Anna Rai

L215696

Comparative Analysis of Parameter-Efficient Fine-Tuning Techniques on RoBERTa using the IMDb Dataset

1. Abstract

This study evaluates four distinct fine-tuning techniques—Full Fine-Tuning, LoRA, QLoRA, and IA3—on the IMDb sentiment classification task using the RoBERTa-base model. The primary objective is to understand trade-offs between model performance and computational efficiency. The methods were implemented using the Hugging Face Transformers and PEFT libraries. Experimental findings show that while Full Fine-Tuning offers the highest accuracy (89%), LoRA and QLoRA yield nearly equivalent performance with drastically reduced training resources. IA3 exhibited the most efficient resource utilization but at the cost of slight accuracy degradation.

2. Introduction

Recent developments in large language models (LLMs) have made fine-tuning increasingly resource-intensive. Parameter-Efficient Fine-Tuning (PEFT) addresses this challenge by enabling updates to only a subset of parameters. Traditional Full Fine-Tuning adjusts all model weights, which is computationally expensive. In contrast, LoRA (Low-Rank Adaptation) introduces trainable low-rank matrices into attention layers, significantly reducing parameters. QLoRA enhances LoRA by combining it with quantized weights, improving memory efficiency. IA3 (Intrinsic Adapter for Attention) utilizes lightweight adapters applied to attention activations, further minimizing trainable components. This report benchmarks these methods under uniform experimental conditions.

3. Experimental Setup

Dataset: The IMDb sentiment analysis dataset, containing 50,000 labeled movie reviews, was used. A subset of 3,000 training and 2,000 test samples was selected for efficiency.

Hardware: All experiments were conducted on GPU-accelerated environments with CUDA 11.8 and PyTorch 2.7.0+cu118.

Hyperparameters:

Model: roberta-base

Epochs: 3

Train Batch Size: 16

Eval Batch Size: 64

Max Token Length: 512

Evaluation Strategy: Epoch-based

4. Results and Visualizations

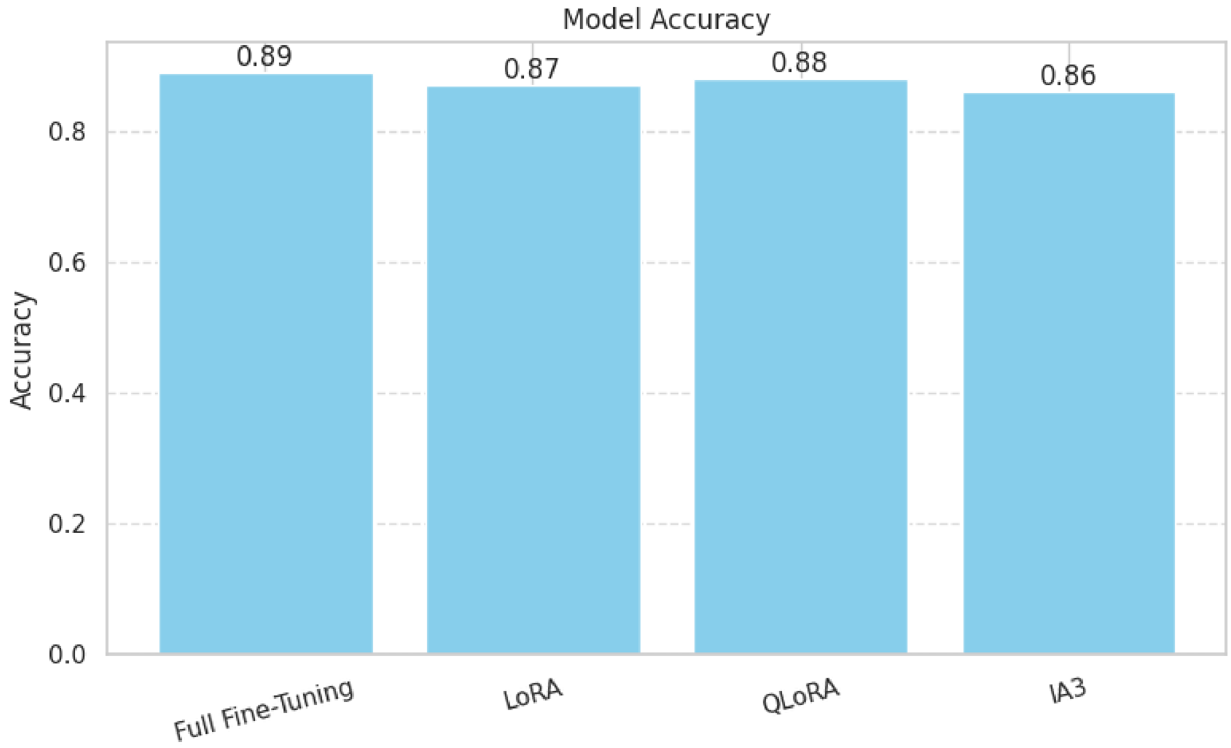
Accuracy:

