AdaMoLE: Fine-Tuning Large Language Models with Adaptive Mixture of Low-Rank Adaptation Experts

Zefang Liu (Georgia Institute of Technology), **Jiahua Luo** (University of Macau)
Contact: liuzefang@gatech.edu

Introduction

- **Objective:** Introduce **AdaMoLE**, a novel method for fine-tuning Large Language Models (LLMs) using an adaptive mixture of experts (MoE) with the Low-Rank Adaptation (LoRA).
- **Problem:** Existing fine-tuning methods such as static top-k expert selection do not adapt to the varying complexity of tasks in LLMs.
- **Solution:** AdaMoLE dynamically activates LoRA experts based on task complexity, ensuring adaptive fine-tuning for a wide range of natural language tasks.

Methodology

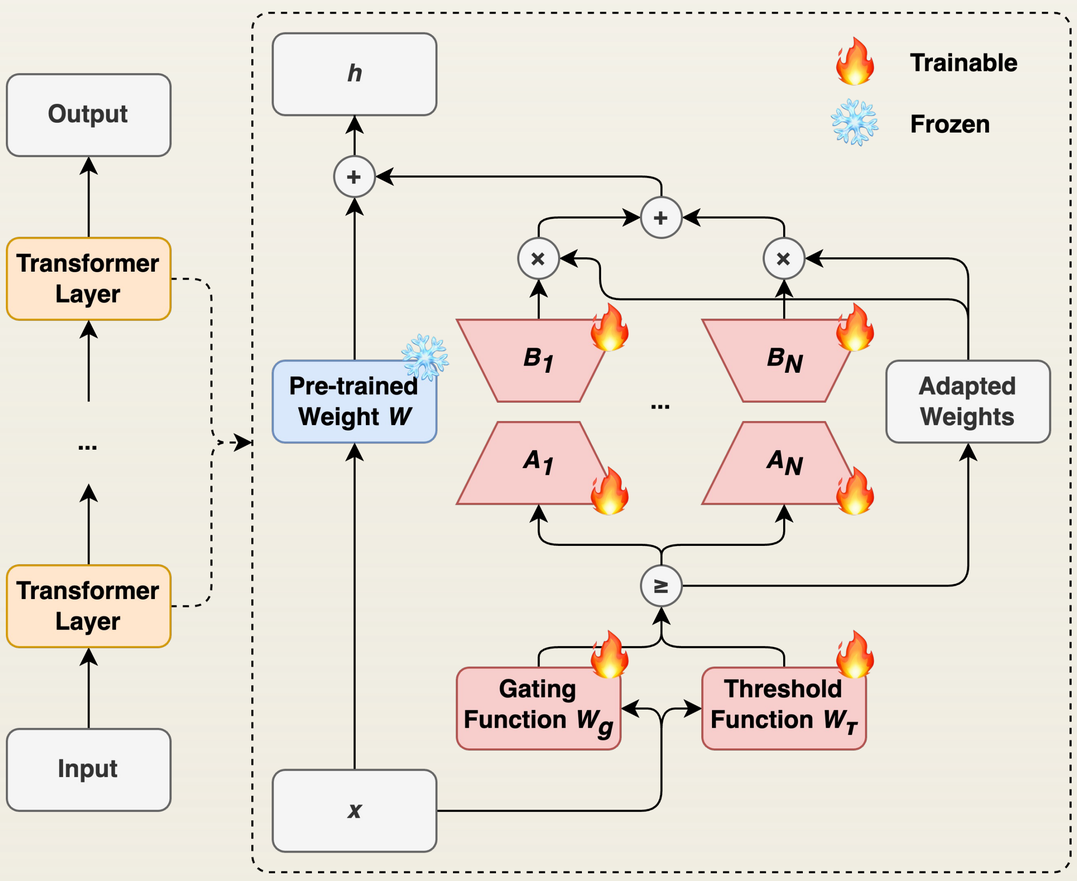
Core Concept:
AdaMoLE replaces static LoRA modules with multiple LoRA experts and an adaptive gating function, allowing the model to select the most appropriate experts based on the input context.

