares memory with the source tensor. CRITICAL for KV-cache operations in transformers.

Usage

```
ggml_view_4d(ctx, a, ne0, ne1, ne2, ne3, nb1, nb2, nb3, offset = 0)
```

Arguments

ctx	GGML context
a	Source tensor
ne0	Size of dimension 0
ne1	Size of dimension 1
ne2	Size of dimension 2
ne3	Size of dimension 3
nb1	Stride for dimension 1 (in bytes)
nb2	Stride for dimension 2 (in bytes)
nb3	Stride for dimension 3 (in bytes)
offset	Byte offset from the start of tensor data

Value

View tensor

ggml_view_tensor	<i>View Tensor</i>
------------------	--------------------

Description

Creates a view of the tensor (shares data, no copy)

Usage

```
ggml_view_tensor(ctx, src)
```

Arguments

ctx	GGML context
src	Source tensor

Value

View tensor (shares data with src)

Examples

```
ctx <- ggml_init(16 * 1024 * 1024)
a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
view <- ggml_view_tensor(ctx, a)
# view shares data with a
ggml_free(ctx)
```

ggml_vulkan_available *Check if Vulkan support is available*

Description

Returns TRUE if the package was compiled with Vulkan support. To enable Vulkan, install libvulkan-dev and glslc, then reinstall ggmlR.

Usage

```
ggml_vulkan_available()
```

Value

Logical indicating if Vulkan is available

Examples

```
ggml_vulkan_available()
```

ggml_vulkan_backend_name
Get Vulkan backend name

Description

Returns the name of the Vulkan backend (includes device info).

Usage

```
ggml_vulkan_backend_name(backend)
```

Arguments

backend Vulkan backend pointer

Value

Character string with backend name

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
  backend <- ggml_vulkan_init(0)
  print(ggml_vulkan_backend_name(backend))
  ggml_vulkan_free(backend)
}
```

ggml_vulkan_device_caps

Get Vulkan device capabilities

Description

Returns hardware capabilities for the specified Vulkan device.

Usage

```
ggml_vulkan_device_caps(device = 0L)
```

Arguments

device Device index (0-based, default 0)

Value

Named list: coopmat_support, coopmat1_fa_support, fp16, subgroup_size, subgroup_no_shmem

ggml_vulkan_device_count

Get number of Vulkan devices

Description

Returns the number of available Vulkan-capable GPU devices.

Usage

```
ggml_vulkan_device_count()
```

Value

Integer count of Vulkan devices (0 if Vulkan not available)

Examples

```
if (ggml_vulkan_available()) {
    ggml_vulkan_device_count()
}
```

```
ggml_vulkan_device_description
    Get Vulkan device description
```

Description

Returns a human-readable description of the specified Vulkan device.

Usage

```
ggml_vulkan_device_description(device = 0L)
```

Arguments

device Device index (0-based)

Value

Character string with device description

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
    ggml_vulkan_device_description(0)
}
```

```
ggml_vulkan_device_groups
    Probe Vulkan device groups (NVLink / multi-GPU peer access)
```

Description

Enumerates Vulkan device groups (VK_KHR_device_group, also known as Linked Device Adapter / LDA) and probes whether the driver reports direct peer memory access between the physical GPUs in each group.

Usage

```
ggml_vulkan_device_groups()
```


Details

A device group with more than one device *and* peer copy/generic memory features is the prerequisite for true GPU-to-GPU transfers routed over NVLink (or PCIe peer-to-peer) through a single device-group logical device — as opposed to sharing memory as an opaque fd between independent devices, which the driver may route through host memory.

This is a diagnostic only: it does not create any long-lived device group. On a machine with a single GPU it reports zero multi-device groups. Device groups for compute are effectively an NVIDIA feature; AMD/RADV typically reports only single-device groups.

Value

A named list with `n_groups` (integer, number of device groups reported by the driver) and `report` (character, a human-readable per-group diagnostic including peer memory features). Use `cat()` on `report` to read it.

Examples

```
if (ggml_vulkan_available()) {  
  g <- ggml_vulkan_device_groups()  
  cat(g$report)  
}
```

ggml_vulkan_device_memory
Get Vulkan device memory

Description

Returns free and total memory for the specified Vulkan device.

Usage

```
ggml_vulkan_device_memory(device = 0L)
```

Arguments

device	Device index (0-based)
--------	------------------------

Value

Named list with 'free' and 'total' memory in bytes

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
  mem <- ggml_vulkan_device_memory(0)
  cat("Free:", mem$free / 1e9, "GB\n")
  cat("Total:", mem$total / 1e9, "GB\n")
}
```

ggml_vulkan_free	<i>Free Vulkan backend</i>
------------------	----------------------------

Description

Releases resources associated with the Vulkan backend.

Usage

```
ggml_vulkan_free(backend)
```

Arguments

backend	Vulkan backend pointer from ggml_vulkan_init()
---------	--

Value

NULL (invisible)

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
  backend <- ggml_vulkan_init(0)
  ggml_vulkan_free(backend)
}
```

ggml_vulkan_hard_exit_available	<i>Is the Vulkan hard-exit path compiled in?</i>
---------------------------------	--

Description

Reports whether this build of ggmlR was compiled with `-DGGML_VK_HARD_EXIT`, which is what makes `ggml_vulkan_shutdown(hard = TRUE)` actually call `_exit()`.

Usage

```
ggml_vulkan_hard_exit_available()
```

Details

The hard-exit path is **disabled by default**: CRAN Repository Policy forbids a package from terminating the user's R session, so the released package must not link `_exit()`. Builds from source can opt in with R CMD INSTALL . --configure-args="--enable-hard-exit" (on Windows, set `Sys.setenv(GGML_VK_HARD_EXIT = "1")` before installing, because R there ignores `configure.args`).

When it is not compiled in, `ggml_vulkan_shutdown(hard = TRUE)` performs the normal teardown and emits a [warning](#) rather than silently ignoring the request.

Value

TRUE if the hard-exit path is available, otherwise FALSE.

See Also

[ggml_vulkan_shutdown](#)

Examples

```
ggml_vulkan_hard_exit_available()
```

ggml_vulkan_init	<i>Initialize Vulkan backend</i>
------------------	----------------------------------

Description

Creates a Vulkan backend for the specified device. The backend must be freed with `ggml_vulkan_free()` when done.

Usage

```
ggml_vulkan_init(device = 0L)
```

Arguments

device Device index (0-based, default 0)

Value

Vulkan backend pointer

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
  backend <- ggml_vulkan_init(0)
  print(ggml_vulkan_backend_name(backend))
  ggml_vulkan_free(backend)
}
```

`ggml_vulkan_is_backend`*Check if backend is Vulkan*

Description

Returns TRUE if the given backend is a Vulkan backend.

Usage

```
ggml_vulkan_is_backend(backend)
```

Arguments

backend	Backend pointer
---------	-----------------

Value

Logical indicating if backend is Vulkan

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {  
  vk_backend <- ggml_vulkan_init(0)  
  cpu_backend <- ggml_backend_cpu_init()  
  
  ggml_vulkan_is_backend(vk_backend) # TRUE  
  ggml_vulkan_is_backend(cpu_backend) # FALSE  
  
  ggml_vulkan_free(vk_backend)  
  ggml_backend_free(cpu_backend)  
}
```

`ggml_vulkan_list_devices`*List all Vulkan devices*

Description

Returns detailed information about all available Vulkan devices.

Usage

```
ggml_vulkan_list_devices()
```

Value

List of device information (index, name, memory)

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() > 0) {
  devices <- ggml_vulkan_list_devices()
  print(devices)
}
```

ggml_vulkan_p2p_selftest

Opaque-fd device-to-device P2P self-test

Description

Exercises the VK_KHR_external_memory_fd transport used by Vulkan tensor parallelism to move data between GPUs. A byte pattern is written on the source device, its memory is exported as an opaque fd and imported on the destination device, copied into a local buffer there, then read back and compared.

Usage

```
ggml_vulkan_p2p_selftest(
  src_device,
  dst_device,
  bytes = 64L * 1024L * 1024L,
  iters = 50L,
  transport = c("host-staging", "opaque-fd", "device-group")
)
```

Arguments

src_device	Source GPU index (0-based).
dst_device	Destination GPU index (0-based). Equal to src_device for a loopback sanity check.
bytes	Transfer size in bytes (default 64 MiB).
iters	Number of copies to time for the bandwidth measurement (default 50).
transport	Cross-device transport to exercise: "host-staging" (default, portable device->host->device copy — correct on every driver, PCIe + RAM bounded), "opaque-fd" (VK_KHR_external_memory_fd P2P; works on AMD/RADV but does NOT alias memory cross-device on the NVIDIA proprietary driver — reads back as zeros), or "device-group" (experimental NVIDIA LDA). host-staging is the transport Stage E3 uses.

Details

When `src_device == dst_device` the test runs in *loopback* mode (export and import on the same GPU) — a sanity check of the fd mechanism that touches no inter-device link. When the devices differ it runs *cross-device*: after verifying correctness it times `iters` device-to-device copies and reports the achieved bandwidth.

Interpreting bandwidth: a measured rate above the PCIe 3.0 x16 ceiling (~16 GB/s) is *empirical* evidence that a faster physical link (e.g. NVLink) carried the bytes. It is **not** a claim that Vulkan used an NVLink API — Vulkan exposes no call to query the route, so the conclusion is inferred from the rate alone, never asserted from the API.

Value

A named list: `status` (integer, 0 = data verified, <0 = failure), `gbps` (numeric, measured cross-device bandwidth; 0 for loopback or on failure) and `report` (character diagnostic, incl. the NVLink-vs-PCIe inference).

Examples

```
if (ggml_vulkan_available()) {
  # loopback sanity check on device 0
  r <- ggml_vulkan_p2p_selftest(0L, 0L)
  cat(r$report)
  # cross-device P2P (requires >= 2 GPUs)
  if (ggml_vulkan_device_count() >= 2) {
    r <- ggml_vulkan_p2p_selftest(0L, 1L)
    cat(r$report)
  }
}
```

<code>ggml_vulkan_shutdown</code>	<i>Explicitly tear down the Vulkan instance and devices</i>
-----------------------------------	---

Description

Waits for all Vulkan devices to go idle, then destroys the devices and the Vulkan instance. Call this at the end of a script that used multiple GPUs (tensor / pipeline parallelism), *before* the process exits.

Usage

```
ggml_vulkan_shutdown(hard = FALSE, status = 0L)
```

Arguments

<code>hard</code>	Logical. If TRUE, call <code>_exit(status)</code> after teardown to skip all exit handlers and guarantee no exit-time segfault; never returns. Requires a build with <code>--enable-hard-exit</code> (see above); otherwise it warns and returns normally. Default FALSE (safe to call mid-session).
<code>status</code>	Integer process exit code used only when <code>hard = TRUE</code> (passed to <code>_exit()</code>). Defaults to 0L (success). If you call this from an error path (e.g. a <code>tryCatch</code> handler after a stage failed), pass a non-zero status so the failed run does not exit 0 and mask the failure. Ignored when <code>hard = FALSE</code> .

Details

Why this exists: with several Vulkan devices active, R's own teardown at process exit can run *after* the Vulkan loader / Mesa ICD libraries have been unmapped, so the device destructors call into unmapped memory and the process crashes with a segfault *after* your results were already printed. Calling `ggml_vulkan_shutdown()` while the process is still alive runs the teardown while the loader is still mapped, avoiding that crash. It is only needed for exit cleanliness — your computed results are unaffected either way.

Idempotent and a no-op if Vulkan was never initialized. After this call the next Vulkan operation transparently re-initializes the instance from scratch, so it is safe to call mid-session too.

The plain teardown alone does not always win the exit-time race: with backends created inside pipeline / tensor-parallel forwards, the loader-static-destruction segfault (address ending `f1ba0`) can still fire flakily *after* your results are printed, because no R exit hook runs before R unmaps the loader. For a guaranteed-clean exit, pass `hard = TRUE` as the **last** statement of a standalone script: after teardown it calls `_exit(0)`, terminating the process immediately without running R's / the C runtime's exit handlers or unmapping any shared library — so there is no loader teardown phase left to crash. Because it bypasses R's normal shutdown, only use `hard = TRUE` at the very end of a script, never mid-session or in a package/interactive context.

Value

Invisibly NULL. Does not return when `hard = TRUE` *and* the hard-exit path is compiled in (see the section above).

Hard exit is opt-in at build time

The `_exit()` path is **compiled out by default**, because CRAN Repository Policy forbids a package from terminating the user's R session. In such a build `hard = TRUE` performs the normal teardown and emits a [warning](#) — it is never silently ignored — so the exit-time race described above can still fire. To compile it in, build from source with R CMD INSTALL . --configure-args="--enable-hard-exit" (on Windows set `Sys.setenv(GGML_VK_HARD_EXIT = "1")` first, since R ignores `configure.args` there). Check the current build with [ggml_vulkan_hard_exit_available\(\)](#).

See Also

[ggml_vulkan_hard_exit_available](#)

Examples

