Computational Research in Boston and Beyond Seminar

Extremely-efficient fine-tuning of LLMs

Friday, April 4, 2025
12:00 PM - 1:00 PM
Zoom: https://mit.zoom.us/j/91933017072

Praneeth Vepakomma
(MIT/IDSS and MBZUAI)

Abstract

Large Language Models (LLMs) have reshaped generative AI, but fully fine-tuning these massive architectures is quite expensive in computational and communication resources. Low-Rank Adaptation (LoRA) partially mitigates these challenges, yet conventional LoRA often struggles to match the performance of full fine-tuning. In this talk, I introduce **LoRA-SB (LoRA Silver Bullet)**, a novel approach that injects a constrained update space into LoRA's framework, enabling optimal scaling for high-rank gradient directions that mimic full fine-tuning in a low-rank space, and meets the performance of full fine-tuning. We theoretically prove that our initialization strategy provides an optimal low-rank approximation of the initial gradient and preserves critical update directions throughout training. Extensive experiments on mathematical reasoning, commonsense inference, and language understanding tasks show that LoRA-SB exceeds the performance of standard LoRA while requiring 27–90× fewer trainable parameters and comprehensively outperforms LoRA-XS. Our findings demonstrate that it is not only possible but also highly effective to simulate full fine-tuning in low-rank subspaces, offering significant efficiency gains at no loss in accuracy. Additionally, we introduce **Fed-SB**, a federated extension of LoRA-SB that employs direct averaging of the small matrix R to guarantee exact updates and drastically reduce communication costs—independent of the number of clients—by up to 230×. Fed-SB further enhances privacy-utility-communication efficiency trade-offs by lowering noise requirements and avoiding noise amplification. Overall, it establishes a new Pareto frontier for efficient, scalable federated fine-tuning in both private and non-private settings.