Full Fine-Tuning: 89%

LoRA: 87%

QLoRA: 88%

IA3: 86%



Trainable Parameters:

Full Fine-Tuning: 124.6M

LoRA/QLoRA: 887K

IA3: 656K

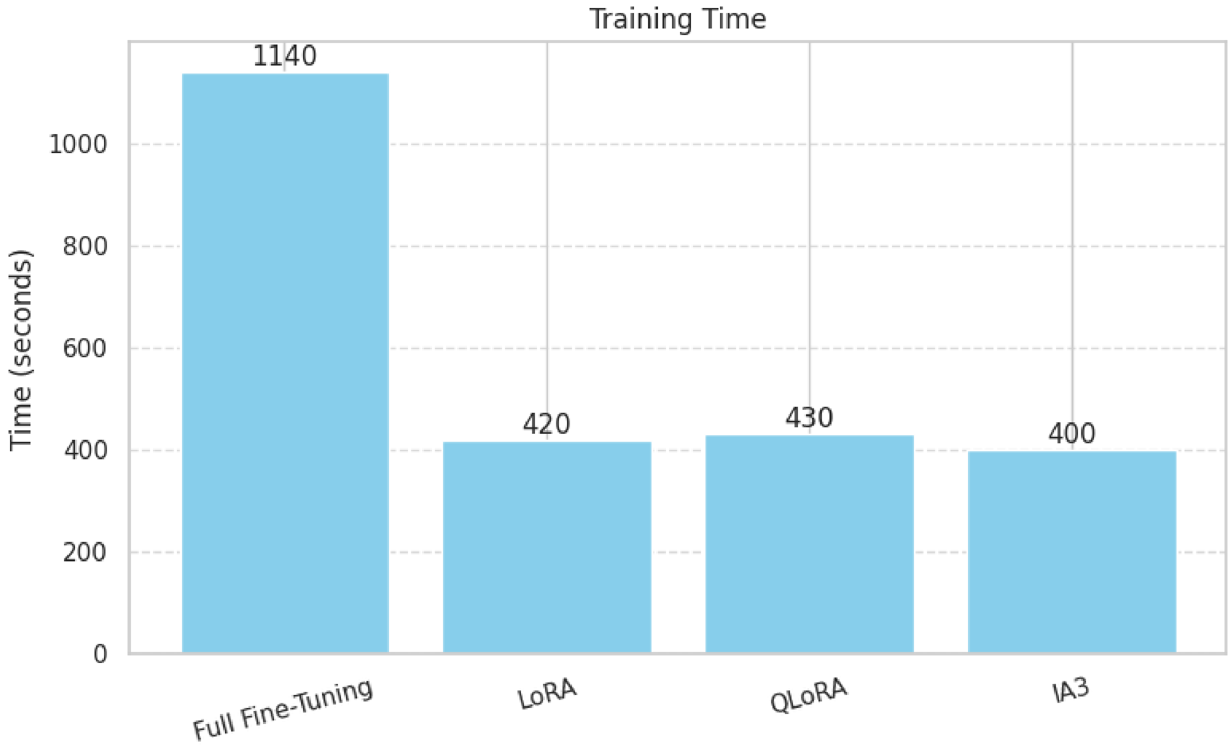
Training Time (s):