```

if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  W <- matrix(rnorm(2048 * 64), nrow = 2048)
  X <- matrix(rnorm(4 * 64), nrow = 4)
  Y <- ggml_vulkan_split_mul_mat(W, X, n_devices = 2)
  ggml_vulkan_shutdown() # clean up before the script exits
}

```

ggml_vulkan_split_buffer_type

Create a Vulkan tensor-split buffer type

Description

Builds (or fetches from a cache) a ggml buffer type that row-splits a weight tensor across `n_devices` GPUs for tensor parallelism. This is the factory the graph allocator talks to (Stage E4): a weight placed in this buffer type has its rows distributed across the devices, each holding a contiguous band (see [ggml_vulkan_split_row_ranges](#)). Non-split tensors fall back to `main_device`.

Usage

```

ggml_vulkan_split_buffer_type(
  n_devices = ggml_vulkan_device_count(),
  main_device = 0L,
  weights = NULL,
  device_ids = NULL,
  transport = c("host-staging", "opaque-fd", "device-group"),
  probe = NULL
)

```

Arguments

<code>n_devices</code>	Number of GPUs to split across (default: all available). Ignored if <code>device_ids</code> is given (its length is used instead).
<code>main_device</code>	Device index (0-based) holding non-split fallbacks (default 0).
<code>weights</code>	Optional numeric vector of length <code>n_devices</code> giving the relative row share per device (e.g. <code>c(3, 1)</code>). <code>NULL</code> (default) splits evenly.
<code>device_ids</code>	Optional integer vector of physical GPU indices (0-based) the split slots map onto, e.g. <code>c(2, 3)</code> . <code>NULL</code> (default) uses <code>0..n_devices-1</code> . When supplied, <code>n_devices</code> is its length.
<code>transport</code>	Cross-device gather transport for the eventual compute: "host-staging" (default), "opaque-fd" or "device-group".
<code>probe</code>	A numeric vector <code>c(N, K)</code> describing a probe weight tensor (N rows, K columns) whose split allocation size to report, or <code>NULL</code> (default) to skip. Useful for inspecting how much total VRAM a given weight would occupy across the split.

Details

The buffer type is cached in C and lives for the rest of the session; the returned external pointer has no finalizer and must not be freed.

Value

A named list: ptr (external pointer to the cached buffer type), name (its config-derived name) and alloc_size (sum of the padded per-device slice sizes for probe, in bytes; 0 if no probe).

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  bt <- ggml_vulkan_split_buffer_type(n_devices = 2, probe = c(2048, 64))
  bt$name
  bt$alloc_size # total bytes for a 2048x64 f32 weight across 2 devices
}
```

ggml_vulkan_split_mul_mat

Tensor-parallel matrix multiply across Vulkan devices

Description

Computes $Y = X \%* \% t(W)$ with the rows of the weight matrix W split across $n_devices$ GPUs (true tensor parallelism), the activations X broadcast to every device, and the resulting column slices gathered back into a single matrix. This is Stage E3 of the Vulkan tensor-parallelism work: it exercises the row-split + per-device compute + cross-device gather path end to end on real hardware.

Usage

```
ggml_vulkan_split_mul_mat(
  W,
  X,
  n_devices = ggml_vulkan_device_count(),
  weights = NULL,
  device_ids = NULL,
  transport = c("host-staging", "opaque-fd", "device-group")
)
```

Arguments

W	Weight matrix, $N \times K$ (N output features, K input features). Its rows are split across the devices.
X	Activation matrix, $M \times K$ (M samples, K input features), broadcast to every device.

<code>n_devices</code>	Number of GPUs to split across (default: all available). Ignored if <code>device_ids</code> is given (its length is used instead).
<code>weights</code>	Optional numeric vector of length <code>n_devices</code> giving the relative share of <code>W</code> 's rows per device (e.g. <code>c(3, 1)</code> for a 3:1 split). <code>NULL</code> (default) splits the rows evenly.
<code>device_ids</code>	Optional integer vector of physical GPU indices (0-based) to split across, e.g. <code>c(2, 3)</code> for the second replica of a TPxDP layout. <code>NULL</code> (default) uses devices <code>0..n_devices-1</code> . When supplied, <code>n_devices</code> is taken from its length.
<code>transport</code>	Cross-device gather transport: "host-staging" (default, portable and correct on every driver), "opaque-fd" or "device-group" (see ggml_vulkan_p2p_selftest).

Details

Each device owns a contiguous band of `W`'s rows (via the same row-split math as [ggml_vulkan_split_row_ranges](#)) and produces the matching band of `Y`'s columns; the bands are gathered through the transport (host staging by default). This is a validation / correctness entry point, not a hot inference path — it stands up one backend per device per call and round-trips through host memory.

Value

The $M \times N$ result matrix, equal to `X %*% t(W)` up to the GPU's floating-point accumulation. Errors if the split compute fails; the C-level diagnostic is attached as attribute "report".

Examples

```
if (ggml_vulkan_available() && ggml_vulkan_device_count() >= 2) {
  W <- matrix(rnorm(8 * 4), nrow = 8) # 8 output features x 4 inputs
  X <- matrix(rnorm(3 * 4), nrow = 3) # 3 samples x 4 inputs
  Y <- ggml_vulkan_split_mul_mat(W, X, n_devices = 2)
  max(abs(Y - X %*% t(W)))           # ~1e-3 (GPU f16 accumulation)
}
```

`ggml_vulkan_split_row_ranges`

Compute tensor-parallel row-split ranges

Description

Pure arithmetic helper for the Vulkan tensor-parallel split buffer type: given a number of tensor rows and a per-device weight vector, returns the half-open row range `[row_low, row_high)` owned by each device. Row boundaries are rounded down to a fixed granularity, and the last device always covers up to `nrows`, so the ranges are contiguous, non-overlapping and cover every row exactly once.

Usage

```
ggml_vulkan_split_row_ranges(nrows, n_devices, weights = NULL)
```

Arguments

nrows	Number of rows in the tensor (integer).
n_devices	Number of devices to split across (integer, >= 1).
weights	Optional numeric vector of length n_devices giving the relative share of rows per device. NULL (default) splits evenly.

Details

This touches no GPU and is exposed mainly to verify the split logic. It is the math behind row-split tensor parallelism (a weight matrix distributed across several GPUs), independent of any actual multi-GPU allocation.

Value

A named list with row_low and row_high: numeric vectors of length n_devices holding 0-based, half-open row ranges.

Examples

```
if (ggml_vulkan_available()) {
  # 4096 rows across 2 devices, evenly
  ggml_vulkan_split_row_ranges(4096L, 2L)
  # weighted 3:1
  ggml_vulkan_split_row_ranges(4096L, 2L, weights = c(3, 1))
}
```

ggml_vulkan_status	<i>Print Vulkan status</i>
--------------------	----------------------------

Description

Prints information about Vulkan availability and devices.

Usage

```
ggml_vulkan_status()
```

Value

NULL (invisible), prints status to console

Examples

```
ggml_vulkan_status()
```

ggml_win_part	<i>Window Partition (Graph)</i>
---------------	---------------------------------

Description

Partitions a tensor into non-overlapping windows of size w.

Usage

ggml_win_part(ctx, a, w)

Arguments

ctx	GGML context
a	Input tensor
w	Window size

Value

Partitioned tensor

ggml_win_unpart	<i>Window Un-partition (Graph)</i>
-----------------	------------------------------------

Description

Reassembles windowed partitions produced by [ggml_win_part](#).

Usage

ggml_win_unpart(ctx, a, w0, h0, w)

Arguments

ctx	GGML context
a	Input tensor
w0	Original width
h0	Original height
w	Window size

Value

Un-partitioned tensor

ggml_with_temp_ctx	<i>Execute with Temporary Context</i>
--------------------	---------------------------------------

Description

Creates a temporary context, executes code, and frees it automatically. Useful when you need to create large temporary tensors.

Usage

```
ggml_with_temp_ctx(mem_size, expr)
```

Arguments

mem_size	Context memory size in bytes
expr	Expression to evaluate with the temporary context

Value

Result of the expression

Examples

```
# Create tensors in temporary context
result <- ggml_with_temp_ctx(1024 * 1024, {
  a <- ggml_new_tensor_1d(ctx, GGML_TYPE_F32, 10)
  ggml_set_f32(a, 1:10)
  ggml_get_f32(a)
})
```

gguf_free	<i>Free GGUF Resources</i>
-----------	----------------------------

Description

Explicitly frees the internal GGUF context. Called automatically by the garbage collector, but can be called manually to release memory sooner.

Usage

```
gguf_free(x)
```

Arguments

x	A gguf object.
---	----------------

Value

Called for its side effect (releases the GGUF context); invisibly returns NULL.

gguf_load	<i>Load a GGUF File</i>
-----------	-------------------------

Description

Opens a GGUF file and reads its metadata. By default also reads tensor data into memory; with `meta_only = TRUE` only the header and key-value metadata are read (no tensor data is allocated), which is cheap and enough for inspecting architecture / type fields. Returns an S3 object of class "gguf" wrapping the internal pointer.

Usage

```
gguf_load(path, meta_only = FALSE)
```

Arguments

- `path` Path to a .gguf file.
- `meta_only` If TRUE, read only the header and metadata without allocating tensor data. Metadata, tensor names and tensor info remain available; `gguf_tensor_data` is not (reload with `meta_only = FALSE`). Default FALSE.

Value

An object of class "gguf".

gguf_metadata	<i>Get GGUF Metadata</i>
---------------	--------------------------

Description

Returns all key-value metadata pairs from a GGUF file as a named list.

Usage

```
gguf_metadata(x)
```

Arguments

- `x` A gguf object from `gguf_load`.

Value

A named list of metadata values.

gguf_tensor_data	<i>Extract Tensor Data</i>
------------------	----------------------------

Description

Dequantizes (if needed) and returns tensor weights as an R numeric array with dimensions matching the tensor shape.

Usage

```
gguf_tensor_data(x, name)
```

Arguments

- | | |
|------|--------------------------|
| x | A gguf object. |
| name | Tensor name (character). |

Value

A numeric array.

gguf_tensor_info	<i>Get Tensor Info</i>
------------------	------------------------

Description

Returns name, shape, type, and size in bytes for a single tensor.

Usage

```
gguf_tensor_info(x, name)
```

Arguments

- | | |
|------|--------------------------|
| x | A gguf object. |
| name | Tensor name (character). |

Details

When the file was opened with `meta_only = TRUE`, the per-dimension shape is NA (the public GGUF API does not expose tensor dimensions without allocating tensors); name, type and size_bytes are still returned.

Value

A list with elements name, shape, type, size_bytes.

gguf_tensor_names	List Tensor Names in a GGUF File
-------------------	----------------------------------

Description

List Tensor Names in a GGUF File

Usage

gguf_tensor_names(x)

Arguments

x A gguf object.

Value

Character vector of tensor names.

glance.ggmlr_parsnip_model	One-row summary of a fitted ggml parsnip model
----------------------------	--

Description

Returns a single-row [tibble][tibble::tibble] summarising the fitted model, in broom 'glance()' style.

Usage

```
## S3 method for class 'ggmlr_parsnip_model'
glance(x, ...)
```

Arguments

x A fitted 'ggmlr_parsnip_model' (the engine object inside a parsnip fit; e.g. from 'extract_fit_engine()').

... Unused; for generic compatibility.

Value

A one-row tibble with columns: 'mode', 'n_features', 'n_layers', 'total_params', 'optimizer', 'loss', 'backend', 'epochs', 'fit_time' (wall seconds) and 'final_loss' (last training loss, 'NA' if no history).

Examples

```
ggml_set_n_threads(1L) # deterministic, single OpenMP pool
spec <- parsnip::mlp(hidden_units = 8L, epochs = 3L) |>
  parsnip::set_engine("ggml", backend = "cpu") |>
  parsnip::set_mode("regression")
fit_obj <- parsnip::fit(spec, mpg ~ ., data = mtcars)
generics::glance(parsnip::extract_fit_engine(fit_obj))
```

iq2xs_free_impl	<i>Free IQ2 Quantization Tables</i>
-----------------	-------------------------------------

Description

Frees lookup tables for IQ2 quantization types.

