Assignment 3 Report: Fine-Tuning Pretrained Transformers

A0236492Y Tan Yun Xiu

1. Introduction

This report explores the fine-tuning of pretrained Transformer models for a downstream NLP task. The goal is to evaluate two fine-tuning strategies:

(1) Full Fine-Tuning, where all model parameters are updated, and

(2) (2) LoRA Fine-tuning, a parameter-efficient method.

The DistilBERT model was trained on the IMDb sentiment classification dataset, and the results are analyzed in terms of accuracy, F1-score, and model efficiency.

2. Dataset

The dataset I decided to use is the IMDb dataset contains 50,000 labeled movie reviews (25,000 for training and 25,000 for testing). Each review is categorized as positive or negative. It is balanced, clean, and commonly used in NLP tasks, making it suitable for benchmarking fine-tuning methods on text classification.

3. Model and Methods

The base model is DistilBERT (distilbert-base-uncased), a 6-layer Transformer model with approximately 67 million parameters. DistilBERT retains about 95% of BERT's performance while being more lightweight. Two fine-tuning strategies were implemented:

1. Full Fine-Tuning: All model parameters are updated during training.
2. LoRA Fine-Tuning: A parameter-efficient approach that trains low-rank adapter matrices on attention layers (q_lin, k_lin, v_lin). LoRA configuration: rank r = 8, alpha = 32, dropout = 0.1.

Training was performed for 3 epochs using a batch size of 16 and a learning rate of 2e-5 on a Kaggle T4 GPU. Evaluation metrics include accuracy and F1-score. Random seed 42 was used for reproducibility.

4. Results

Strategy	Accuracy	F1-score	Trainable Params (M)	Reduction (%)
Full Fine-Tuning	0.9307	0.9312	66.96	—

LoRA Fine-Tuning	0.9009	0.9015	0.81	98.8

The results show that LoRA achieved nearly equivalent performance to full fine-tuning while reducing trainable parameters by approximately 99%. This results in faster training and lower GPU memory usage.

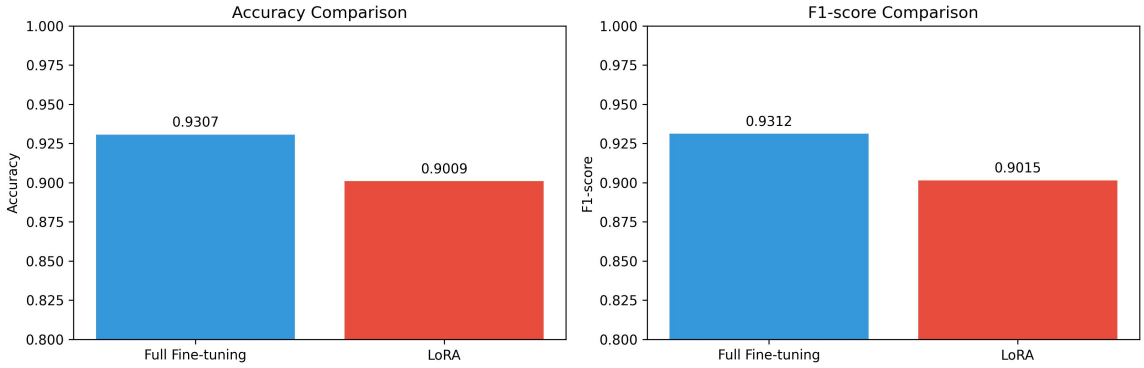

Figure 1. Accuracy and F1-score comparison between full and LoRA fine-tuning.

