Efficient Fine-Tuning Techniques for Large Language Models

Yunus Emre Demirdağ
Big Data and Artificial Lab, Fırat University, Elazığ, Türkiye

Betül AY
Associate Professor, Fırat University, Elazığ, Türkiye

Abstract

Large Language Models (LLMs) have gained significant attention, particularly following the release of GPT-3. With the emergence of open-source alternatives, the need for efficient fine-tuning methods has become increasingly important. However, fine-tuning LLMs presents several challenges, including high storage requirements due to large parameter sizes and computational inefficiencies that slow down the training process. Addressing these challenges is crucial for making fine-tuning more accessible and cost-effective.

To overcome these issues, various **Parameter-Efficient Fine-Tuning (PEFT)** techniques have been developed. These methods enable fine-tuning with significantly fewer parameters, reducing memory usage, improving training speed, and lowering computational costs. Among the most prominent PEFT techniques are **Prompt Tuning, Low-Rank Adaptation (LoRA), Adapters, and Prefix Tuning**.

- **Prompt Tuning** modifies input prompts rather than model weights, enabling efficient adaptation with minimal computational overhead.
- **LoRA** introduces low-rank matrix updates to pre-trained models, significantly reducing the number of trainable parameters.
- **Adapters** add small trainable modules between model layers, allowing targeted fine-tuning without altering the entire model.
- **Prefix Tuning** conditions the model on learnable prefix embeddings, optimizing performance while keeping most model weights frozen.

These techniques provide scalable solutions for fine-tuning LLMs, making them more practical for real-world applications by addressing storage, speed, and efficiency challenges. This paper explores these approaches in depth, highlighting their advantages and trade-offs in fine-tuning large-scale models.

Keywords

Large Language Models (LLM), Fine-Tuning, Parameter-Efficient Fine-Tuning (PEFT), Compute Capability, Memory Usage.