Usage

```
iq2xs_free_impl(type)
```

Arguments

type GGML type constant

Value

NULL invisibly

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

iq2xs_init_impl	<i>Initialize IQ2 Quantization Tables</i>
-----------------	---

Description

Initializes lookup tables for IQ2 quantization types. Must be called before using iq2_xxs, iq2_xs, or iq2_s quantization.

Usage

iq2xs_init_impl(type)

Arguments

type GGML type constant (e.g., GGML_TYPE_IQ2_XXS())

Value

NULL invisibly

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

iq3xs_free_impl	<i>Free IQ3 Quantization Tables</i>
-----------------	-------------------------------------

Description

Frees lookup tables for IQ3 quantization types.

Usage

iq3xs_free_impl(grid_size)

Arguments

grid_size Grid size for IQ3

Value

NULL invisibly

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

iq3xs_init_impl	<i>Initialize IQ3 Quantization Tables</i>
-----------------	---

Description

Initializes lookup tables for IQ3 quantization types. Must be called before using iq3_xxs or iq3_s quantization.

Usage

```
iq3xs_init_impl(grid_size)
```

Arguments

grid_size	Grid size for IQ3 (typically 256)
-----------	-----------------------------------

Value

NULL invisibly

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

lr_scheduler_cosine *Cosine-annealing learning rate scheduler*

Description

Varies the learning rate following a cosine curve from lr_max down to lr_min over T_max steps. Restarts (SGDR-style) if restart = TRUE.

Usage

```
lr_scheduler_cosine(optimizer, T_max, lr_min = 0, restart = FALSE)
```

Arguments

optimizer	Optimizer environment.
T_max	Number of steps for one cosine cycle.
lr_min	Minimum learning rate (default 0).
restart	Logical; if TRUE restart after T_max steps.

Value

An lr_scheduler_cosine environment

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
opt <- optimizer_adam(list(w = w), lr = 0.1)
sch <- lr_scheduler_cosine(opt, T_max = 50L)
for (epoch in 1:50) sch$step()
```

lr_scheduler_step *Step-decay learning rate scheduler*

Description

Multiplies the optimizer learning rate by gamma every step_size calls to \$step().

Usage

```
lr_scheduler_step(optimizer, step_size, gamma = 0.1)
```

Arguments

- optimizer An ag_optimizer_adam or ag_optimizer_sgd environment.
- step_size Decay every this many steps (epochs).
- gamma Multiplicative decay factor (default 0.1).

Value

An lr_scheduler_step environment

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
opt <- optimizer_adam(list(w = w), lr = 0.1)
sch <- lr_scheduler_step(opt, step_size = 10L, gamma = 0.5)
for (epoch in 1:30) sch$step()
opt$lr # 0.1 * 0.5^3 = 0.0125
```

nn_topo_sort	<i>Topologically sort nodes reachable from output nodes</i>
--------------	---

Description

Topologically sort nodes reachable from output nodes

Usage

```
nn_topo_sort(outputs)
```

Arguments

- outputs List of output ggml_tensor_node objects

Value

Named list: nodes in topological order (inputs first, outputs last)

Value

A named list where names are input tensor names and values are integer vectors of dimension sizes (-1 for dynamic dimensions).

onnx_load	<i>Load an ONNX model</i>
-----------	---------------------------

Description

Parses an .onnx file, builds a ggml computation graph, and allocates tensors on the specified device. Weights are loaded via memory-mapped file (zero-copy where possible).

Usage

```
onnx_load(  
    path,  
    device = NULL,  
    input_shapes = NULL,  
    n_threads = NULL,  
    dtype = "f32"  
)
```

Arguments

path	Path to .onnx file.
device	Backend device: "vulkan" (default if available) or "cpu".
input_shapes	Optional named list of integer vectors specifying fixed shapes for inputs with dynamic dimensions. Names must match input tensor names. Each shape must include all dimensions including batch, e.g. <code>list(image = c(1L, 3L, 224L, 224L))</code> . Required when the model has dynamic dimensions and no default shape.
n_threads	Number of CPU threads. NULL (default) reads <code>getOption("ggmlR.n_threads")</code> ; if that is also unset, uses <code>parallel::detectCores() - 1</code> (minimum 1).
dtype	Weight precision: "f32" (default) or "f16". When "f16", large weight tensors (≥ 256 elements) are stored in half-precision for faster Vulkan compute and lower VRAM usage. Small tensors (bias, scalars, batch-norm params) remain in F32 for numerical stability. Inputs and outputs are always F32.

Value

An opaque model object (external pointer) for use with `onnx_run()`, `onnx_summary()`, and `onnx_inputs()`.

onnx_run	<i>Run ONNX model inference</i>
----------	---------------------------------

Description

Run ONNX model inference

Usage

onnx_run(model, inputs)

Arguments

- | | |
|--------|--|
| model | An onnx_model object from onnx_load(). |
| inputs | A named list of numeric vectors/matrices. Names must match the model’s input tensor names. Use onnx_inputs() to see expected names and shapes. |

Value

A named list of output tensors (numeric vectors with dim attributes for multi-dimensional outputs).

onnx_summary	<i>ONNX model summary</i>
--------------	---------------------------

Description

Returns metadata about a loaded ONNX model.

Usage

onnx_summary(model)

Arguments

- | | |
|-------|--|
| model | An onnx_model object from onnx_load(). |
|-------|--|

Value

A list with ir_version, opset_version, producer, graph_name, n_nodes, n_initializers, and ops.

optimizer_adam	Create an Adam optimizer
----------------	--------------------------

Description

Create an Adam optimizer

Usage

```
optimizer_adam(params, lr = 0.001, beta1 = 0.9, beta2 = 0.999, eps = 1e-08)
```

Arguments

- | | |
|--------|-------------------------------------|
| params | Named list of ag_param tensors |
| lr | Learning rate (default 1e-3) |
| beta1 | First moment decay (default 0.9) |
| beta2 | Second moment decay (default 0.999) |
| eps | Stability constant (default 1e-8) |

Value

An optimizer environment

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
opt <- optimizer_adam(list(w = w), lr = 1e-3)
```

optimizer_sgd	Create an SGD optimizer
---------------	-------------------------

Description

Create an SGD optimizer

Usage

```
optimizer_sgd(params, lr = 0.01, momentum = 0)
```

Arguments

- | | |
|----------|--------------------------------|
| params | Named list of ag_param tensors |
| lr | Learning rate (default 0.01) |
| momentum | Momentum factor (default 0) |

Value

An optimizer environment

Examples

```
w <- ag_param(matrix(runif(4), 2, 2))
opt <- optimizer_sgd(list(w = w), lr = 0.01)
```

plot.ggml_history	<i>Plot training history</i>
-------------------	------------------------------

Description

Plots loss and accuracy curves over epochs.

Usage

```
## S3 method for class 'ggml_history'
plot(x, ...)
```

Arguments

x	A ggml_history object
...	Additional arguments (ignored)

Value

The history object (invisibly).

predict.ggml_sequential_model	<i>Predict with a Trained Model</i>
-------------------------------	-------------------------------------

Description

Generates predictions from a trained model. Uses the standard R [predict](#) generic for compatibility with keras3 and the broader R ecosystem.

Usage

```
## S3 method for class 'ggml_sequential_model'
predict(object, x, batch_size = 32L, ...)

## S3 method for class 'ggml_functional_model'
predict(object, x, batch_size = 32L, ...)
```

Arguments

object	A trained model object.
x	Input data (matrix, array, or list for multi-input models).
batch_size	Batch size for inference (default 32).
...	Additional arguments (ignored).

Value

Matrix of predictions.

print.ag_tensor	<i>Print method for ag_tensor</i>
-----------------	-----------------------------------

Description

Print method for ag_tensor

Usage

```
## S3 method for class 'ag_tensor'  
print(x, ...)
```

Arguments

x	An ag_tensor
...	Ignored

Value

The input x, returned invisibly (called for its side effect of printing).

print.ggml_functional_model	<i>Print method for ggml_functional_model</i>
-----------------------------	---

Description

Print method for ggml_functional_model

Usage

```
## S3 method for class 'ggml_functional_model'  
print(x, ...)
```

Arguments

<code>x</code>	A <code>ggml_functional_model</code> object
<code>...</code>	Additional arguments (ignored)

Value

The model object (invisibly).

<code>print.ggml_history</code>	<i>Print method for ggml_history</i>
---------------------------------	--------------------------------------

Description

Print method for `ggml_history`

Usage

```
## S3 method for class 'ggml_history'
print(x, ...)
```

Arguments

<code>x</code>	A <code>ggml_history</code> object
<code>...</code>	Additional arguments (ignored)

Value

The history object (invisibly).

<code>print.ggml_sequential_model</code>	<i>Print method for ggml_sequential_model</i>
--	---

Description

Prints a summary of the model architecture including layer types, output shapes, and parameter counts.

Usage

```
## S3 method for class 'ggml_sequential_model'
print(x, ...)
```

Arguments

- x A ggml_sequential_model object
- ... Additional arguments (ignored)

Value

The model object (invisibly).

print.onnx_model	<i>Print ONNX model summary</i>
------------------	---------------------------------

Description

Print ONNX model summary

Usage

```
## S3 method for class 'onnx_model'  
print(x, ...)
```

Arguments

- x An onnx_model object.
- ... Ignored.

Value

Invisibly returns x.

quantize_iq2_xxs	<i>Quantize Data (IQ)</i>
------------------	---------------------------

Description

Quantizes float data to IQ format. IQ formats require importance matrix initialization before use (see iq2xs_init_impl, iq3xs_init_impl).

Usage

```
quantize_iq2_xxs(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq2_xs(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq2_s(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq3_xxs(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq3_s(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq1_s(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq1_m(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq4_n1(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_iq4_xs(src_data, n_rows, n_per_row, imatrix = NULL)
```

Arguments

src_data	Numeric vector of float values to quantize
n_rows	Number of rows
n_per_row	Number of elements per row
imatrix	Optional importance matrix (numeric vector or NULL)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_mxfp4	<i>Quantize Data (MXFP4)</i>
----------------	------------------------------

Description

Quantizes float data to MXFP4 (microscaling FP4) format.

Usage

quantize_mxfp4(src_data, n_rows, n_per_row, imatrix = NULL)

Arguments

- src_data Numeric vector of float values to quantize
- n_rows Number of rows
- n_per_row Number of elements per row
- imatrix Optional importance matrix (numeric vector or NULL)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_nvfp4	<i>Quantize Data (NVFP4)</i>
----------------	------------------------------

Description

Quantizes float data to NVFP4 format (NVIDIA FP4 with UE4M3 per-sub-block scale).

Usage

quantize_nvfp4(src_data, n_rows, n_per_row, imatrix = NULL)

Arguments

- src_data Numeric vector of float values to quantize
- n_rows Number of rows
- n_per_row Number of elements per row (must be multiple of 64)
- imatrix Optional importance matrix (currently ignored)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_q1_0	<i>Quantize Data (Q1_0)</i>
---------------	-----------------------------

Description

Quantizes float data to Q1_0 format (1-bit-per-weight sign quantization).

Usage

```
quantize_q1_0(src_data, n_rows, n_per_row, imatrix = NULL)
```

Arguments

src_data	Numeric vector of float values to quantize
n_rows	Number of rows
n_per_row	Number of elements per row (must be multiple of 128)
imatrix	Optional importance matrix (currently ignored)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_q2_K	<i>Quantize Data (K-quants)</i>
---------------	---------------------------------

Description

Quantizes float data to K-quant format with optional importance matrix. K-quants provide better quality/size tradeoffs than basic quants.

Usage

```
quantize_q2_K(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q3_K(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q4_K(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q5_K(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q6_K(src_data, n_rows, n_per_row, imatrix = NULL)
```

Arguments

src_data	Numeric vector of float values to quantize
n_rows	Number of rows
n_per_row	Number of elements per row
imatrix	Optional importance matrix (numeric vector or NULL)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_q4_0	<i>Quantize Data (Q4_0)</i>
---------------	-----------------------------

Description

Quantizes float data to Q4_0 format with optional importance matrix.

Usage

```
quantize_q4_0(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q4_1(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q5_0(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q5_1(src_data, n_rows, n_per_row, imatrix = NULL)
quantize_q8_0(src_data, n_rows, n_per_row, imatrix = NULL)
```

Arguments

src_data	Numeric vector of float values to quantize
n_rows	Number of rows
n_per_row	Number of elements per row
imatrix	Optional importance matrix (numeric vector or NULL)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_row_iq3_xxs_ref

Quantize Row Reference (IQ)

Description

Basic row-level IQ quantization.

Usage