Key Components:

- **Low-Rank Adaptation (LoRA):** Efficiently fine-tunes model parameters using low-rank matrices.
- **Mixture of Experts (MoE):** Distributes input across multiple experts, each specializing in different aspects of the task.
- **Adaptive Threshold Network:** Dynamically adjusts the number of activated experts by setting a threshold based on input complexity. The adaptive threshold is computed as:

$$\tau = \tau_{\max} \cdot \sigma(W_\tau x + b_\tau),$$

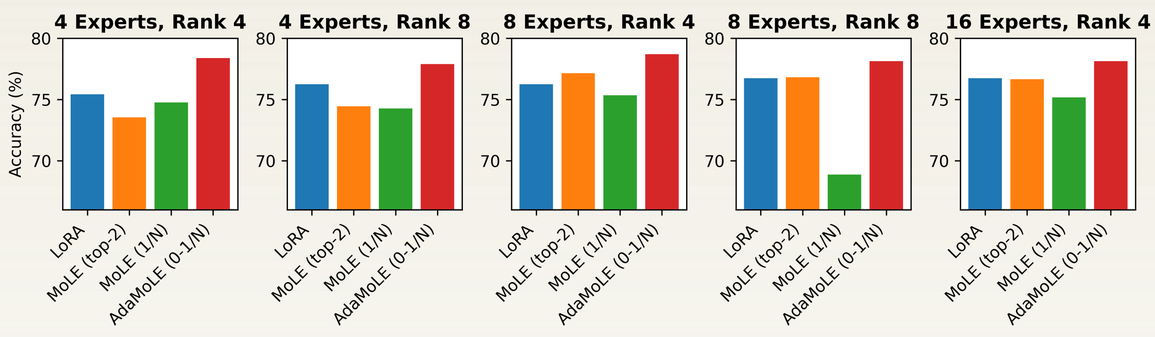
where σ is the sigmoid activation function, W_τ and b_τ are learnable parameters.



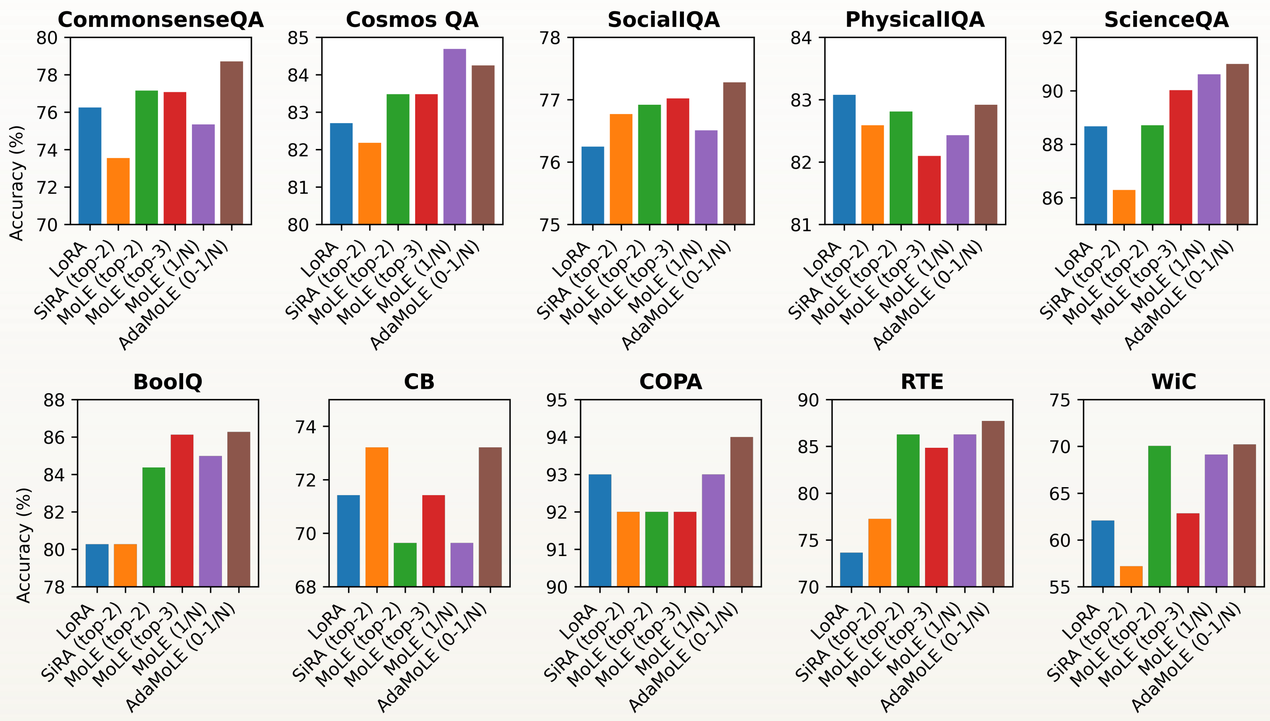
Experimental Settings

- **Baseline Models:** Low-Rank Adaptation (LoRA), Mixture of LoRA Experts (MoLE) with top-k expert selection or hard thresholding, and Sparse Mixture of Low-Rank Adaptation (SiRA).
- **Datasets:** commonsense reasoning tasks (CommonsenseQA, SocialIQA, etc.) and NLP tasks (SuperGLUE benchmarks).

