Tensor Decomposition on Language Models

Katherine Zhao

INSPIRE Lab, Rutgers University

June 2026

Costs of Fine-Tuning Models

- Foundation models like Llama-3-8B and GPT-4 contain billions of parameters.
- Fine-tuning primarily operates through multiplication across tensors.
- Updating matrices during tuning requires sizable memory allocation.

Current Work

- Create efficient fine-tuning method on resource-constrained hardware, expanding access to models with memory constraints.
- Evaluate cost of inference and precision to ensure quality is relatively maintained with reduction in computation cost.

Citations

- Hu, Edward J., et al. “Lora: Low-Rank Adaptation of Large Language Models.” arXiv.Org, 16 Oct. 2021, arxiv.org/abs/2106.09685.
- Taki, Batoul, et al. “Structured Low-Rank Tensors for Generalized Linear Models.” arXiv.Org, 5 Aug. 2023, arxiv.org/abs/2308.02922.

Acknowledgments

This work was carried out as a part of the 2026 DIMACS REU program at Rutgers University, supported by NSF Grant CCF-2447342.