```
quantize_row_iq3_xxs_ref(src_data, n_elements)
```

```
quantize_row_iq4_n1_ref(src_data, n_elements)
```

```
quantize_row_iq4_xs_ref(src_data, n_elements)
```

```
quantize_row_iq3_s_ref(src_data, n_elements)
```

```
quantize_row_iq2_s_ref(src_data, n_elements)
```

Arguments

src_data Numeric vector of float values to quantize

n_elements Number of elements to quantize

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_row_mxfp4_ref
<i>Quantize Row Reference (MXFP4)</i>

Description

Basic row-level MXFP4 quantization.

Usage

quantize_row_mxfp4_ref(src_data, n_elements)

Arguments

src_data	Numeric vector of float values to quantize
n_elements	Number of elements to quantize

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_row_q2_K_ref	<i>Quantize Row Reference (K-quants)</i>
-----------------------	--

Description

Basic row-level K-quant quantization without importance matrix.

Usage

quantize_row_q2_K_ref(src_data, n_elements)

quantize_row_q3_K_ref(src_data, n_elements)

quantize_row_q4_K_ref(src_data, n_elements)

```
quantize_row_q5_K_ref(src_data, n_elements)
```

```
quantize_row_q6_K_ref(src_data, n_elements)
```

```
quantize_row_q8_K_ref(src_data, n_elements)
```

Arguments

src_data	Numeric vector of float values to quantize
n_elements	Number of elements to quantize

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_row_q4_0_ref *Quantize Row Reference (Basic)*

Description

Basic row-level quantization without importance matrix. These are reference implementations.

Usage

```
quantize_row_q4_0_ref(src_data, n_elements)
```

```
quantize_row_q4_1_ref(src_data, n_elements)
```

```
quantize_row_q5_0_ref(src_data, n_elements)
```

```
quantize_row_q5_1_ref(src_data, n_elements)
```

```
quantize_row_q8_0_ref(src_data, n_elements)
```

```
quantize_row_q8_1_ref(src_data, n_elements)
```

Arguments

src_data	Numeric vector of float values to quantize
n_elements	Number of elements to quantize

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

[quantize_row_tq1_0_ref](#)

Quantize Row Reference (Ternary)

Description

Basic row-level ternary quantization.

Usage

```
quantize_row_tq1_0_ref(src_data, n_elements)
```

```
quantize_row_tq2_0_ref(src_data, n_elements)
```

Arguments

src_data	Numeric vector of float values to quantize
n_elements	Number of elements to quantize

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_tq1_0\(\)](#)

quantize_tq1_0	<i>Quantize Data (Ternary)</i>
----------------	--------------------------------

Description

Quantizes float data to ternary format with optional importance matrix.

Usage

```
quantize_tq1_0(src_data, n_rows, n_per_row, imatrix = NULL)

quantize_tq2_0(src_data, n_rows, n_per_row, imatrix = NULL)
```

Arguments

src_data	Numeric vector of float values to quantize
n_rows	Number of rows
n_per_row	Number of elements per row
imatrix	Optional importance matrix (numeric vector or NULL)

Value

Raw vector of quantized data

See Also

Other quantization: [dequantize_row_iq2_xxs\(\)](#), [dequantize_row_mxfp4\(\)](#), [dequantize_row_nvfp4\(\)](#), [dequantize_row_q1_0\(\)](#), [dequantize_row_q2_K\(\)](#), [dequantize_row_q4_0\(\)](#), [dequantize_row_tq1_0\(\)](#), [ggml_quant_block_info\(\)](#), [iq2xs_free_impl\(\)](#), [iq2xs_init_impl\(\)](#), [iq3xs_free_impl\(\)](#), [iq3xs_init_impl\(\)](#), [quantize_iq2_xxs\(\)](#), [quantize_mxfp4\(\)](#), [quantize_nvfp4\(\)](#), [quantize_q1_0\(\)](#), [quantize_q2_K\(\)](#), [quantize_q4_0\(\)](#), [quantize_row_iq3_xxs_ref\(\)](#), [quantize_row_mxfp4_ref\(\)](#), [quantize_row_q2_K_ref\(\)](#), [quantize_row_q4_0_ref\(\)](#), [quantize_row_tq1_0_ref\(\)](#)

rope_types	<i>RoPE Mode Constants</i>
------------	----------------------------

Description

RoPE (Rotary Position Embedding) Type Constants

Usage

```
GGML_ROPE_TYPE_NORM  
  
GGML_ROPE_TYPE_NEOX  
  
GGML_ROPE_TYPE_MROPE  
  
GGML_ROPE_TYPE_VISION
```

Format

Integer constants

An object of class integer of length 1.

An object of class integer of length 1.

An object of class integer of length 1.

Details

Constants for RoPE (Rotary Position Embedding) modes used in transformer models. Different models use different RoPE implementations.

- GGML_ROPE_TYPE_NORM (0): Standard RoPE as in original paper (LLaMA, Mistral)
- GGML_ROPE_TYPE_NEOX (2): GPT-NeoX style RoPE with different interleaving
- GGML_ROPE_TYPE_MROPE (8): Multi-RoPE for multimodal models (Qwen2-VL)
- GGML_ROPE_TYPE_VISION (24): Vision model RoPE variant

Value

An integer constant representing a RoPE type

Examples

```
GGML_ROPE_TYPE_NORM    # 0 - Standard RoPE (LLaMA, Mistral)  
GGML_ROPE_TYPE_NEOX    # 2 - GPT-NeoX style  
GGML_ROPE_TYPE_MROPE    # 8 - Multi-RoPE (Qwen2-VL)  
GGML_ROPE_TYPE_VISION  # 24 - Vision models
```

`RunGGML.SingleCellExperiment`*Run a GGML GPU operation on a Seurat object*

Description

High-level, Seurat-style entry point: extracts the expression matrix from the object, runs the requested operation on the GGML backend (Vulkan GPU with CPU fallback) and writes the result back into the object. Returns the updated object, so it slots into a `%>%` / `|>` pipeline next to `Seurat::RunPCA()`. Methods exist for `Seurat` and `SingleCellExperiment` objects (and a `.default` for bare matrices).

Usage

```
## S3 method for class 'SingleCellExperiment'
RunGGML(
  object,
  op = "embed",
  assay = NULL,
  layer = NULL,
  n_components = 50L,
  reduction_name = "ggml",
  device = "auto",
  genes = NULL,
  cells = NULL,
  reduction = NULL,
  dims = NULL,
  ...
)

RunGGML(
  object,
  op = "embed",
  assay = NULL,
  layer = NULL,
  n_components = 50L,
  reduction_name = "ggml",
  device = "auto",
  genes = NULL,
  cells = NULL,
  reduction = NULL,
  dims = NULL,
  ...
)

## Default S3 method:
RunGGML(
```

```

    object,
    op = "embed",
    assay = NULL,
    layer = NULL,
    n_components = 50L,
    reduction_name = "ggml",
    device = "auto",
    genes = NULL,
    cells = NULL,
    ...
)

## S3 method for class 'Seurat'
RunGGML(
  object,
  op = "embed",
  assay = NULL,
  layer = NULL,
  n_components = 50L,
  reduction_name = "ggml",
  device = "auto",
  genes = NULL,
  cells = NULL,
  reduction = NULL,
  dims = NULL,
  ...
)

```

Arguments

object	A Seurat or SingleCellExperiment object, or a bare feature-by-cell matrix/dgCMatix (the .default method returns a ggml_result instead of an object).
op	Operation name; see ggml_ops_registry . Default "embed".
assay	Assay to read (Seurat); defaults to the object's default assay.
layer	Layer/slot to read; default "data".
n_components	Number of components for "embed" (PCA). Default 50.
reduction_name	Name of the reduction slot to create. Default "ggml".
device	"auto" (default), "vulkan" or "cpu".
genes, cells	Optional feature/cell subsets passed to extraction.
reduction	Optional name of an existing reduction to use as the input feature space (e.g. "pca" for "umap" / "neighbors"), instead of an assay layer. Seurat: <code>Embeddings()</code> ; SCE: <code>reducedDim()</code> .
dims	Optional integer vector selecting columns of reduction.
...	Additional parameters forwarded to the engine. A notable one is <code>chunk_size</code> : for <code>op = "scale"</code> and <code>op = "embed"</code> (PCA), passing an integer streams the sparse input in blocks of that many cells, densifying one block at a time so the full dense

features-by-cells matrix (tens of GB at scale) is never held. Results are identical to the non-chunked path. `op = "normalize"` is already sparse, so `chunk_size` is a no-op there.

Details

Supported operations include "embed" (PCA), "normalize", "scale", "umap", "neighbors" and "largest_gene" (per-cell highest-expressed gene QC metric); see [ggml_ops_registry](#).

Value

For a Seurat object, the updated object with a new reduction. For a bare matrix, a [ggml_result](#).

Examples

```
## Not run:
library(Seurat)
pbmc <- RunGGML(pbmc, op = "embed", n_components = 30)
DimPlot(pbmc, reduction = "ggml")

## End(Not run)
```

```
summary.ggml_sequential_model
```

Summary method for ggml_sequential_model

Description

Prints a detailed summary including input shape, layer details, trainable/non-trainable parameter counts, and memory estimate.

Usage

```
## S3 method for class 'ggml_sequential_model'
summary(object, ...)
```

Arguments

<code>object</code>	A <code>ggml_sequential_model</code> object
<code>...</code>	Additional arguments (ignored)

Value

The model object (invisibly).

tidy.ggmlr_parsnip_model

Tidy a fitted ggml parsnip model into a per-layer table

Description

Returns one row per layer of the underlying sequential network, in broom style. Useful for comparing architectures across experiments in a R Markdown / Quarto report.

Usage

```
## S3 method for class 'ggmlr_parsnip_model'
tidy(x, ...)
```

Arguments

x	A fitted ‘ggmlr_parsnip_model’ (the engine object inside a parsnip fit; e.g. from ‘extract_fit_engine()’).
...	Unused; for generic compatibility.

Value

A [tibble][tibble::tibble] with columns: ‘layer’ (name), ‘type’, ‘units’ (output units, ‘NA’ if not applicable), ‘activation’, ‘output_shape’ (character), ‘params’ (trainable parameter count) and ‘trainable’ (logical).

Examples

```
ggml_set_n_threads(1L) # deterministic, single OpenMP pool
spec <- parsnip::mlp(hidden_units = 8L, epochs = 3L) |>
  parsnip::set_engine("ggml", backend = "cpu") |>
  parsnip::set_mode("regression")
fit_obj <- parsnip::fit(spec, mpg ~ ., data = mtcars)
generics::tidy(parsnip::extract_fit_engine(fit_obj))
```

with_grad_tape

Run code with gradient tape enabled

Description

Records all `ag_*` operations inside `expr` for later `backward()`. When the default device is "gpu", the ggml context is reset at the start of each tape.

Usage

```
with_grad_tape(expr)
```

Arguments

`expr` Expression to evaluate under gradient tape

Value

Value of last expression in `expr` (invisibly)

Examples

