Compression-Aware Computing for Scalable and Sustainable AI

Zhaozhuo Xu[1]

[1]Stevens Institute of Technology
1 Castle Point Terrace
Hoboken, NJ 07030
zxu79@stevens.edu

Abstract

This talk explores the challenge of customizing large-scale AI models, particularly generative AI, on cost-effective devices with limited memory and energy resources. Modern AI models demand substantial computational power, often relying on specialized hardware such as GPUs. To address this, the talk introduces compression-aware computing, a framework enabling AI models to recognize and adapt to their compressed states while preserving performance (Xu et al. 2024). Compression-aware computing integrates compression techniques like sparsification (Liu et al. 2023; Zhou et al. 2024), quantization (Liu et al. 2024b; Zhang et al. 2024a,b), and low-rank decomposition (Liu et al. 2024a) to enhance the efficiency and accuracy of AI models, broadening these models' accessibility across diverse devices. Additionally, this talk highlights one rationale of scalable and sustainable AI in advancing Alzheimer's research by facilitating the analysis of large single-cell transcriptomics datasets for gene-gene interaction discovery (Wu et al. 2024).

References

Liu, Z.; Desai, A.; Liao, F.; Wang, W.; Xie, V.; Xu, Z.; Kyrillidis, A.; and Shrivastava, A. 2023. Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Liu, Z.; Wang, G.; Zhong, S. H.; Xu, Z.; Zha, D.; Tang, R. R.; Jiang, Z. S.; Zhou, K.; Chaudhary, V.; Xu, S.; et al. 2024a. Winner-take-all column row sampling for memory efficient adaptation of language model. *Advances in Neural Information Processing Systems*, 36.

Liu, Z.; Yuan, J.; Jin, H.; Zhong, S.; Xu, Z.; Braverman, V.; Chen, B.; and Hu, X. 2024b. KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Wu, Y.; Yang, Y.; Liu, Z.; Li, Z.; Pahwa, K.; Li, R.; Zheng, W.; Hu, X.; and Xu, Z. 2024. Weighted Diversified Sampling for Efficient Data-Driven Single-Cell Gene-Gene Interaction Discovery. *arXiv preprint arXiv:2410.15616*.

Xu, Z.; Liu, Z.; Chen, B.; Zhong, S.; Tang, Y.; Wang, J.; Zhou, K.; Hu, X.; and Shrivastava, A. 2024. Soft Prompt Recovers Compressed LLMs, Transferably. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhang, T.; Yi, J.; Xu, Z.; and Shrivastava, A. 2024a. KV Cache is 1 Bit Per Channel: Efficient Large Language Model Inference with Coupled Quantization. *Advances in Neural Information Processing Systems*.

Zhang, T.; Yi, J. W.; Yao, B.; Xu, Z.; and Shrivastava, A. 2024b. Nomad-attention: Efficient llm inference on cpus through multiply-add-free attention. *Advances in Neural Information Processing Systems*.

Zhou, Y.; Chen, Z.; Xu, Z.; Lin, V.; and Chen, B. 2024. Sirius: Contextual sparsity with correction for efficient llms. *Advances in Neural Information Processing Systems*.