Rethinking the Bit Length of Post-Training Quantization for LLM Accuracy and Hardware Efficiency

Yanchen Li[1], Chenlin Shi[1&2], Shinobu Miwa[2], Kentaro Sano[1]

[1]RIKEN Center for Computational Science, [2]The University of Electro-Communications

Introduction

Background

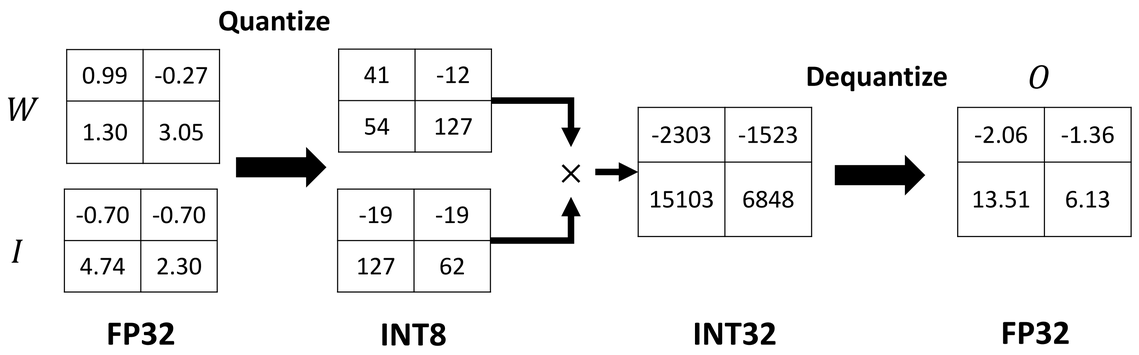
- Large language models (LLMs) consume too much memory and energy for inference
- Quantization [1] compresses model weights and activations from high precisions (e.g., FP32) to lower precisions (e.g., INT8, FP8)

Problem

- Standard 2-multiple bit lengths are often too coarse, providing a poor trade-off for preserving model accuracy.

Expected Result

- Provide fine-grain quantization precisions for higher LLM inference accuracy
- Achieve lower energy consumption or smaller area for the inference on hardware

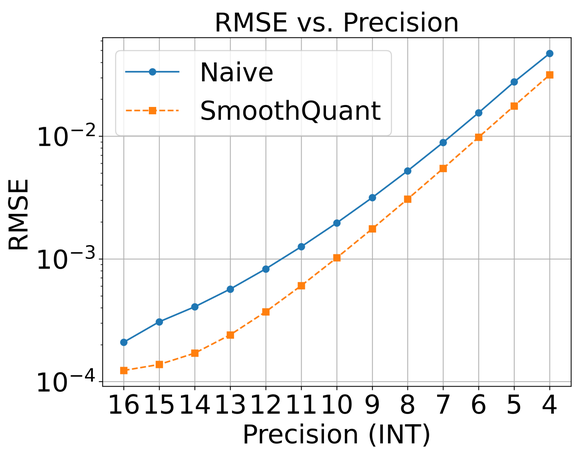


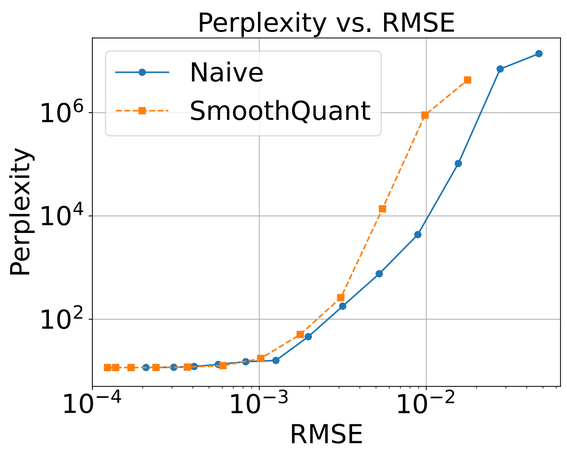
Inference Accuracy

- Eight-bit precision (INT8) is conventionally considered sufficient to preserve the inference accuracy of quantized LLMs. However, our experiments challenge this assumption.