```
w <- ag_param(matrix(c(1, 0, 0, 1), 2, 2))
x <- ag_tensor(matrix(c(1, 2), 2, 1))
y <- ag_tensor(matrix(c(1, 2), 2, 1))
with_grad_tape({
  out <- ag_matmul(w, x)
  loss <- ag_mse_loss(out, y)
})
backward(loss)
```

Index

* backend

ggml_backend_buffer_clear, [61](#)
ggml_backend_buffer_get_usage, [63](#)
ggml_backend_buffer_is_host, [64](#)
ggml_backend_buffer_is_multi_buffer, [65](#)
ggml_backend_buffer_reset, [66](#)
ggml_backend_buffer_set_usage, [67](#)
ggml_backend_buffer_usage_any, [68](#)
ggml_backend_buffer_usage_compute, [69](#)
ggml_backend_buffer_usage_weights, [70](#)
ggml_backend_dev_by_name, [75](#)
ggml_backend_dev_by_type, [76](#)
ggml_backend_dev_count, [77](#)
ggml_backend_dev_description, [78](#)
ggml_backend_dev_get, [79](#)
ggml_backend_dev_get_props, [80](#)
ggml_backend_dev_init, [81](#)
ggml_backend_dev_memory, [82](#)
ggml_backend_dev_name, [83](#)
ggml_backend_dev_offload_op, [84](#)
ggml_backend_dev_supports_buft, [85](#)
ggml_backend_dev_supports_op, [86](#)
ggml_backend_dev_type, [87](#)
ggml_backend_device_register, [71](#)
ggml_backend_device_type_accel, [72](#)
ggml_backend_device_type_cpu, [73](#)
ggml_backend_device_type_gpu, [74](#)
ggml_backend_device_type_igpu, [75](#)
ggml_backend_event_free, [88](#)
ggml_backend_event_new, [89](#)
ggml_backend_event_record, [90](#)
ggml_backend_event_synchronize, [91](#)
ggml_backend_event_wait, [92](#)
ggml_backend_get_device, [93](#)
ggml_backend_graph_compute_async, [95](#)

ggml_backend_graph_plan_compute, [96](#)
ggml_backend_graph_plan_create, [97](#)
ggml_backend_graph_plan_free, [98](#)
ggml_backend_init_best, [99](#)
ggml_backend_init_by_name, [99](#)
ggml_backend_init_by_type, [100](#)
ggml_backend_load, [101](#)
ggml_backend_load_all, [102](#)
ggml_backend_meta_device, [103](#)
ggml_backend_multi_buffer_alloc_buffer, [104](#)
ggml_backend_multi_buffer_set_usage, [105](#)
ggml_backend_reg_by_name, [108](#)
ggml_backend_reg_count, [109](#)
ggml_backend_reg_dev_count, [109](#)
ggml_backend_reg_dev_get, [110](#)
ggml_backend_reg_get, [111](#)
ggml_backend_reg_name, [112](#)
ggml_backend_register, [107](#)
ggml_backend_synchronize, [122](#)
ggml_backend_tensor_copy_async, [123](#)
ggml_backend_tensor_get_async, [124](#)
ggml_backend_tensor_set_async, [126](#)
ggml_backend_unload, [128](#)

* callbacks

ggml_callback_early_stopping, [131](#)
ggml_schedule_cosine_decay, [343](#)
ggml_schedule_reduce_on_plateau, [344](#)
ggml_schedule_step_decay, [345](#)

* cpu_features

ggml_cpu_features, [144](#)
ggml_cpu_get_rvv_vlen, [145](#)
ggml_cpu_get_sve_cnt, [145](#)
ggml_cpu_has_amx_int8, [146](#)
ggml_cpu_has_arm_fma, [146](#)

- ggml_cpu_has_avx, 147
- ggml_cpu_has_avx2, 148
- ggml_cpu_has_avx512, 148
- ggml_cpu_has_avx512_bf16, 149
- ggml_cpu_has_avx512_vbmi, 149
- ggml_cpu_has_avx512_vnni, 150
- ggml_cpu_has_avx_vnni, 151
- ggml_cpu_has_bmi2, 151
- ggml_cpu_has_dotprod, 152
- ggml_cpu_has_f16c, 152
- ggml_cpu_has_fma, 153
- ggml_cpu_has_fp16_va, 154
- ggml_cpu_has_llamafile, 154
- ggml_cpu_has_matmul_int8, 155
- ggml_cpu_has_neon, 155
- ggml_cpu_has_riscv_v, 156
- ggml_cpu_has_sme, 157
- ggml_cpu_has_sse3, 157
- ggml_cpu_has_ssse3, 158
- ggml_cpu_has_sve, 159
- ggml_cpu_has_vsx, 159
- ggml_cpu_has_vxe, 160
- ggml_cpu_has_wasm_simd, 160
- * **datasets**
 - GGML_GLU_OP_REGLU, 208
 - ggml_pool_1d, 310
 - GGML_SORT_ORDER_ASC, 367
 - GGML_TYPE_F32, 387
 - ggml_upscale, 392
 - rope_types, 439
- * **graph**
 - ggml_graph_view, 215
 - ggml_op_can_inplace, 305
- * **logging**
 - ggml_abort_is_r_enabled, 51
 - ggml_log_is_r_enabled, 248
 - ggml_log_set_default, 249
 - ggml_log_set_r, 249
 - ggml_set_abort_callback_default, 347
 - ggml_set_abort_callback_r, 347
- * **op_info**
 - ggml_get_unary_op, 206
 - ggml_op_desc, 306
 - ggml_op_name, 306
 - ggml_op_symbol, 307
 - ggml_unary_op_name, 390
- * **optimization**
 - ggml_fit_opt, 181
 - ggml_opt_alloc, 271
 - ggml_opt_context_optimizer_type, 272
 - ggml_opt_dataset_data, 272
 - ggml_opt_dataset_free, 273
 - ggml_opt_dataset_get_batch, 274
 - ggml_opt_dataset_init, 275
 - ggml_opt_dataset_labels, 276
 - ggml_opt_dataset_ndata, 276
 - ggml_opt_dataset_shuffle, 277
 - ggml_opt_dataset_weights, 278
 - ggml_opt_default_params, 279
 - ggml_opt_epoch, 279
 - ggml_opt_eval, 281
 - ggml_opt_fit, 282
 - ggml_opt_free, 283
 - ggml_opt_get_lr, 284
 - ggml_opt_grad_acc, 285
 - ggml_opt_init, 285
 - ggml_opt_init_for_fit, 287
 - ggml_opt_inputs, 288
 - ggml_opt_labels, 288
 - ggml_opt_loss, 289
 - ggml_opt_loss_type_cross_entropy, 290
 - ggml_opt_loss_type_mean, 290
 - ggml_opt_loss_type_mse, 291
 - ggml_opt_loss_type_sum, 292
 - ggml_opt_loss_type_weighted_mse, 292
 - ggml_opt_ncorrect, 293
 - ggml_opt_optimizer_name, 294
 - ggml_opt_optimizer_type_adamw, 294
 - ggml_opt_optimizer_type_sgd, 295
 - ggml_opt_outputs, 296
 - ggml_opt_pred, 296
 - ggml_opt_prepare_alloc, 297
 - ggml_opt_reset, 298
 - ggml_opt_result_accuracy, 299
 - ggml_opt_result_free, 299
 - ggml_opt_result_init, 300
 - ggml_opt_result_loss, 301
 - ggml_opt_result_ndata, 301
 - ggml_opt_result_pred, 302
 - ggml_opt_result_reset, 303
 - ggml_opt_set_lr, 304
 - ggml_opt_static_graphs, 304

*** quantization**

dequantize_row_iq2_xxs, 42
 dequantize_row_mxfp4, 43
 dequantize_row_nvfp4, 43
 dequantize_row_q1_0, 44
 dequantize_row_q2_K, 45
 dequantize_row_q4_0, 46
 dequantize_row_tq1_0, 46
 ggml_quant_block_info, 320
 iq2xs_free_impl, 417
 iq2xs_init_impl, 418
 iq3xs_free_impl, 418
 iq3xs_init_impl, 419
 quantize_iq2_xxs, 429
 quantize_mxfp4, 430
 quantize_nvfp4, 431
 quantize_q1_0, 432
 quantize_q2_K, 433
 quantize_q4_0, 434
 quantize_row_iq3_xxs_ref, 435
 quantize_row_mxfp4_ref, 436
 quantize_row_q2_K_ref, 436
 quantize_row_q4_0_ref, 437
 quantize_row_tq1_0_ref, 438
 quantize_tq1_0, 439

*** rope**

ggml_rope_ext_inplace, 335
 ggml_rope_multi, 336
 ggml_rope_multi_inplace, 338

*** softmax**

ggml_soft_max_ext_back_inplace, 365
 ggml_soft_max_ext_inplace, 366

*** tensor_layout**

ggml_are_same_stride, 58
 ggml_can_repeat, 132
 ggml_count_equal, 142
 ggml_is_contiguous_0, 225
 ggml_is_contiguous_1, 226
 ggml_is_contiguous_2, 226
 ggml_is_contiguous_channels, 227
 ggml_is_contiguous_rows, 228
 ggml_is_contiguously_allocated, 225

*** tensor**

ggml_are_same_layout, 57
 ggml_get_op_params, 203
 ggml_get_op_params_f32, 203

ggml_get_op_params_i32, 204
 ggml_set_op_params, 353
 ggml_set_op_params_f32, 353
 ggml_set_op_params_i32, 354

*** type_system**

ggml_blk_size, 129
 ggml_ftype_to_ggml_type, 188
 ggml_is_quantized, 229
 ggml_type_name, 389
 ggml_type_sizef, 390
 %**%, ANY, ggml_matrix-method
 (ggml_matrix-class), 253
 %**%, ggml_matrix, ANY-method
 (ggml_matrix-class), 253
 %**%, ggml_matrix, ggml_matrix-method
 (ggml_matrix-class), 253

ag_add, 15
 ag_batch_norm, 15
 ag_clamp, 16
 ag_cross_entropy_loss, 16
 ag_dataloader, 17
 ag_default_device, 18
 ag_default_dtype, 18
 ag_device, 19
 ag_dropout, 19
 ag_dtype, 20
 ag_embedding, 20
 ag_eval, 21
 ag_exp, 21
 ag_gradcheck, 22
 ag_linear, 23
 ag_load_model, 23, 30, 31
 ag_log, 24
 ag_matmul, 25
 ag_mean, 25
 ag_mse_loss, 26
 ag_mul, 26
 ag_multihead_attention, 27
 ag_param, 28
 ag_pow, 29
 ag_relu, 29
 ag_reshape, 30
 ag_save_model, 23, 24, 30
 ag_scale, 31
 ag_sequential, 32
 ag_sigmoid, 32
 ag_softmax, 33
 ag_softmax_cross_entropy_loss, 33

- ag_sub, 34
- ag_sum, 34
- ag_tanh, 35
- ag_tensor, 35
- ag_to_device, 36
- ag_train, 36
- ag_transpose, 37
- as.matrix.ggml_matrix, 37
- as_gpu_matrix, 38, 252, 253
- augment.ggmlr_parsnip_model, 38

- backward, 39

- clip_grad_norm, 40
- compile.ggml_functional_model
 (compile.ggml_sequential_model),
 41
- compile.ggml_sequential_model, 41
- crossprod,ggml_matrix-method
 (ggml_matrix-class), 253

- dequantize_row_iq1_m
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq1_s
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq2_s
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq2_xs
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq2_xxs, 42, 43–47, 320,
 417–419, 430–439
- dequantize_row_iq3_s
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq3_xxs
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq4_n1
 (dequantize_row_iq2_xxs), 42
- dequantize_row_iq4_xs
 (dequantize_row_iq2_xxs), 42
- dequantize_row_mxfp4, 42, 43, 44–47, 320,
 417–419, 430–439
- dequantize_row_nvfp4, 42, 43, 43, 44–47,
 320, 417–419, 430–439
- dequantize_row_q1_0, 42–44, 44, 45–47,
 320, 417–419, 430–439
- dequantize_row_q2_K, 42–44, 45, 46, 47,
 320, 417–419, 430–439
- dequantize_row_q3_K
 (dequantize_row_q2_K), 45
- dequantize_row_q4_0, 42–45, 46, 47, 320,
 417–419, 430–439
- dequantize_row_q4_1
 (dequantize_row_q4_0), 46
- dequantize_row_q4_K
 (dequantize_row_q2_K), 45
- dequantize_row_q5_0
 (dequantize_row_q4_0), 46
- dequantize_row_q5_1
 (dequantize_row_q4_0), 46
- dequantize_row_q5_K
 (dequantize_row_q2_K), 45
- dequantize_row_q6_K
 (dequantize_row_q2_K), 45
- dequantize_row_q8_0
 (dequantize_row_q4_0), 46
- dequantize_row_q8_K
 (dequantize_row_q2_K), 45
- dequantize_row_tq1_0, 42–46, 46, 320,
 417–419, 430–439
- dequantize_row_tq2_0
 (dequantize_row_tq1_0), 46
- dim,ggml_matrix-method
 (ggml_matrix-class), 253
- dp_train, 47

- evaluate.ggml_functional_model
 (evaluate.ggml_sequential_model),
 49
- evaluate.ggml_sequential_model, 49

- fit.ggml_functional_model
 (fit.ggml_sequential_model), 50
- fit.ggml_sequential_model, 50

- ggml_abort_is_r_enabled, 51, 248, 249,
 347
