Fine-Tuning Models

Submitted By:
Jawaria Asim
21L-5157

1. Abstract

This project focuses on fine-tuning and parameter-efficient fine-tuning (PEFT) methods for sentiment classification using the IMDb dataset. We compared four strategies: Full Fine-Tuning, LoRA, QLoRA, and Adapter Tuning (IA3), all based on the `roberta-base` model. While Full Fine-Tuning achieved the highest accuracy (93.3%), LoRA and QLoRA significantly reduced the number of trainable parameters and training time, with minimal drops in performance. IA3 offered a trade-off between performance and efficiency. The results emphasize that PEFT methods can offer competitive results while drastically lowering computational costs.

2. Introduction

Large language models (LLMs) like RoBERTa have transformed NLP tasks, but their fine-tuning demands are heavy in both computational resources and time. PEFT offers a solution by updating only a small subset of parameters while keeping the base model mostly frozen.

- **Full Fine-Tuning** involves updating all parameters and typically gives the best performance but at a high cost.
- **LoRA (Low-Rank Adaptation)** inserts trainable low-rank matrices into attention layers, achieving good performance with fewer parameters.
- **QLoRA** combines 4-bit quantization of the base model with LoRA for extremely low memory usage while maintaining good performance.
- **IA3 Adapter Tuning** modifies internal layers with additive adapters and freezes the rest, offering efficiency and simplicity.

3. Experimental Setup

Dataset: IMDb movie reviews with binary sentiment labels. A subset was used for quicker training:

- Training samples: 3000
- Testing samples: 2000

Hardware:

- GPU: NVIDIA A100 40GB (I bought Colab Pro)

- Memory Used (peak): 9–12 GB, depending on method
- Frameworks: PyTorch, HuggingFace Transformers & Datasets, PEFT

Hyperparameters (common unless noted):

- Epochs: 3 (LoRA), 5 (others)
- Batch size: 16
- Max length: 512
- Learning rate: 2e-5
- Weight decay: 0.01

4. Results and Visualizations

Method	Accuracy	Train Time (s)	Trainable Params	GPU Memory (MB)
Full Fine-Tuning	0.933	1755.27	124.6M	11559.4
LoRA	0.904	841.87	1.48M	9330.8
QLoRA *(to be added)*	0.89	252.05	296450	4678
IA3 Adapter Tuning	0.8225	1464.94	1.30M	12184.5

5. Analysis and Discussion

- **Full Fine-Tuning** is the gold standard in performance, but the least efficient in terms of memory and time. It trains over 124M parameters, making it unsuitable for low-resource environments.
- **LoRA** provided an impressive trade-off, achieving 90.4% accuracy with only ~1.5M trainable parameters and cutting training time by half.
- **QLoRA** (expected outcome): With quantization, we see a slight drop in accuracy (~89–90%) but a drastic reduction in memory usage (~4GB).
- **IA3** underperformed slightly compared to LoRA, possibly due to less expressive power in the adapter layers, but still viable in low-budget settings.

Use-Case Recommendations:

- **High accuracy, high resource**: Full Fine-Tuning.
- **Balanced trade-off**: LoRA.
- **Very low resource used**: QLoRA.
- **Simple plug-and-play minimal tuning**: IA3.

6. Conclusion and Recommendation

PEFT techniques, especially LoRA and QLoRA, provide compelling alternatives to full model fine-tuning. They are particularly suited for resource-constrained settings without a major loss in accuracy. For future work, incorporating QLoRA results and exploring multi-lingual or domain-specific tasks would extend the value of these methods.