Analysis of the article

[The case for 4-bit precision: k-bit Inference Scaling Laws](#thecasefor4bit)

Daniil Sergeev  
 NUST MISIS  
 Moscow  
 [daniil1706d@gmail.com](mailto:daniil1706d@gmail.com)

ABSTRACT

Recently, there has been a tendency to multiply the number of LLMs parameters, which makes it difficult to use them both for research purposes and in other fields due to incredible computational costs. To address this authors study trade-off of using different low-bit precisions for quantizing LLMs weights by developing inference scaling laws of zero-shot performance in LLMs to determine the bit-precision and model size that maximizes zero-shot performance.

The relevance of the task

Using low-bit quantization is very promising in the context of reducing inference costs without significant loss of quality. Previous literature has focused on 8-bit and 16-bit precision, and this article suggests significant progress in this area by exploring low-bit precision.

The logic and completeness of the proposed solution

The proposed solution looks quite logical, develops the ideas of previous works. It also looks complete, as it answers the proposed questions, it also leads to both positive and negative results (e.g. distribution centering).

Weaknesses of the article

Although the article looks sufficient and complete, it considers only zero-shot performance, but the quality of low-bit quantization in other tasks, such as text classification, translation, CV tasks is not measured. It is important to study the impact of low-bit quantization in various tasks and domains, since quantization when working with, for example, medical data can significantly affect quality.

The authors investigate low-bit quantization, but do so for at least 3-bit accuracy, it would be better to consider cases with fewer bits per parameter, as, for example, in this article (Shuming Ma et al., [2024](#eraofbit1)).

The authors also generalize the results of a study of several LLM families to all Large Language Models, which may not be true (e.g. LLMs with another modality (CV, etc.) can behave differently).

Strengths of the article

It is important to consider the strengths of this article, which include:

1. The authors conducted more than 35 000 experiments, which is a convincing evidence base that 4-bit quantization is almost always the best way to get a trade-off between quality and size.
2. The research includes 4 different data types (int, float, quantile quantization, dynamic exponent quantization)
3. Different methods are researched, which improve quantization precision, such as:
4. Blocking / grouping, but in this paper they use only blocking. Blocking splits the tensor into blocks of size - block\_size and quantize each block independently, which reduces the impact of outliers compared to quantizing the entire tensor at once, but using 16-bit normalization constants and a block size of 64 means, we have an extra 16 bits every 64 parameters or 16/64=0.25 bit-per-parameter additional cost for using block-wise quantization.
5. Outlier-dependent quantization through proxy quantization (instead of just clipping (we can lose information using it) they map outliers to a representative value, preserving their importance).

Interesting conclusions from the article

I would like to mention some, as it seemed to me, interesting conclusions that the authors came to:

1. For a given zero-shot performance, 4-bit precision yields optimal scaling for almost all model families and model scales.
2. Scaling curves are almost parallel, which indicates that bit-level scaling is mostly independent of scale.
3. No scaling improvements for 6 to 8-bit precision (authors suppose that it can be because of sufficient precision of weights (even for outliers) in 6 to 8-bit). Therefore, scaling behavior can only be improved by using less than 6-bit precision rather than enhancing the quantization precision through other methods.
4. Small block size improves scaling (in blocking method) (mostly for 4 bit), because the lower block size the more bits per parameter we add (because using a small block size adds a few extra bits (1024 -> 1 fp16 normalization constant, 64 -> 1 fp16 constant (and therefore 1024/64 = 16 fp16 constant in this case))).
5. Data types improve scaling. In particular, the quantile quantization and float data types provide better scaling than integer and dynamic exponent quantization. It is really interesting that integer quantization is better than float quantization for 5-bit (the float data type can be better or worse depending on the particular bit precision)
6. Outlier-dependent quantization improves stability, but not scaling. It is really important to note that Proxy quantization is only useful for 3-bit precision weights. Proxy quantization is not useful for models that are relatively stable such as 3-bit BLOOM and GPT-2 (because they do not usually have outliers in their weights). And 4-bit precision (without Proxy quantization) still provides better scaling.
7. More parameters and fewer bits (up to 4-bit (less is worse)) better than Fewer parameters and more bits (up to 16-bit (they show only 16))

Authors recommendations

Authors made following recommendations:

1. By default, use 4-bit quantization for LLM inference as it offers the total model bits and zero-shot accuracy trade-offs.
2. Use a block size of 128 or lower (when using a blocking method) to stabilize 4-bit quantization and improve zero-shot performance.
3. Use a floating point or quantile quantization data type. In some cases, integer data types might be preferable to improve inference latency depending on the implementation and hardware support.

**Ideas about future researches**

Here are some ideas about future researches based on this paper:

1. To research the impact of quantization on other tasks (other NLP tasks, CV tasks), domains and models (not only LLMs).
2. To research how the behavior of the model changes depending on quantization (e.g. which neurons in MLP layers light up with quantization and which do not), based on this article (Elena Voita et al., [2019](#elenavoita)).
3. The [original article](#thecasefor4bit) mentions quantization only at the inference stage, but we should also check how the model behaves with quantization at the pretrain stage.
4. Try to quantize in 1-bit precision more than 4B models (because [in the original article](#eraofbit1), they use maximum 4B models to quantize them in 1-bit)

And the next idea I want to describe more thoroughly (in my opinion it is a most promising study based on [this paper](#thecasefor4bit)):

Formulation:

The article (Yinbo et al., [2022](#metalearners)) examines how transformers can serve as meta learners for modeling MLP (or another model) weights. As a further research, the task is proposed: to investigate how various quantization methods (e.g. 8-bit, 4-bit quantization) affect the accuracy and stability of the generation these weights, and the quality of the tasks of the final model (e.g. MLP).

Hypothesis:

Quantization of the model reduces the computational costs required for training and inference the model, while the quality of the generated weights will almost not change if the quantization method and data type are chosen correctly.

Data:

In the [original paper](#metalearners), MLP is used to restore the input image (by feeding coordinates and generated weights, MLP generates 3 outputs – R, G, B – for 2D, and for 3D – density, R, G, B (NeRF)), loss – reconstruction loss, so that we can use datasets with different pictures.

New approach:

Because the first layers process low–level patterns (in the context of images, these are various sticks, simple shapes, and the last layers store more information (because their receptive fields are larger)), then we can try to quantize only the first layers in order to save as much useful information as possible.

Research methods:

1. To research training stage, fine-tuning, inference stage.
2. To use different methods of quantization, such as: post-training quantization, quantization-aware training, inference quantization, based on scaling laws researched in [this](#thecasefor4bit) article.
3. To research not only CV, but also NLP tasks (e.g. text classification).

Metrics:

Accuracy - how much does the quantization of the transformer affect the accuracy of MLP predictions compared to the unquantized transformer in classification tasks, etc.

Reconstruction loss – in the case of researching the effect of transformer quantization on the accuracy of recreating the original image using MLP.

Expected result:

It is expected that with low-bit quantization, the quality of restoring the original image (in the case of CV) will drop slightly, but not critically, while significantly reducing resource consumption.

**REFERENCES**

[1] Tim Dettmers, Luke Zettlemoyer University of Washington, *The case for 4-bit precision: k-bit Inference Scaling Laws.*

[2] S. Ma, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, J. Xue, F. Wei are with Microsoft Research. H. Wang and R. Wang are with University of Chinese Academy of Sciences, *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits*.

[3] Elena Voita, Rico Sennrich, Ivan Titov, *The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives.*

[4] Yinbo Chen and Xiaolong Wang UC San Diego, *Transformers as Meta-Learners for Implicit Neural Representations*