Efficient Fine-Tuning for Medical Question-Answering with Selective Unfreezing and Parameter-Efficient Techniques

Adar Schwarzbach
Department of Computer Science
Duke University
adar.schwarzbach@duke.edu

December 9, 2024

Abstract

Fine-tuning large language models for domain-specific tasks often demands significant computational resources. We explore three distinct fine-tuning approaches on a medical question-answering dataset using the Llama-1B model from the Llama 3 herd of models (Touvron et al., 2023; Grattafiori et al., 2024): Selective Parameter Unfreezing (Lee et al., 2019), LoRA (Low-Rank Adaptation) (Hu et al., 2021), and BOFT (Orthogonal Butterfly Fine-Tuning) (Liu et al., 2024). These methods were evaluated on perplexity and next-token accuracy. Our results demonstrate that Selective Unfreezing achieves the best performance, while LoRA and BOFT provide resource-efficient alternatives with trade-offs in accuracy. This study highlights the efficacy of selective fine-tuning in resource-constrained settings.

1 Introduction

Domain adaptation of large language models (LLMs) is a computationally intensive task due to the sheer scale of parameters involved. Parameter-efficient fine-tuning (PEFT) techniques, such as LoRA (Hu et al., 2021) and BOFT (Liu et al., 2024), have been developed to reduce memory usage while achieving competitive performance. We investigate the application of these techniques alongside Selective Parameter Unfreezing to fine-tune the Llama-1B model for a medical question-answering task. Our study compares the approaches based on perplexity and next-token accuracy, providing insights into their trade-offs and applicability.

2 Methods

2.1 Dataset and Preprocessing

We utilized the MedQuAD (Medical Question Answering Dataset) (Afroz, 2023), a comprehensive dataset of medical question-answer pairs. MedQuAD contains questions curated from 12 trusted National Institutes of Health (NIH) websites, such as cancer.gov and the Genetic and Rare Diseases Information Resource (GARD). These sources ensure the dataset covers a wide range of health-related topics, including treatments, diagnoses, symptoms, and risk factors.

For this study, we focused on the main question-answer pairs provided in the dataset. Rows with missing or empty fields were excluded, ensuring data quality. The cleaned dataset was then split into 80% training and 20% testing subsets.

To prepare the data for fine-tuning, each entry was tokenized using a pre-trained tokenizer from the Hugging Face library. The question and answer fields were concatenated into a single input sequence, with a separator token inserted between the question and answer to preserve context. The resulting sequences were padded or truncated to a fixed length to ensure consistent input dimensions for the model.

2.2 Fine-Tuning Approaches

We explored three approaches for fine-tuning the Llama-1B model:

- **Selective Parameter Unfreezing:** Only the final transformer block and language model head (`lm_head`) were unfrozen for training, while the rest of the model remained frozen (Lee et al., 2019).
- **LoRA:** Low-Rank Adaptation added trainable low-rank matrices to specific layers (`q_proj` and `v_proj`) while freezing the rest of the model (Hu et al., 2021).
- **BOFT:** Orthogonal Butterfly Fine-Tuning applied structured transformations to selected layers (`q_proj` and `v_proj`) with trainable parameters in a subset of transformer blocks (Liu et al., 2024).

2.3 Training Setup

All models were trained for 3 epochs using the AdamW optimizer with BF16 precision for computational efficiency. Training and evaluation were performed on an NVIDIA A100 GPU.

3 Results and Discussion

We evaluated the fine-tuned models based on perplexity (lower is better) and next-token accuracy (higher is better). Table 1 summarizes the results.

Table 1: Performance Comparison of Fine-Tuning Techniques

Technique	**Perplexity**	**Next-Token Accuracy**
Selective Unfreezing	**2.9599**	**0.7329**
LoRA	4.6205	0.6433
BOFT	3.2522	0.7146
Base Model	6.1104	0.5907

3.1 Selective Parameter Unfreezing

This approach provided the best performance, achieving the lowest perplexity and highest next-token accuracy. By focusing training on the final transformer block and `lm_head`, the model effectively adapted to the task-specific data with minimal computational overhead.

3.2 LoRA

LoRA significantly reduced memory usage by introducing low-rank matrices in attention layers. While computationally efficient, it resulted in higher perplexity and lower accuracy compared to Selective Unfreezing.

