Efficient LLM Adaptation with Forward-Only Passes

Baichuan Huang[1], Ananth Balashankar[2], Amir Aminifar[1]

[1]Department of Electrical and Information Technology, Lund University, Sweden
[2]Google DeepMind, USA

Motivation

Artificial intelligence (AI) serving is emerging as a key driver of growth in energy consumption as AI systems are deployed at scale, particularly for Transformer-based large language models (LLMs) inference serving. Large-scale LLM inference generated from real user interactions during LLM serving **motivates us to rethink how the inference process itself** can be leveraged to improve the serving system, particularly **enabling efficient task-specific LLM adaptation**.

Method

Our forward-only approach includes two steps, i.e., forward-only search and rank reduction. The proposed forward-only search, as shown in Fig.1, contains ① Singular Value Decomposition (SVD) with multiple SVD components; ② The sum of activations from all SVD components recovers the standard forward pass; ③ Forward-only passes measure the angle concentration to identify which weight matrix has the highest dispersion that should be applied with rank reduction.

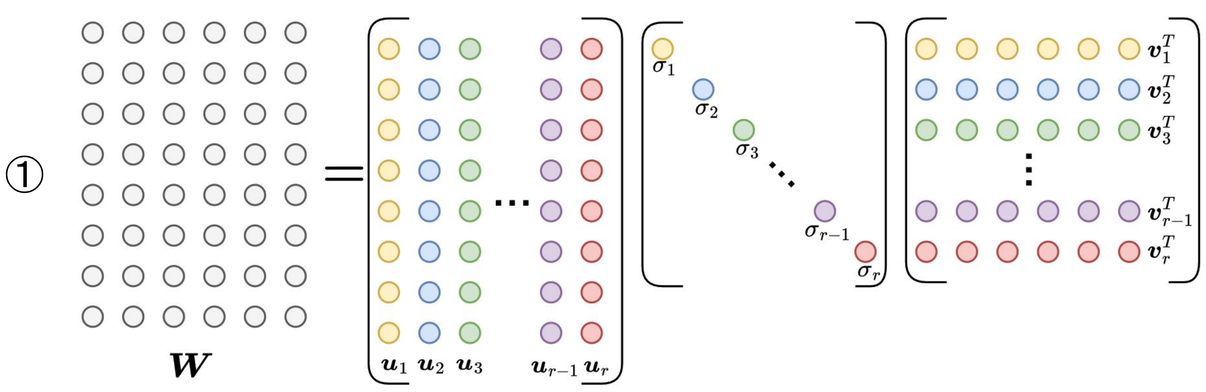


Fig. 1: Efficient LLM Adaptation in LLM inference serving.

Our forward-only search eliminates the need for backpropagated gradients [1] and avoids the computational burden of exhaustive search by sweeping all the weights [2]. After the forward-only search, the rank-reduction strategy is determined by the accuracy evaluated on validation datasets.

Empirical Evaluation

Low Memory & Greater Speedup & Competitive Acc

Our forward-only approach consumes lower peak memory and greater speedup than LASER-Grad [1], and achieves faster computation than LASER [2], while maintaining comparable or even superior downstream accuracy, as shown in Fig. 2.

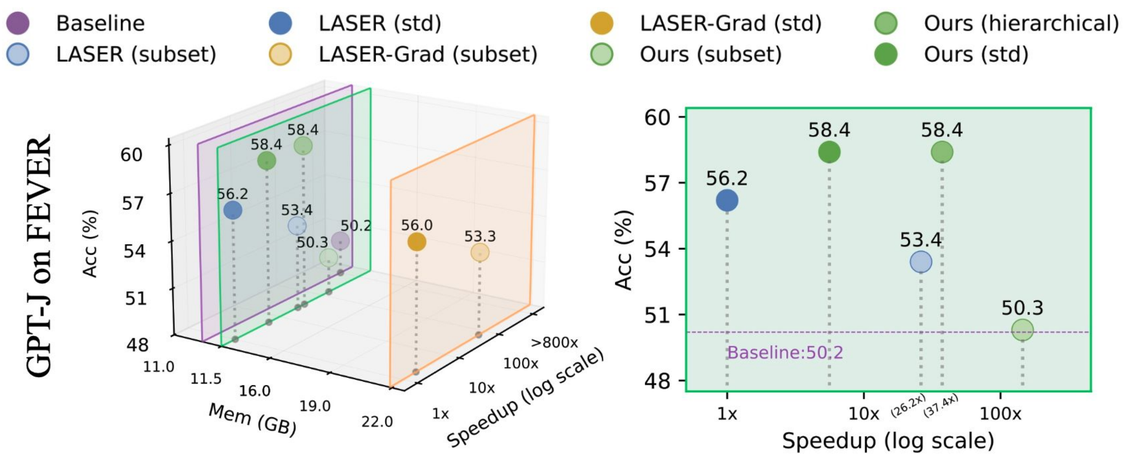


Fig. 2: Experiments across LASER, LASER-Grad, and Ours.

Enhancing Fine-Tuning

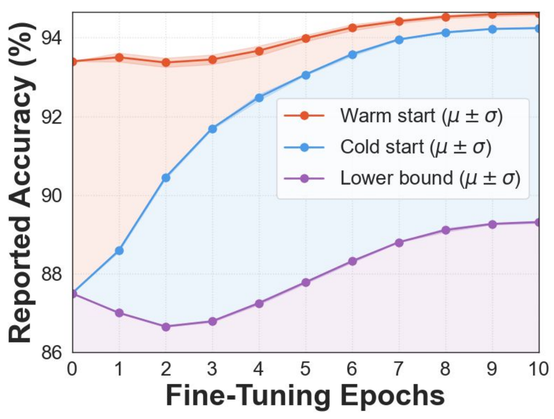


Fig. 3: Final candidate with rank reduction (warm start) from our forward-only approach enhances fine-tuning process. At the early fine-tuning epochs, the warm start performs better than the cold start (cold start is for the final candidate without rank reduction).

Extension to various Models

Epistemic Reasoning	Baseline	Ours	
	Acc	Acc	Speedup
Gemma4-E2B-it	48.5%	**60.2%**	**11.6x**
Qwen2-57B-A14B*	39.9%	**62.1%**	**13.9x**

Table 1: The generalization of our approach.

(refer to our paper for more results)

References

[1] Sreeram, Shiva, et al. Compress to Impress: Efficient LLM Adaptation Using a Single Gradient Step on 100 Samples. NeurIPS, 2025.

[2] Sharma, Pratyusha, et al. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. ICLR, 2024.