Experiment Settings

- Large language model: Llama-3.2-8B [2]
- Task: WikiText2 [3]
- Quantization methods:
  - Naïve: $X^q = \left\lfloor \frac{X}{\Delta} \right\rceil, \Delta = \frac{\max(|X|)}{2^{N-1}-1}$
  - SmoothQaunt [4]: Before applying naïve quantization, smooths the weights and activation by $X_s = X\,\mathrm{diag}(s)^{-s}, W_s = \mathrm{diag}(s)W$
- Metrics:
  - Root mean square error (RMSE) for quantization error
  - Perplexity (cross-entropy loss) for inference accuracy

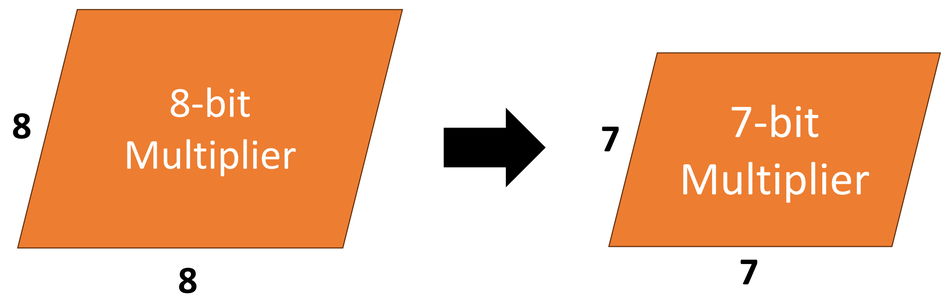




Hardware Implementation

Objectives

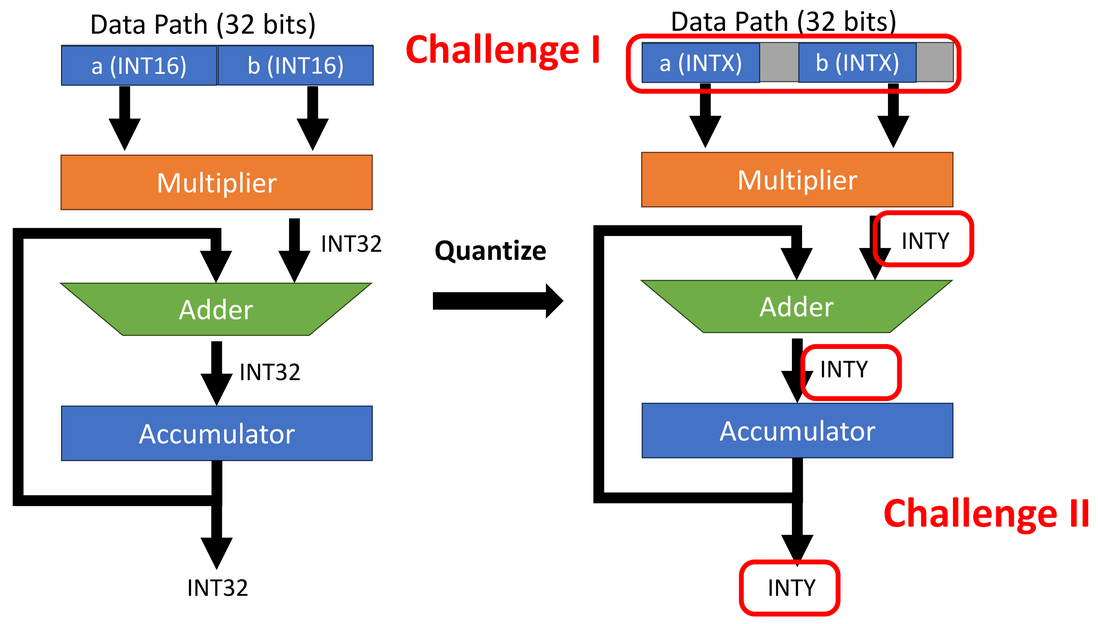
- Save energy and reduce area
  - Simplifying the multipliers in fused multiply-add (FMA)
  - Reducing data transfer



Space Complexity: $O(n^2)$

Challenges

- (I) A data interface mismatch creates a bottleneck.
  - The bandwidth of the data path is wasted without optimization
- (II) The accumulator precision requires a careful trade-off



Potential Solutions

- Design new memory interfaces according to the input lengths
- Optimized batched data transfer strategies
- Select the accumulation precision by analyzing trade-off between accumulator precision, FMA precision, and overall model accuracy

Conclusion

- Our experimental results indicate that standard INT8 precision is insufficient to preserve LLM inference accuracy.
- We show that supporting fine-grained, irregular bit lengths to maintain high LLM accuracy.
- However, implementing these irregular bit lengths efficiently in hardware presents significant challenges, which are left as our future studies.

Contact

Yanchen Li
RIKEN Center for Computational Science
Kobe City, Hyoko, Japan
yanchen.li@riken.jp

References

1. Yao, Zhewei, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers." Advances in neural information processing systems 35 (2022): 27168-27183.
2. Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).
3. Merity, Stephen, Caiming Xiong, James Bradbury, and Richard Socher. "Pointer sentinel mixture models." arXiv preprint arXiv:1609.07843 (2016).
4. Xiao, Guangxuan, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. "Smoothquant: Accurate and efficient post-training quantization for large language models." In International conference on machine learning, pp. 38087-38099. PMLR, 2023.