# Literature Review
## *EasyQuant: An Efficient Data-free Quantization Algorithm for LLMs*

Sylvain Mortgat | 316678 | sylvain.mortgat@epfl.ch
ZEPHYR

## 1 Summary

This paper introduces EasyQuant, a weight quantization algorithm that is both data-free and training-free, propitious to a rapid implementation in Large Language Models (LLMs). The method's idea is easy to understand. It is divided into two main steps: the first, *outlier isolation*, involves identifying and not quantizing weights that significantly deviate from the mean. The authors find this to be crucial for minimizing quantization error. The second step involves quantizing the remaining weights by *optimizing the quantization range* using a gradient-based approach to minimize the reconstruction error.

By eliminating the reliance on additional data, EasyQuant addresses the generalization challenges common to many existing quantization techniques that depend on calibration data, for instance GPTQ (Frantar et al., 2023).

The paper evaluates the algorithm's performance across all BLOOM and LLAMA models on perplexity-based (WikiText2, PTB, C4) and zero-shot tasks (ARC-easy, Arc-challenge, StoryCloze) on INT4 quantization. EasyQuant's performance is benchmarked against traditional methods such as RTN (Round To Nearest) and GPTQ. The authors find that their algorithm consistently surpasses RTN and outperforms GPTQ in most scenarios but not in all of them. In addition, when compared to their floating-point16 (fp16) unquantized counterparts, the quantized models exhibit a relative perplexity increase ranging from 1% to approximately 7% for BLOOM models and from 4% to 39% for LLAMA models. The latter models consistently exhibit their worse results on PTB.

## 2 Strengths

**In Brief** The ideas proposed in this paper are relevant and well motivated with respect to the current quantization challenges of LLMs. Indeed, EasyQuant does not require additional data. By removing potential bias sources, its generalization capacities are guaranteed. The authors also provide novel empirical insights on outlier isolation as well as a way to optimize the quantization range using a gradient method.

**Detailed version** The paper provides interesting insights on the role of outliers in weight quantization for LLMs. While previous research has explored the significance of outliers (Zhu et al., 2023), this study reveals new insights about a distinct mechanism by which outliers influence model performance.

Importantly, outlier isolation (OI) is key but not sufficient, quantization range optimization (QRO) is required to achieve the best performances. The authors demonstrate that outliers act as a "*gating mechanism*" on the quantization range optimization instead of having a direct influence. Indeed, they observe that the perplexity obtained by pruning 1% of the largest weights is the same as that obtained by pruning 1% of the median weights. For them, this shows that there is an indirect mechanism at play when isolating outliers. Another observation highlights the interplay between OI and QRO. When only performing QRO, the model achieves a much worse performance under a small reconstruction error. However, adding OI to QRO, the perplexity of the quantized LLM decreases jointly with the reconstruction error. This behavior is best shown in Figure 2 of the paper.

In addition to these empirical findings, the authors are also interested in the practical deployment of their algorithm. Notably, the overall latency induced by the outlier isolation step is small. By implementing EasyQuant in parallel and running it on 8×A100 GPUs, it "*can finish the quantization for a 176B-sized model within 10 minutes*". In comparison, this is faster than the four GPU-hours[1]

[1] GPTQ only requires one A100 GPU.

required by GPTQ to quantize a model of the same size (Frantar et al., 2023).

## 3 Weaknesses

**In Brief** While the paper is great overall, its impact could have been strengthened with a broader range of experiments. Additionally, the absence of an open-source implementation of EasyQuant limits the opportunity for replication and further exploration by the community.

**Detailed Version** Overall, the paper would benefit from being more detailed and comprehensive regarding some of the choices made. Given the absence of an open-source version of the algorithm, the parameters (learning rate, choice of optimizer for QRO) appear to be underspecified or subjectively determined. An appendix addressing these choices would have been appreciated. Still, it should be possible to reproduce the presented work.

This lack of comprehensiveness is also felt in the experiments conducted. Although the authors test their algorithm on all BLOOM and LLAMA models, they limit their testing to INT4 quantization and focus primarily on evaluations based on perplexity. The performance of EasyQuant would be more convincing if it included additional types of quantization (lower bits cases), explored more models, and utilized a broader range of metrics. Even though the exploration of the distribution of outliers is interesting, it also suffers from incompleteness, as it is only performed on BLOOM7B.

Other comments are more about the paper's form than its substance. The $\mathrm{clamp}$ operator is not defined in Equation 1 (but all other quantities are). There is a typo in the penultimate equation of the gradient derivation in Section 4.1. Inside the inner product, it should be $\lfloor x/s \rceil$ instead of $\lfloor x_i/s \rceil$.

**Ethics** The paper does not present any ethical considerations. This is probably due to the claim that EasyQuant is not biased by additional training data. But, as shown by Li et al. (2024) quantization can influence the ethical behavior of some models. It would be interesting to conduct the same experiments as Li et al. (2024) with EasyQuant, testing different models on ETHICS, TruthfulQA and AdvGLUE.

## References

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Evaluating quantized large language models.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models.