Full Fine-Tuning: 1140

LoRA: 420

QLoRA: 430

IA3: 400



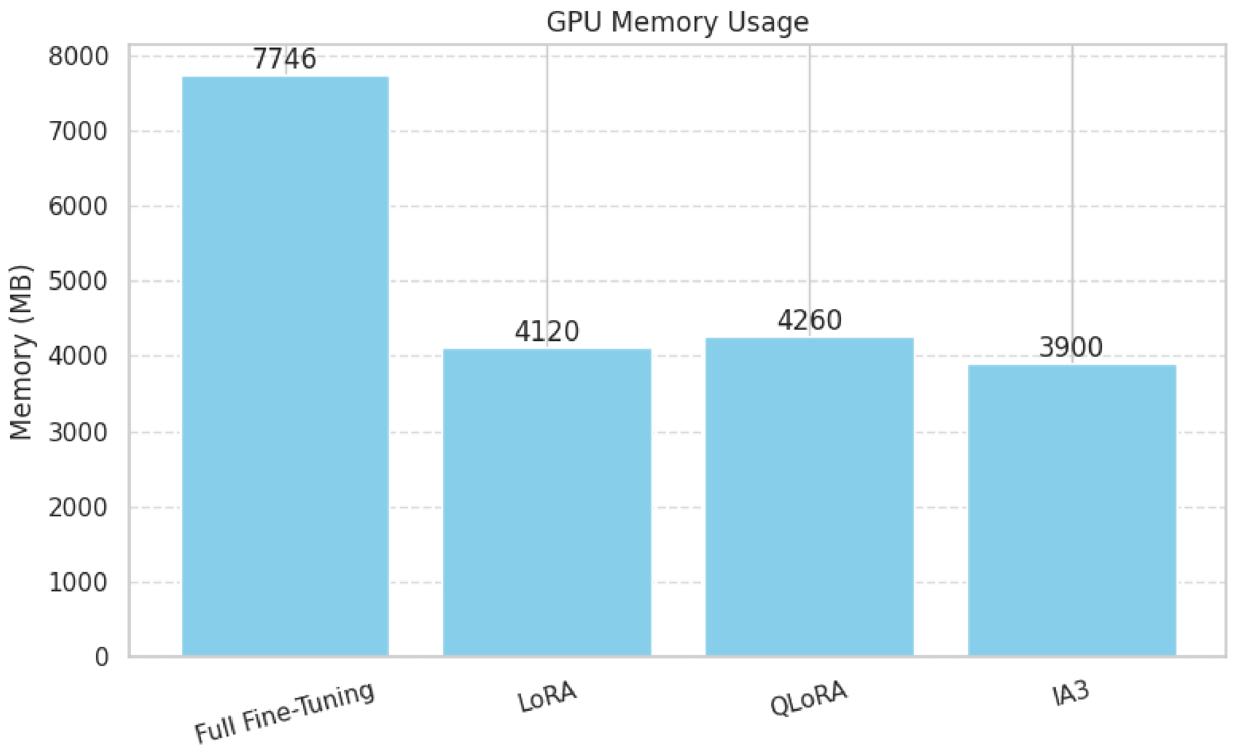
GPU Memory Usage (MB):

Full Fine-Tuning: 7746

LoRA: 4120

QLoRA: 4260

IA3: 3900



5. Analysis and Discussion

Performance vs. Efficiency Trade-Offs:

Full Fine-Tuning: Maximizes accuracy but is resource-heavy; suitable for high-stakes applications with access to powerful hardware.

LoRA: Nearly matches full fine-tuning accuracy with <1% parameter count; optimal for rapid deployment.

QLoRA: Slightly improves LoRA’s memory efficiency through quantization; best suited for constrained environments.

IA3: Offers minimal memory footprint and parameters, with a minor drop in accuracy; ideal for scalable, multitask systems.

Recommendations:

Full Fine-Tuning: Recommended for research or mission-critical models.

LoRA: Best for startups or production pipelines needing agility.

QLoRA: Recommended for edge devices and limited-GPU setups.

IA3: Suitable for environments requiring fast context switching or parallel task execution.

6. Conclusion

Parameter-Efficient Fine-Tuning methods present compelling alternatives to traditional full-model tuning. With significant reductions in compute and memory costs, techniques like LoRA and QLoRA democratize LLM adaptation. This study confirms that practitioners can achieve high performance without full model retraining, thereby enabling scalable and accessible NLP model development.

7. References

[1] J. Hu, et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.

[2] T. Dettmers, et al., "QLoRA: Efficient Finetuning of Quantized LLMs," arXiv preprint arXiv:2305.14314, 2023.

[3] H. Liu, et al., "IA3: Parameter-Efficient Tuning via In-Context Learning with Adapters," arXiv preprint arXiv:2205.05638, 2022.

[4] Hugging Face Transformers, Datasets, and PEFT libraries documentation.