Partial Quantization: Improving Text Generation over Uniform Quantization

Minhyuk Kim*, Heuiseok Lim*† ‡
*Department of Computer Science and Engineering, Korea University, †Human-Inspired AI Research
{mhkim0929, limhseok}@korea.ac.kr

Abstract

Large language models (LLMs) have achieved remarkable success in natural language processing but are often constrained by their significant computational requirements and large model sizes. Quantization offers a promising solution by reducing the bit-width of model parameters, thus decreasing memory and computational demands. This study investigates partial quantization, which applies quantization selectively to specific layers while preserving others at their original precision. Using the KULLM 3 model, which has 48 layers and is based on the Solar (10.7B) model, we tested applying 4-bit quantization to different sections while keeping others in FP16. Our results show that preserving layers 16-31 achieved the best overall quality score compared to a uniformly 8-bit quantized model. This indicates that partial quantization can effectively optimize both performance and efficiency.

Keywords: Quantization, Text Generation

1. Introduction

The use of large language models (LLMs) has surged in the field of natural language processing, delivering remarkable results across a wide range of applications. However, these models often characterized by their massive parameters, demand significant computational resources and prolonged training times. This presents a substantial challenge during deployment and operation, particularly in resource-constrained environments where the use of such models may be severely limited.

A promising solution to address this issue is quantization. Quantization reduces the resource requirements by representing model parameters with fewer bits, enabling more efficient model operation with reduced computational overhead [1]. When applied appropriately, quantization can greatly improve computational efficiency while minimizing performance degradation, making it a crucial technique for enhancing the practicality of large language models.

Recent studies have suggested that not all layers of a model contribute equally to its output [2]. From this perspective, applying uniform quantization across the entire model can lead to inefficiencies and performance loss. To address this, the proposed approach introduces partial quantization, focusing on preserving the language model's ability in text generation. This method prioritizes applying quantization in a way that minimizes performance degradation while maintaining the model's language generation capabilities.

‡Corresponding author

2. Related Work

Quantization techniques are primarily divided into two types: Post Training Quantization (PTQ) and Quantization Aware Training (QAT). PTQ applies quantization after model training, which is simpler but can lead to significant performance degradation [3]. In contrast, QAT integrates quantization during training, allowing the model to adapt and reduce performance loss [4].

This study extends these methods by investigating partial quantization, which selectively quantizes certain parts of the model while preserving others, to balance efficiency and performance in text generation tasks.

3. Experiments

3.1 Model

KULLM 3: In this experiment, we utilized the KULLM 3 model, which is derived from the Solar (10.7B) model and

Table 1. Comparison of Partial Quantization (Preserved Layers) and Uniform Quantization(8-bit)

Criteria	Original	Layers 0-15	Layers 16-31	Layers 32-47	Uniform 8-bit
Fluency	4.87	4.81	4.85	4.88	4.77
Coherence	4.85	4.75	4.81	4.80	4.36
Accuracy	4.44	4.33	4.41	4.36	4.43
Completeness	4.52	4.38	4.45	4.40	4.48
Overall Quality	4.57	4.44	**4.52**	4.46	4.48

instruction-tuned for Korean. The model comprises 48 layers, each represented in fp16 format [5].

3.2 Experimental Setup

The model was divided into three sections: layers 0-15, 16-31, and 32-47. In the first configuration, layers 0-15 were retained in their original FP16 format, while the remaining layers (16-47) were quantized to 4-bit. This process was repeated for the other sections, where each respective set of layers was kept in FP16, while the rest were quantized to 4-bit. Performance was assessed for each configuration, and in all cases, the model size was identical to that of the fully 8-bit quantized model, which is 10.7GB.

3.3 Evaluation

The task involved generating responses based on given prompts. The outputs generated by the model were evaluated using GPT-4, which rated each response on four metrics: fluency, coherence, accuracy, and completeness. Each metric was scored on a 5-point scale, and the average of these scores was computed to produce an overall score for each response.

3.4 Result

The results in Table 1 show that preserving layers 16-31 achieved the highest overall quality score (4.52), outperforming the fully 8-bit quantized model. This indicates that selectively preserving certain layers in FP16 can mitigate performance degradation in text generation, particularly in fluency and coherence, where the fully 8-bit quantized model saw the largest drop.

4. Conclusion

In conclusion, our experiments confirm that applying partial quantization to a model can be effective in maintaining performance while reducing model size. We intend to advance partial quantization using layer-wise quantization strategies in future research, with the goal of achieving optimal efficiency and output quality in text generation tasks.

Acknowledgements

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT). (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

Reference

[1] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, Vol. 36, 2024.

[2] S. Fan, X. Jiang, X. Li, X. Meng, P. Han, S. Shang, A. Sun, Y. Wang, and Z. Wang, "Not all layers of llms are necessary during inference," *arXiv preprint arXiv:2403.02181*, 2024.

[3] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," *arXiv preprint arXiv:2210.17323*, 2022.

[4] Z. Liu, B. Oguz, C. Zhao, E. Chang, P. Stock, Y. Mehdad, Y. Shi, R. Krishnamoorthi, and V. Chandra, "Llm-qat: Data-free quantization aware training for large language models," *arXiv preprint arXiv:2305.17888*, 2023.

[5] S. Lee, T. Lee, J. Lee, Y. Jang, and H. Lim, "Kullm: Learning to construct korean instruction-following large language models," *Annual Conference on Human and Language Technology*, pp. 196–202, 2023.