3.3 BOFT

BOFT balanced performance and efficiency, outperforming LoRA in both perplexity and accuracy but falling short of Selective Unfreezing. Its structured approach to parameter-efficient fine-tuning makes it a viable choice in resource-constrained scenarios.

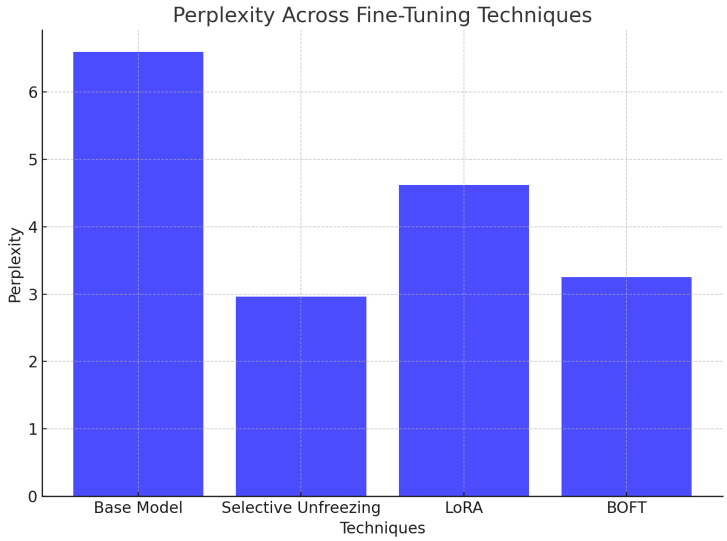

Figure 1: Perplexity comparison of fine-tuning techniques. Lower perplexity indicates better model performance.

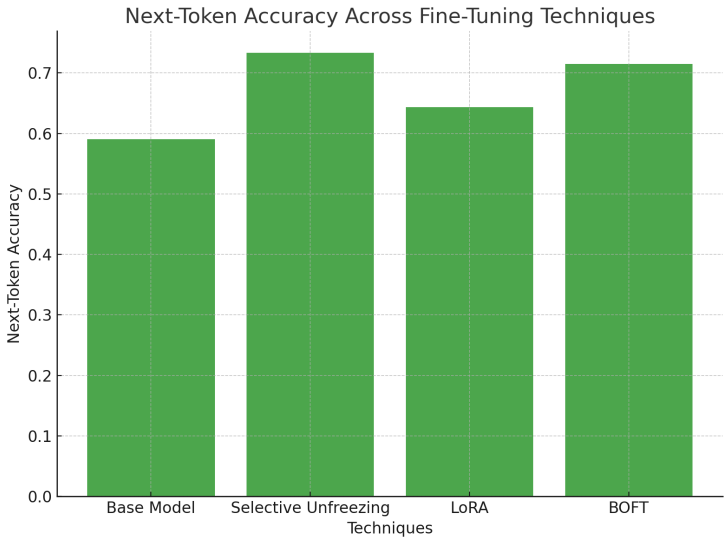

Figure 2: Next-token accuracy comparison of fine-tuning techniques. Selective Unfreezing achieves the best accuracy.

4 Conclusion

Our study demonstrates that there are several less computationally intensive fine-tuning strategies that can boost performance in large language models on domain-specific tasks. Altough Selective Parameter Unfreezing was the most effective technique in this exploration, LoRA and BOFT offered strong results as well.

References

1. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. Retrieved from https://arxiv.org/abs/2302.13971.

2. Hu, E. J., Shen, D., Wallis, P., Allen-Zhu, Z., Li, S., Wang, L., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. Retrieved from https://arxiv.org/abs/2106.09685.

3. Liu, W., Qiu, Z., Feng, Y., Xiu, Y., Xue, Y., Yu, L., et al. (2024). Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization. Retrieved from https://boft.wyliu.com.

4. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 Herd of Models. Retrieved from https://arxiv.org/abs/2407.21783.

5. Afroz, P. (2023). MedQuAD: Medical Question Answer Dataset for AI Research. Retrieved from https://www.kaggle.com/datasets/pythonafroz/medquad-medical-question-answer-for-ai-research.

6. Lee, J., Tang, R., Lin, J. (2019). What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. Retrieved from https://arxiv.org/abs/1911.03090.

Advisor Acknowledgment

The author would like to acknowledge the guidance and support provided by Dr. Pranam Chatterjee, Department of Computer Science, Duke University.