- ggml_abs, 52
- ggml_abs_inplace, 52
- ggml_add, 53
- ggml_add1, 54
- ggml_add_inplace, 55
- ggml_add_rel_pos, 55
- ggml_apply, 56
- ggml_arange, 57
- ggml_are_same_layout, 57, 203, 204, 353,
 354
- ggml_are_same_shape, 58

- `ggml_are_same_stride`, [58](#), [132](#), [143](#),
[225–228](#)
- `ggml_argmax`, [59](#)
- `ggml_argsort`, [60](#)
- `ggml_backend_alloc_ctx_tensors`, [60](#)
- `ggml_backend_buffer_clear`, [61](#), [63–65](#),
[67–70](#), [72–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_buffer_free`, [62](#)
- `ggml_backend_buffer_get_size`, [62](#)
- `ggml_backend_buffer_get_usage`, [61](#), [63](#),
[64](#), [65](#), [67–70](#), [72–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_buffer_is_host`, [61](#), [63](#), [64](#),
[65](#), [67–70](#), [72–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_buffer_is_multi_buffer`,
[61](#), [63](#), [64](#), [65](#), [67–70](#), [72–92](#),
[94–102](#), [104–112](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_buffer_name`, [66](#)
- `ggml_backend_buffer_reset`, [61](#), [63–65](#), [66](#),
[67–70](#), [72–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_buffer_set_usage`, [61](#),
[63–65](#), [67](#), [68](#), [68–70](#), [72–92](#),
[94–102](#), [104–112](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_buffer_usage_any`, [61](#),
[63–65](#), [67](#), [68](#), [69](#), [70](#), [72–92](#),
[94–102](#), [104–112](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_buffer_usage_compute`, [61](#),
[63–65](#), [67](#), [68](#), [69](#), [70](#), [72–92](#),
[94–102](#), [104–112](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_buffer_usage_weights`, [61](#),
[63–65](#), [67–69](#), [70](#), [72–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_cpu_init`, [70](#)
- `ggml_backend_cpu_set_n_threads`, [71](#)
- `ggml_backend_dev_by_name`, [61](#), [63–65](#),
[67–70](#), [72–75](#), [75](#), [77–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_by_type`, [61](#), [63–65](#),
[67–70](#), [72–76](#), [76](#), [77–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_count`, [61](#), [63–65](#), [67–70](#),
[72–77](#), [77](#), [78–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_description`, [61](#), [63–65](#),
[67–70](#), [72–77](#), [78](#), [79–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_get`, [61](#), [63–65](#), [67–70](#),
[72–78](#), [79](#), [80–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_get_props`, [61](#), [63–65](#),
[67–70](#), [72–79](#), [80](#), [81–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_init`, [61](#), [63–65](#), [67–70](#),
[72–80](#), [81](#), [82–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_memory`, [61](#), [63–65](#),
[67–70](#), [72–81](#), [82](#), [83–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_name`, [61](#), [63–65](#), [67–70](#),
[72–82](#), [83](#), [84–92](#), [94–102](#), [104–112](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_offload_op`, [61](#), [63–65](#),
[67–70](#), [72–83](#), [84](#), [85–92](#), [94–102](#),
[104–112](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_supports_buft`, [61](#),
[63–65](#), [67–70](#), [72–84](#), [85](#), [86–92](#),
[94–102](#), [104–113](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_dev_supports_op`, [61](#), [63–65](#),
[67–70](#), [72–85](#), [86](#), [87–92](#), [94–102](#),
[104–113](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_dev_type`, [61](#), [63–65](#), [67–70](#),
[72–86](#), [87](#), [88–92](#), [94–102](#), [104–113](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_device_register`, [61](#), [63–65](#),
[67–70](#), [71](#), [73–92](#), [94–102](#), [104–113](#),
[122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_device_type_accel`, [61](#),
[63–65](#), [67–70](#), [72](#), [72](#), [73–92](#),
[94–102](#), [104–113](#), [122](#), [123](#), [125](#),
[127](#), [128](#)
- `ggml_backend_device_type_cpu`, [61](#), [63–65](#),
[67–70](#), [72](#), [73](#), [73](#), [74–92](#), [94–102](#),
[104–113](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_device_type_gpu`, [61](#), [63–65](#),
[67–70](#), [72](#), [73](#), [74](#), [75–92](#), [94–102](#),
[104–113](#), [122](#), [123](#), [125](#), [127](#), [128](#)
- `ggml_backend_device_type_igpu`, [61](#),
[63–65](#), [67–70](#), [72–74](#), [75](#), [76–92](#),

- [94–102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_event_free, 61, 63–65, 67–70, 72–87, 88, 89–92, 94–102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_event_new, 61, 63–65, 67–70, 72–88, 89, 90–92, 94–102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_event_record, 61, 63–65, 67–70, 72–89, 90, 91, 92, 94–102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_event_synchronize, 61, 63–65, 67–70, 72–90, 91, 92, 94–102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_event_wait, 61, 63–65, 67–70, 72–91, 92, 94, 95, 97–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_free, 93](#)
- [ggml_backend_get_device, 61, 63–65, 67–70, 72–92, 93, 95, 97–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_graph_compute, 94](#)
- [ggml_backend_graph_compute_async, 61, 63–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_graph_plan_compute, 61, 63–65, 67–70, 72–92, 94, 95, 96, 97–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_graph_plan_create, 61, 63–65, 67–70, 72–92, 94, 95, 97, 97, 98–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_graph_plan_free, 61, 63–65, 67–70, 72–92, 94, 95, 97, 98, 99–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_init_best, 61, 63–65, 67–70, 72–92, 94, 95, 97, 98, 99, 100–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_init_by_name, 62–65, 67–70, 72–92, 94, 95, 97–99, 99, 101–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_init_by_type, 62–65, 67–70, 72–92, 94, 95, 97–100, 100, 102–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_load, 62–65, 67–70, 72–92, 94, 95, 97–101, 101, 103–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_load_all, 62–65, 67–70, 72–92, 94, 95, 97–102, 102, 104–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_meta_device, 62–65, 67–70, 72–92, 94, 95, 97–103, 103, 105–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_multi_buffer_alloc_buffer, 62–65, 67–70, 72–92, 94, 95, 97–104, 104, 106–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_multi_buffer_set_usage, 62–65, 67–70, 72–92, 94, 95, 97–105, 105, 107–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_name, 106](#)
- [ggml_backend_reg_by_name, 62–65, 67–70, 72–92, 94, 95, 97–107, 108, 109–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_reg_count, 62–65, 67–70, 72–92, 94, 95, 97–108, 109, 110–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_reg_dev_count, 62–65, 67–70, 72–92, 94, 95, 97–109, 109, 111–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_reg_dev_get, 62–65, 67–70, 72–92, 94, 95, 97–110, 110, 112, 113, 122, 123, 125, 127, 128](#)
- [ggml_backend_reg_get, 62–65, 67–70, 72–92, 94, 95, 97–111, 111, 113, 122, 123, 125, 127, 128](#)
- [ggml_backend_reg_name, 62–65, 67–70, 72–92, 94, 95, 97–112, 112, 122, 123, 125, 127, 128](#)
- [ggml_backend_register, 62–65, 67–70, 72–92, 94, 95, 97–106, 107, 108–113, 122, 123, 125, 127, 128](#)
- [ggml_backend_sched_alloc_graph, 113](#)
- [ggml_backend_sched_free, 113](#)
- [ggml_backend_sched_get_backend, 114](#)
- [ggml_backend_sched_get_n_backends, 115](#)
- [ggml_backend_sched_get_n_copies, 115](#)
- [ggml_backend_sched_get_n_splits, 116](#)
- [ggml_backend_sched_get_tensor_backend, 116](#)
- [ggml_backend_sched_graph_compute, 117](#)
- [ggml_backend_sched_graph_compute_async, 118](#)
- [ggml_backend_sched_new, 118](#)
- [ggml_backend_sched_reserve, 119](#)

- ggml_backend_sched_reset, 120
- ggml_backend_sched_set_tensor_backend, 121
- ggml_backend_sched_synchronize, 121
- ggml_backend_synchronize, 62–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 125, 127, 128
- ggml_backend_tensor_copy_async, 62–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 125, 127, 128
- ggml_backend_tensor_get_and_sync, 124
- ggml_backend_tensor_get_async, 62–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 124, 127, 128
- ggml_backend_tensor_get_data, 125
- ggml_backend_tensor_get_f32_first, 126
- ggml_backend_tensor_set_async, 62–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 125, 126, 128
- ggml_backend_tensor_set_data, 127
- ggml_backend_unload, 62–65, 67–70, 72–92, 94, 95, 97–113, 122, 123, 125, 127, 128
- ggml_batch_norm, 129
- ggml_blk_size, 129, 188, 229, 389, 390
- ggml_build_forward_expand, 130
- ggml_callback_early_stopping, 131, 343–345
- ggml_can_repeat, 59, 132, 143, 225–228
- ggml_ceil, 132
- ggml_ceil_inplace, 133
- ggml_clamp, 133
- ggml_compile, 41, 165, 181
- ggml_compile
 - (ggml_compile.ggml_functional_model), 134
- ggml_compile.ggml_functional_model, 134
- ggml_concat, 135
- ggml_cont, 136
- ggml_conv_1d, 136
- ggml_conv_1d_dw, 137
- ggml_conv_2d, 138
- ggml_conv_2d_direct, 138
- ggml_conv_2d_dw, 139
- ggml_conv_2d_dw_direct, 140
- ggml_conv_transpose_1d, 141
- ggml_conv_transpose_2d_p0, 141
- ggml_cos, 142
- ggml_count_equal, 59, 132, 142, 225–228
- ggml_cpu_add, 143
- ggml_cpu_features, 144, 145–161
- ggml_cpu_get_rvv_vlen, 144, 145, 146–161
- ggml_cpu_get_sve_cnt, 144, 145, 145, 146–161
- ggml_cpu_has_amx_int8, 144–146, 146, 147–161
- ggml_cpu_has_arm_fma, 144–146, 146, 147–161
