

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:

- $\hat{A} \hat{A} \hat{A} 2 Q4_0$ 3.56G +0.2166 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} \hat{A} 3 Q4_1$ 3.90G +0.1585 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} \hat{A} 8 Q5_0$ 4.33G +0.0683 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} \hat{A} 9 Q5_1$ 4.70G +0.0349 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} 10 Q2_K$ 2.63G +0.6717 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} 12 Q3_K$ alias for Q3_K_M
- $\hat{A} \hat{A} 11 Q3_K_S$ 2.75G +0.5551 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} 12 Q3_K_M$ 3.07G +0.2496 ppl @ LLaMA-v1-7B
- $\hat{A} \hat{A} 13 Q3_K_L$ 3.35G +0.1764 ppl @ LLaMA-v1-7B

- `^ ^ 15 Q4_K` alias for `Q4_K_M`
- `^ ^ 14 Q4_K_S` 3.59G +0.0992 ppl @ LLaMA-v1-7B
- `^ ^ 15 Q4_K_M` 3.80G +0.0532 ppl @ LLaMA-v1-7B
- `^ ^ 17 Q5_K` alias for `Q5_K_M`
- `^ ^ 16 Q5_K_S` 4.33G +0.0400 ppl @ LLaMA-v1-7B
- `^ ^ 17 Q5_K_M` 4.45G +0.0122 ppl @ LLaMA-v1-7B
- `^ ^ 18 Q6_K` 5.15G -0.0008 ppl @ LLaMA-v1-7B
- `^ ^ ^ 7 Q8_0` 6.70G +0.0004 ppl @ LLaMA-v1-7B
- `^ ^ ^ 1 F16` 13.00G @ 7B
- `^ ^ ^ 0 F32` 26.00G @ 7B
- `COPY` Only copy tensors, no quantizing.

SEE ALSO

`llamafile(1)`, `llamafile-perplexity(1)`, `llava-quantize(1)`, `zipalign(1)`, `unzip(1)`