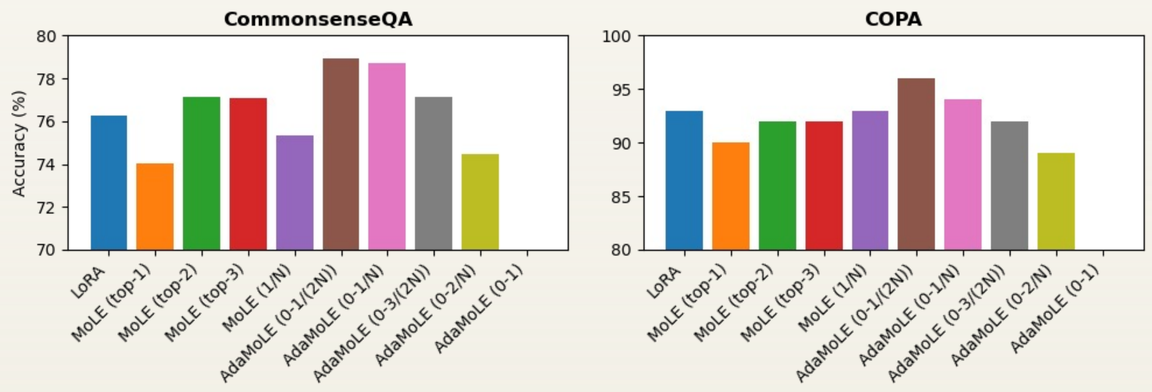
Numbers of Experts



Experimental Results

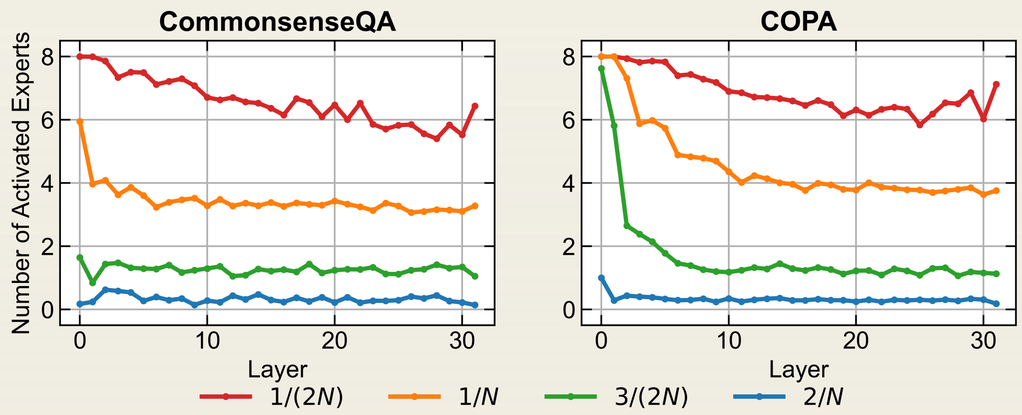


Threshold Sensitivity



AdaMoLE achieves the highest accuracy with a dynamic threshold range of [0, 1/(2N)], but a balance between performance and computational load can be made by adjusting the threshold.

Expert Activation



Most experts are activated in the lower layers, with refined engagement at higher layers, reducing computational load while maintaining task performance.

Conclusion

- **Conclusion:** AdaMoLE demonstrates superior performance and adaptability compared to existing MoE-based fine-tuning methods by dynamically selecting experts based on the task complexity.
- **Future Work:** Further investigation into expert interaction and minimizing computational overhead.

References

Hu, E. J., et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*.
Shazeer, N., et al. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*.
Fedus, W., Zoph, B., Shazeer, N. (2022). *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*.
Wu, X., Huang, S., Wei, F. (2023). *Mixture of LoRA Experts*.
Liu, Z., Luo, J. (2024). *AdaMoLE: Fine-Tuning Large Language Models with Adaptive Mixture of Low-Rank Adaptation Experts*.

SCAN ME