- ggml_cpu_has_avx, 144–147, 147, 148–161
- ggml_cpu_has_avx2, 144–147, 148, 149–161
- ggml_cpu_has_avx512, 144–148, 148, 149–161
- ggml_cpu_has_avx512_bf16, 144–149, 149, 150–161
- ggml_cpu_has_avx512_vbmi, 144–149, 149, 150–161
- ggml_cpu_has_avx512_vnni, 144–150, 150, 151–161
- ggml_cpu_has_avx_vnni, 144–150, 151, 152–161
- ggml_cpu_has_bmi2, 144–151, 151, 152–161
- ggml_cpu_has_dotprod, 144–152, 152, 153–161
- ggml_cpu_has_f16c, 144–152, 152, 153–161
- ggml_cpu_has_fma, 144–153, 153, 154–161
- ggml_cpu_has_fp16_va, 144–153, 154, 155–161
- ggml_cpu_has_llamafile, 144–154, 154, 155–161
- ggml_cpu_has_matmul_int8, 144–155, 155, 156–161
- ggml_cpu_has_neon, 144–155, 155, 156–161
- ggml_cpu_has_riscv_v, 144–156, 156, 157–161
- ggml_cpu_has_sme, 144–156, 157, 158–161
- ggml_cpu_has_sse3, 144–157, 157, 158–161
- ggml_cpu_has_ssse3, 144–158, 158, 159–161
- ggml_cpu_has_sve, 144–158, 159, 160, 161
- ggml_cpu_has_vsx, 144–159, 159, 160, 161
- ggml_cpu_has_vxe, 144–160, 160, 161
- ggml_cpu_has_wasm_simd, 144–160, 160
- ggml_cpu_mul, 161
- ggml_cpy, 162
- ggml_crossprod, 162, 252

ggml_cycles, 163
 ggml_cycles_per_ms, 163
 ggml_default_mlp, 164
 ggml_dense, 165
 ggml_diag, 166
 ggml_diag_mask_inf, 167
 ggml_diag_mask_inf_inplace, 168
 ggml_diag_mask_zero, 168
 ggml_div, 169
 ggml_div_inplace, 170
 ggml_dup, 170
 ggml_dup_inplace, 171
 ggml_dup_tensor, 171
 ggml_element_size, 172
 ggml_elu, 172
 ggml_elu_inplace, 173
 ggml_embedding, 173
 ggml_estimate_memory, 174
 ggml_evaluate, 49
 ggml_evaluate
 (ggml_evaluate.ggml_functional_model), ggml_get_op_params_i32, 58, 203, 204, 204,
 175
 ggml_evaluate.ggml_functional_model,
 175
 ggml_exp, 176
 ggml_exp_inplace, 177
 ggml_extract, 177
 ggml_fit, 50, 51, 165, 386
 ggml_fit
 (ggml_fit.ggml_functional_model),
 179
 ggml_fit.ggml_functional_model, 179
 ggml_fit_opt, 181, 181, 271–281, 283–305
 ggml_flash_attn_back, 183
 ggml_flash_attn_ext, 184
 ggml_floor, 185
 ggml_floor_inplace, 186
 ggml_free, 186
 ggml_freeze_weights, 187
 ggml_ftype_to_ggml_type, 129, 188, 229,
 389, 390
 ggml_gallocr_alloc_graph, 188
 ggml_gallocr_free, 189
 ggml_gallocr_get_buffer_size, 189
 ggml_gallocr_new, 190
 ggml_gallocr_reserve, 191
 ggml_geglu, 191
 ggml_geglu_quick, 192
 ggml_geglu_split, 192
 ggml_gelu, 193
 ggml_gelu_erf, 194
 ggml_gelu_inplace, 194
 ggml_gelu_quick, 195
 ggml_get_f32, 196
 ggml_get_f32_nd, 197
 ggml_get_first_tensor, 197
 ggml_get_i32, 198
 ggml_get_i32_nd, 198
 ggml_get_layer, 199
 ggml_get_max_tensor_size, 199
 ggml_get_mem_size, 200
 ggml_get_n_threads, 202
 ggml_get_name, 201
 ggml_get_next_tensor, 201
 ggml_get_no_alloc, 202
 ggml_get_op_params, 58, 203, 204, 353, 354
 ggml_get_op_params_f32, 58, 203, 203, 204,
 353, 354
 (ggml_evaluate.ggml_functional_model), ggml_get_op_params_i32, 58, 203, 204, 204,
 353, 354
 ggml_get_rel_pos, 204
 ggml_get_rows, 205
 ggml_get_rows_back, 206
 ggml_get_unary_op, 206, 306, 307, 391
 ggml_glu, 207
 GGML_GLU_OP_GEGLU (GGML_GLU_OP_REGLU),
 208
 GGML_GLU_OP_GEGLU_ERF
 (GGML_GLU_OP_REGLU), 208
 GGML_GLU_OP_GEGLU_QUICK
 (GGML_GLU_OP_REGLU), 208
 GGML_GLU_OP_REGLU, 208
 GGML_GLU_OP_SWIGLU (GGML_GLU_OP_REGLU),
 208
 GGML_GLU_OP_SWIGLU_OAI
 (GGML_GLU_OP_REGLU), 208
 ggml_glu_split, 209
 ggml_graph_compute, 209
 ggml_graph_compute_with_ctx, 210
 ggml_graph_dump_dot, 211
 ggml_graph_get_tensor, 212
 ggml_graph_n_nodes, 213
 ggml_graph_node, 212
 ggml_graph_overhead, 213
 ggml_graph_print, 214
 ggml_graph_reset, 214

- ggml_graph_view, 215, 305
- ggml_group_norm, 216
- ggml_group_norm_inplace, 216
- ggml_gru, 217
- ggml_hardsigmoid, 218
- ggml_hardswish, 219
- ggml_im2col, 219
- ggml_init, 220
- ggml_init_auto, 221
- ggml_inject, 222
- ggml_input, 223
- ggml_is_available, 224
- ggml_is_contiguous, 224
- ggml_is_contiguous_0, 59, 132, 143, 225, 225, 226–228
- ggml_is_contiguous_1, 59, 132, 143, 225, 226, 226, 227, 228
- ggml_is_contiguous_2, 59, 132, 143, 225, 226, 226, 227, 228
- ggml_is_contiguous_channels, 59, 132, 143, 225–227, 227, 228
- ggml_is_contiguous_rows, 59, 132, 143, 225–227, 228
- ggml_is_contiguously_allocated, 59, 132, 143, 225, 226–228
- ggml_is_permuted, 228
- ggml_is_quantized, 129, 188, 229, 389, 390
- ggml_is_transposed, 230
- ggml_l2_norm, 230
- ggml_l2_norm_inplace, 231
- ggml_layer_add, 232
- ggml_layer_batch_norm, 232
- ggml_layer_concatenate, 233
- ggml_layer_conv_1d, 234
- ggml_layer_conv_2d, 235
- ggml_layer_dense, 164, 165, 236
- ggml_layer_dropout, 165, 237
- ggml_layer_embedding, 239
- ggml_layer_flatten, 240
- ggml_layer_global_average_pooling_2d, 240
- ggml_layer_global_max_pooling_2d, 241
- ggml_layer_gru, 242
- ggml_layer_lstm, 243
- ggml_layer_max_pooling_2d, 244
- ggml_leaky_relu, 245
- ggml_load_model, 246, 392
- ggml_load_weights, 246
- ggml_log, 247
- ggml_log_inplace, 248
- ggml_log_is_r_enabled, 51, 248, 249, 347
- ggml_log_set_default, 51, 248, 249, 249, 347
- ggml_log_set_r, 51, 248, 249, 249, 347
- ggml_lstm, 250
- ggml_marshal_model, 250, 392
- ggml_matmul, 38, 162, 163, 251, 252, 253, 377
- ggml_matmul_f64, 252
- ggml_matrix, 37, 38
- ggml_matrix-class, 253
- ggml_mean, 254
- ggml_model, 255
- ggml_model_backend, 256, 386
- ggml_model_sequential, 165, 256
- ggml_mul, 257
- ggml_mul_inplace, 258
- ggml_mul_mat, 258
- ggml_mul_mat_id, 259
- ggml_n_dims, 270
- ggml_nbytes, 260
- ggml_neg, 261
- ggml_neg_inplace, 262
- ggml_nelements, 262
- ggml_new_f32, 263
- ggml_new_i32, 264
- ggml_new_tensor, 264
- ggml_new_tensor_1d, 265
- ggml_new_tensor_2d, 266
- ggml_new_tensor_3d, 266
- ggml_new_tensor_4d, 267
- ggml_norm, 268
- ggml_norm_inplace, 269
- ggml_nrows, 269
- ggml_op_can_inplace, 215, 305
- ggml_op_desc, 207, 306, 307, 391
- ggml_op_name, 207, 306, 306, 307, 391
- GGML_OP_POOL_AVG (ggml_pool_1d), 310
- GGML_OP_POOL_MAX (ggml_pool_1d), 310
- ggml_op_symbol, 207, 306, 307, 307, 391
- ggml_ops_registry, 270, 340, 377, 442, 443
- ggml_opt_alloc, 182, 271, 272–281, 283–305
- ggml_opt_context_optimizer_type, 182, 271, 272, 273–281, 283–305
- ggml_opt_dataset_data, 182, 271, 272, 272, 273–281, 283–305

- `ggml_opt_dataset_free`, [182](#), [271–273](#), [273](#), [274–281](#), [283–305](#)
- `ggml_opt_dataset_get_batch`, [182](#), [271–273](#), [274](#), [275–281](#), [283–305](#)
- `ggml_opt_dataset_init`, [182](#), [271–274](#), [275](#), [276–281](#), [283–305](#)
- `ggml_opt_dataset_labels`, [182](#), [271–275](#), [276](#), [277–281](#), [283–305](#)
- `ggml_opt_dataset_ndata`, [182](#), [271–276](#), [276](#), [278–281](#), [283–305](#)
- `ggml_opt_dataset_shuffle`, [182](#), [271–277](#), [277](#), [278–281](#), [283–305](#)
- `ggml_opt_dataset_weights`, [182](#), [271–278](#), [278](#), [279–281](#), [283–305](#)
- `ggml_opt_default_params`, [182](#), [271–278](#), [279](#), [280](#), [281](#), [283–305](#)
- `ggml_opt_epoch`, [182](#), [271–279](#), [279](#), [281](#), [283–305](#)
- `ggml_opt_eval`, [182](#), [271–280](#), [281](#), [283–305](#)
- `ggml_opt_fit`, [182](#), [271–281](#), [282](#), [284–305](#)
- `ggml_opt_free`, [182](#), [271–281](#), [283](#), [283](#), [284–305](#)
- `ggml_opt_get_lr`, [183](#), [271–281](#), [283](#), [284](#), [284](#), [285–305](#)
- `ggml_opt_grad_acc`, [183](#), [271–281](#), [283](#), [284](#), [285](#), [286–305](#)
- `ggml_opt_init`, [183](#), [271–281](#), [283–285](#), [285](#), [287–305](#)
- `ggml_opt_init_for_fit`, [183](#), [271–281](#), [283–286](#), [287](#), [288–305](#)
- `ggml_opt_inputs`, [183](#), [271–281](#), [283–287](#), [288](#), [289–305](#)
- `ggml_opt_labels`, [183](#), [271–281](#), [283–288](#), [288](#), [289–305](#)
- `ggml_opt_loss`, [183](#), [271–281](#), [283–289](#), [289](#), [290–305](#)
- `ggml_opt_loss_type_cross_entropy`, [183](#), [271–281](#), [283–289](#), [290](#), [291–305](#)
- `ggml_opt_loss_type_mean`, [183](#), [271–281](#), [283–290](#), [290](#), [291–305](#)
- `ggml_opt_loss_type_mse`, [183](#), [271–281](#), [283–291](#), [291](#), [292–305](#)
- `ggml_opt_loss_type_sum`, [183](#), [271–281](#), [283–291](#), [292](#), [293–305](#)
- `ggml_opt_loss_type_weighted_mse`, [183](#), [271–281](#), [283–292](#), [292](#), [293–305](#)
- `ggml_opt_ncorrect`, [183](#), [271–281](#), [283–293](#), [293](#), [294–305](#)
- `ggml_opt_optimizer_name`, [183](#), [271–281](#), [283–293](#), [294](#), [295–305](#)
- `ggml_opt_optimizer_type_adamw`, [183](#), [271–281](#), [283–294](#), [294](#), [295–305](#)
- `ggml_opt_optimizer_type_sgd`, [183](#), [271–281](#), [283–295](#), [295](#), [296–305](#)
- `ggml_opt_outputs`, [183](#), [271–281](#), [283–295](#), [296](#), [297–305](#)
- `ggml_opt_pred`, [183](#), [271–281](#), [283–296](#), [296](#), [298–305](#)
- `ggml_opt_prepare_alloc`, [183](#), [271–281](#), [283–297](#), [297](#), [298–305](#)
- `ggml_opt_reset`, [183](#), [271–281](#), [283–298](#), [298](#), [299–305](#)
- `ggml_opt_result_accuracy`, [183](#), [271–281](#), [283–298](#), [299](#), [300–305](#)
- `ggml_opt_result_free`, [183](#), [271–281](#), [283–299](#), [299](#), [301–305](#)
- `ggml_opt_result_init`, [183](#), [271–281](#), [283–300](#), [300](#), [301–305](#)
- `ggml_opt_result_loss`, [183](#), [271–281](#), [283–301](#), [301](#), [302–305](#)
- `ggml_opt_result_ndata`, [183](#), [271–281](#), [283–301](#), [301](#), [303–305](#)
- `ggml_opt_result_pred`, [183](#), [271–281](#), [283–302](#), [302](#), [303–305](#)
- `ggml_opt_result_reset`, [183](#), [271–281](#), [283–303](#), [303](#), [304](#), [305](#)
