NAME

llamafile-quantize — large language model quantizer

SYNOPSIS

`llamafile-quantize` [flags...] *`model-f32.gguf`* [*`model-quant.gguf`*] *`type`* [*`nthreads`*]

DESCRIPTION

`llamafile-quantize` converts large language model weights from the float32 or float16 formats into smaller data types from 2 to 8 bits in size.

OPTIONS

The following flags are available:

`--allow-requantize`
: Allows requantizing tensors that have already been quantized. Warning: This can severely reduce quality compared to quantizing from 16bit or 32bit

`--leave-output-tensor`
: Will leave output.weight un(re)quantized. Increases model size but may also increase quality, especially when requantizing

`--pure`
: Disable k-quant mixtures and quantize all tensors to the same type

ARGUMENTS

The following positional arguments are accepted:

`model-f32.gguf`
: Is the input file, which contains the unquantized model weights in either the float32 or float16 format.

`model-quant.gguf`
: Is the output file, which will contain quantized weights in the desired format. If this path isn't specified, it'll default to [inp path]/ggml-model-[ftype].gguf.

`type`
: Is the desired quantization format, which may be the integer id of a supported quantization type, or its name. See the quantization types section below for acceptable formats.

`nthreads`
: Number of threads to use during computation (default: nproc/2)

QUANTIZATION TYPES

The following quantization types are available. This table shows the ID of the quantization format, its name, the file size of 7B model weights that use it, and finally the amount of quality badness it introduces as measured by the llamafile-perplexity tool averaged over 128 chunks with the TinyLLaMA 1.1B v1.0 Chat model. Rows are ordered in accordance with how recommended the quantization format is for general usage.

- 18 Q6_K 5.6gb +0.0446 ppl (q6 kawrakow)
- 7 Q8_0 7.2gb +0.0022 ppl (q8 gerganov)
- 1 F16 14gb +0.0000 ppl (best but biggest)
- 8 Q5_0 4.7gb +0.0817 ppl (q5 gerganov zero)
- 17 Q5_K_M 4.8gb +0.0836 ppl (q5 kawrakow medium)
- 16 Q5_K_S 4.7gb +0.1049 ppl (q5 kawrakow small)
- 15 Q4_K_M 4.1gb +0.3132 ppl (q4 kawrakow medium)
- 14 Q4_K_S 3.9gb +0.3408 ppl (q4 kawrakow small)
- 13 Q3_K_L 3.6gb +0.5736 ppl (q3 kawrakow large)
- 12 Q3_K_M 3.3gb +0.7612 ppl (q3 kawrakow medium)
- 11 Q3_K_S 3.0gb +1.3834 ppl (q3 kawrakow small)

- 10 Q2_K 2.6gb +4.2359 ppl (tiniest hallucinates most)
- 32 BF16 14gb +0.0000 ppl (canonical but cpu/cuda only)
- 0 F32 27gb 9.0952 ppl (reference point)
- 2 Q4_0 3.9gb +0.3339 ppl (legacy)
- 3 Q4_1 4.3gb +0.4163 ppl (legacy)
- 9 Q5_1 5.1gb +0.1091 ppl (legacy)
- 12 Q3_K alias for Q3_K_M
- 15 Q4_K alias for Q4_K_M
- 17 Q5_K alias for Q5_K_M
- COPY Only copy tensors, no quantizing.

SEE ALSO

llamafile(1), *llamafile-imatrix*(1), *llamafile-perplexity*(1), *llava-quantize*(1), *zipalign*(1), *unzip*(1)