- `ggml_opt_set_lr`, [183](#), [271–281](#), [283–303](#), [304](#), [305](#)
- `ggml_opt_static_graphs`, [183](#), [271–281](#), [283–304](#), [304](#)
- `ggml_out_prod`, [307](#)
- `ggml_pad`, [308](#)
- `ggml_pad_reflect_1d`, [309](#)
- `ggml_permute`, [309](#)
- `ggml_pool_1d`, [310](#)
- `ggml_pool_2d`, [311](#)
- `ggml_pop_layer`, [312](#)
- `ggml_pp_dp_forward`, [313](#)
- `ggml_pp_forward`, [313](#), [314](#), [314](#)
- `ggml_predict`
(`ggml_predict.ggml_functional_model`), [315](#)
- `ggml_predict.ggml_functional_model`, [315](#)
- `ggml_predict_classes`, [316](#)
- `ggml_print_mem_status`, [317](#)

- ggml_print_objects, 317
- ggml_quant_block_info, 42–47, 320, 417–419, 430–439
- ggml_quantize_chunk, 318
- ggml_quantize_free, 319
- ggml_quantize_init, 319
- ggml_quantize_requires_imatrix, 320
- ggml_reglu, 321
- ggml_reglu_split, 321
- ggml_relu, 322
- ggml_relu_inplace, 323
- ggml_repeat, 323
- ggml_repeat_back, 324
- ggml_reset, 324
- ggml_reshape_1d, 325
- ggml_reshape_2d, 326
- ggml_reshape_3d, 326
- ggml_reshape_4d, 327
- ggml_result, 222, 223, 328, 340, 442, 443
- ggml_rms_norm, 328
- ggml_rms_norm_back, 329
- ggml_rms_norm_inplace, 330
- ggml_roll, 330
- ggml_rope, 331
- ggml_rope_ext, 332
- ggml_rope_ext_back, 334
- ggml_rope_ext_inplace, 335, 337, 339
- ggml_rope_inplace, 336
- ggml_rope_multi, 336, 336, 339
- ggml_rope_multi_inplace, 336, 337, 338
- GGML_ROPE_TYPE_MROPE (rope_types), 439
- GGML_ROPE_TYPE_NEOX (rope_types), 439
- GGML_ROPE_TYPE_NORM (rope_types), 439
- GGML_ROPE_TYPE_VISION (rope_types), 439
- ggml_round, 339
- ggml_round_inplace, 339
- ggml_run, 340, 377
- ggml_save_model, 251, 340
- ggml_save_weights, 341
- ggml_scale, 342
- ggml_scale_inplace, 342
- GGML_SCALE_MODE_BICUBIC (ggml_upscale), 392
- GGML_SCALE_MODE_BILINEAR (ggml_upscale), 392
- GGML_SCALE_MODE_NEAREST (ggml_upscale), 392
- ggml_schedule_cosine_decay, 131, 343, 344, 345
- ggml_schedule_reduce_on_plateau, 131, 343, 344, 345
- ggml_schedule_step_decay, 131, 343, 344, 345
- ggml_set, 345
- ggml_set_1d, 346
- ggml_set_2d, 346
- ggml_set_abort_callback_default, 51, 248, 249, 347, 347
- ggml_set_abort_callback_r, 51, 248, 249, 347, 347
- ggml_set_f32, 348
- ggml_set_f32_nd, 349
- ggml_set_i32, 349
- ggml_set_i32_nd, 350
- ggml_set_input, 350
- ggml_set_n_threads, 352
- ggml_set_name, 351
- ggml_set_no_alloc, 351
- ggml_set_omp_threads, 352
- ggml_set_op_params, 58, 203, 204, 353, 354
- ggml_set_op_params_f32, 58, 203, 204, 353, 353, 354
- ggml_set_op_params_i32, 58, 203, 204, 353, 354, 354
- ggml_set_output, 355
- ggml_set_param, 355
- ggml_set_seed, 356
- ggml_set_zero, 357
- ggml_sgn, 357
- ggml_sigmoid, 358
- ggml_sigmoid_inplace, 359
- ggml_silu, 359
- ggml_silu_back, 360
- ggml_silu_inplace, 360
- ggml_sin, 361
- ggml_soft_max, 363
- ggml_soft_max_ext, 364
- ggml_soft_max_ext_back, 365
- ggml_soft_max_ext_back_inplace, 365, 366
- ggml_soft_max_ext_inplace, 366, 366
- ggml_soft_max_inplace, 367
- ggml_softplus, 362
- ggml_softplus_inplace, 362
- GGML_SORT_ORDER_ASC, 367
- GGML_SORT_ORDER_DESC

- (GGML_SORT_ORDER_ASC), 367
- ggml_sqr, 368
- ggml_sqr_inplace, 370
- ggml_sqrt, 369
- ggml_sqrt_inplace, 370
- ggml_step, 371
- ggml_sub, 371
- ggml_sub_inplace, 372
- ggml_sum, 373
- ggml_sum_rows, 373
- ggml_swiglu, 374
- ggml_swiglu_split, 375
- ggml_tanh, 375
- ggml_tanh_inplace, 376
- ggml_task, 270, 340, 376
- ggml_tcrossprod, 252, 377
- ggml_tensor_copy, 378
- ggml_tensor_nb, 378
- ggml_tensor_num, 379
- ggml_tensor_overhead, 379
- ggml_tensor_set_f32_scalar, 380
- ggml_tensor_shape, 380
- ggml_tensor_type, 381
- ggml_test, 381
- ggml_time_init, 382
- ggml_time_ms, 383
- ggml_time_us, 383
- ggml_timestep_embedding, 382
- ggml_top_k, 384
- ggml_tp_dp_forward, 313–315, 384
- ggml_training_history, 256, 386
- ggml_transpose, 386
- GGML_TYPE_BF16 (GGML_TYPE_F32), 387
- GGML_TYPE_F16 (GGML_TYPE_F32), 387
- GGML_TYPE_F32, 387
- GGML_TYPE_I32 (GGML_TYPE_F32), 387
- ggml_type_name, 129, 188, 229, 389, 390
- GGML_TYPE_Q2_K (GGML_TYPE_F32), 387
- GGML_TYPE_Q3_K (GGML_TYPE_F32), 387
- GGML_TYPE_Q4_0 (GGML_TYPE_F32), 387
- GGML_TYPE_Q4_1 (GGML_TYPE_F32), 387
- GGML_TYPE_Q4_K (GGML_TYPE_F32), 387
- GGML_TYPE_Q5_K (GGML_TYPE_F32), 387
- GGML_TYPE_Q6_K (GGML_TYPE_F32), 387
- GGML_TYPE_Q8_0 (GGML_TYPE_F32), 387
- ggml_type_size, 389
- ggml_type_sizef, 129, 188, 229, 389, 390
- ggml_unary_op_name, 207, 306, 307, 390
- ggml_unfreeze_weights, 391
- ggml_unmarshal_model, 251, 392
- ggml_upscale, 392
- ggml_used_mem, 394
- ggml_version, 394
- ggml_view_1d, 395
- ggml_view_2d, 395
- ggml_view_3d, 396
- ggml_view_4d, 397
- ggml_view_tensor, 397
- ggml_vulkan_available, 398
- ggml_vulkan_backend_name, 398
- ggml_vulkan_device_caps, 399
- ggml_vulkan_device_count, 399
- ggml_vulkan_device_description, 400
- ggml_vulkan_device_groups, 400
- ggml_vulkan_device_memory, 401
- ggml_vulkan_free, 402
- ggml_vulkan_hard_exit_available, 402, 407
- ggml_vulkan_init, 403
- ggml_vulkan_is_backend, 404
- ggml_vulkan_list_devices, 404
- ggml_vulkan_p2p_selftest, 405, 410
- ggml_vulkan_shutdown, 402, 403, 406
- ggml_vulkan_split_buffer_type, 408
- ggml_vulkan_split_mul_mat, 314, 384, 385, 409
- ggml_vulkan_split_row_ranges, 408, 410, 410
- ggml_vulkan_status, 411
- ggml_win_part, 412, 412
- ggml_win_unpart, 412
- ggml_with_temp_ctx, 413
- gguf_free, 413
- gguf_load, 414, 414
- gguf_metadata, 414
- gguf_tensor_data, 414, 415
- gguf_tensor_info, 415
- gguf_tensor_names, 416
- glance.ggmrlr_parsnip_model, 416
- iq2xs_free_impl, 42–47, 320, 417, 418, 419, 430–439
- iq2xs_init_impl, 42–47, 320, 417, 418, 419, 430–439
- iq3xs_free_impl, 42–47, 320, 417, 418, 418, 419, 430–439

- iq3xs_init_impl, [42–47](#), [320](#), [417–419](#), [419](#),
[430–439](#)
- lr_scheduler_cosine, [420](#)
- lr_scheduler_step, [420](#)
- nn_topo_sort, [421](#)
- onnx_device_info, [422](#)
- onnx_inputs, [422](#)
- onnx_load, [423](#)
- onnx_run, [424](#)
- onnx_summary, [424](#)
- optimizer_adam, [425](#)
- optimizer_sgd, [425](#)
- plot.ggml_history, [426](#)
- predict, [426](#)
- predict.ggml_functional_model
(predict.ggml_sequential_model),
[426](#)
- predict.ggml_sequential_model, [426](#)
- print.ag_tensor, [427](#)
- print.ggml_functional_model, [427](#)
- print.ggml_history, [428](#)
- print.ggml_sequential_model, [428](#)
- print.onnx_model, [429](#)
- quantize_iq1_m(quantize_iq2_xxs), [429](#)
- quantize_iq1_s(quantize_iq2_xxs), [429](#)
- quantize_iq2_s(quantize_iq2_xxs), [429](#)
- quantize_iq2_xs(quantize_iq2_xxs), [429](#)
- quantize_iq2_xxs, [42–47](#), [320](#), [417–419](#),
[429](#), [431–439](#)
- quantize_iq3_s(quantize_iq2_xxs), [429](#)
- quantize_iq3_xxs(quantize_iq2_xxs), [429](#)
- quantize_iq4_nl(quantize_iq2_xxs), [429](#)
- quantize_iq4_xs(quantize_iq2_xxs), [429](#)
- quantize_mxfp4, [42–47](#), [320](#), [417–419](#), [430](#),
[430](#), [432–439](#)
- quantize_nvfp4, [42–47](#), [320](#), [417–419](#), [430](#),
[431](#), [431](#), [432–439](#)
- quantize_q1_0, [42–47](#), [320](#), [417–419](#),
[430–432](#), [432](#), [433–439](#)
- quantize_q2_K, [42–47](#), [320](#), [417–419](#),
[430–432](#), [433](#), [434–439](#)
- quantize_q3_K(quantize_q2_K), [433](#)
- quantize_q4_0, [42–47](#), [320](#), [417–419](#),
[430–433](#), [434](#), [435–439](#)
- quantize_q4_1(quantize_q4_0), [434](#)
- quantize_q4_K(quantize_q2_K), [433](#)
- quantize_q5_0(quantize_q4_0), [434](#)
- quantize_q5_1(quantize_q4_0), [434](#)
- quantize_q5_K(quantize_q2_K), [433](#)
- quantize_q6_K(quantize_q2_K), [433](#)
- quantize_q8_0(quantize_q4_0), [434](#)
- quantize_row_iq2_s_ref
(quantize_row_iq3_xxs_ref), [435](#)
- quantize_row_iq3_s_ref
(quantize_row_iq3_xxs_ref), [435](#)
- quantize_row_iq3_xxs_ref, [42–47](#), [320](#),
[417–419](#), [430–434](#), [435](#), [436–439](#)
- quantize_row_iq4_nl_ref
(quantize_row_iq3_xxs_ref), [435](#)
- quantize_row_iq4_xs_ref
(quantize_row_iq3_xxs_ref), [435](#)
- quantize_row_mxfp4_ref, [42–47](#), [320](#),
[417–419](#), [430–435](#), [436](#), [437–439](#)
- quantize_row_q2_K_ref, [42–47](#), [320](#),
[417–419](#), [430–436](#), [436](#), [438](#), [439](#)
- quantize_row_q3_K_ref
(quantize_row_q2_K_ref), [436](#)
- quantize_row_q4_0_ref, [42–47](#), [320](#),
[417–419](#), [430–437](#), [437](#), [438](#), [439](#)
- quantize_row_q4_1_ref
(quantize_row_q4_0_ref), [437](#)
- quantize_row_q4_K_ref
(quantize_row_q2_K_ref), [436](#)
- quantize_row_q5_0_ref
(quantize_row_q4_0_ref), [437](#)
- quantize_row_q5_1_ref
(quantize_row_q4_0_ref), [437](#)
- quantize_row_q5_K_ref
(quantize_row_q2_K_ref), [436](#)
- quantize_row_q6_K_ref
(quantize_row_q2_K_ref), [436](#)
- quantize_row_q8_0_ref
(quantize_row_q4_0_ref), [437](#)
- quantize_row_q8_1_ref
(quantize_row_q4_0_ref), [437](#)
- quantize_row_q8_K_ref
(quantize_row_q2_K_ref), [436](#)
- quantize_row_tq1_0_ref, [42–47](#), [320](#),
[417–419](#), [430–438](#), [438](#), [439](#)
- quantize_row_tq2_0_ref
(quantize_row_tq1_0_ref), [438](#)
- quantize_tq1_0, [42–47](#), [320](#), [417–419](#),

[430–438](#), [439](#)
quantize_tq2_0 (quantize_tq1_0), [439](#)
rope_types, [439](#)
RunGML (RunGML.SingleCellExperiment),
 [441](#)
RunGML.SingleCellExperiment, [441](#)
set.seed, [356](#)
show, ggml_matrix-method
 (ggml_matrix-class), [253](#)
summary.ggml_sequential_model, [443](#)
tcrossprod, ggml_matrix-method
 (ggml_matrix-class), [253](#)
tidy.ggmlr_parsnip_model, [444](#)
warning, [403](#), [407](#)
with_grad_tape, [444](#)