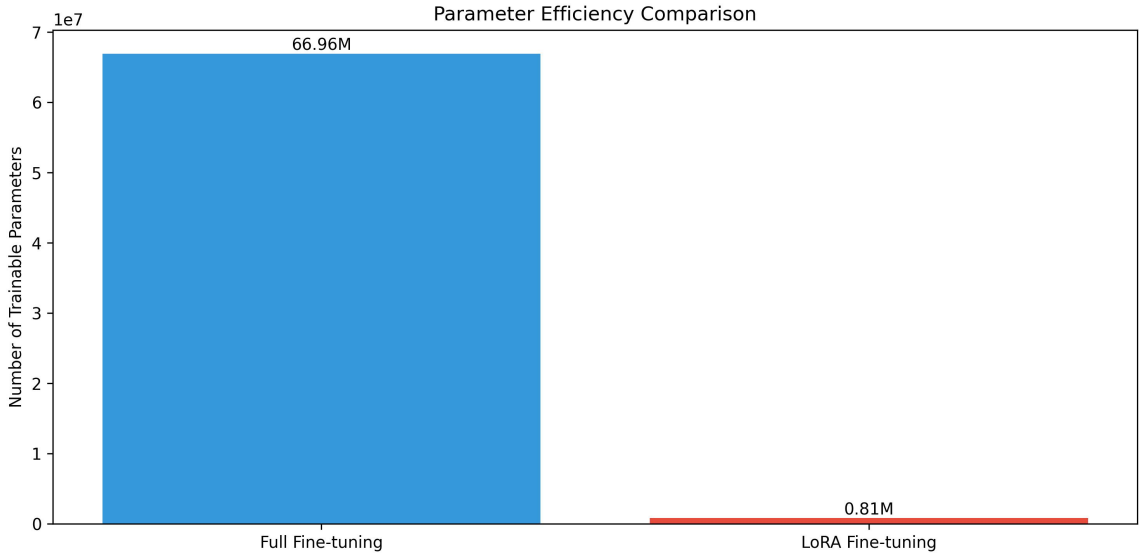

Figure 2. Comparison of trainable parameters between full fine-tuning and LoRA fine-tuning.

5. Discussion

Both fine-tuning methods performed effectively for IMDb sentiment classification. Full fine-tuning achieved slightly better performance (Accuracy = 0.9307, F1 = 0.9312) but required updating all 67 million parameters. LoRA fine-tuning required updating only 0.8 million parameters (1.2% of total) while maintaining Accuracy = 0.9009 and F1 = 0.9015. This

highlights LoRA's efficiency in achieving strong results with a fraction of the computational cost.

The trade-off between performance and efficiency makes LoRA an ideal method for scenarios with limited resources. Although it achieved slightly lower accuracy, the parameter savings make it much more practical for model deployment.

5. Key Takeaways and Limitations

Takeaway

First, the DistilBERT model demonstrated strong performance on the IMDb text classification task, achieving an accuracy of 0.93 and an F1-score of 0.93 through full fine-tuning. Meanwhile, the LoRA fine-tuning strategy achieved comparable results (accuracy 0.90, F1-score 0.90) while reducing trainable parameters by approximately 98.8%.

This substantial reduction illustrates the effectiveness of parameter-efficient approaches such as LoRA in maintaining high performance with significantly lower computational cost. Both methods produced stable results across epochs, confirming the reproducibility of the experimental setup.

Overall, the findings suggest that LoRA offers a practical and efficient alternative to full fine-tuning, particularly in environments with limited resources.

Limitations

The experiment was restricted to the IMDb dataset, which represents a general benchmark rather than a domain-specific corpus. Future research could extend this work to specialized domains such as biomedical or financial texts to assess the generalizability of LoRA.

Additionally, only one Transformer architecture (DistilBERT) was explored; larger models such as RoBERTa or DeBERTa could yield different performance–efficiency trade-offs. The evaluation focused primarily on accuracy and F1-score, which, while informative, could be complemented by precision, recall, and confusion matrix analyses for a more detailed assessment.

Furthermore, the LoRA configuration (rank = 8, $\alpha = 32$) was fixed throughout, may tune these hyperparameters or applying LoRA to different model layers may enhance results.

6. Conclusion

Both fine-tuning methods achieved strong sentiment classification results. LoRA provided significant parameter savings (98.8%) with only a small performance drop (about 3% accuracy difference). Full fine-tuning remains slightly superior but is computationally expensive. LoRA offers a compelling balance between efficiency and performance for Transformer fine-tuning.

7. References

1. Hugging Face Transformers Documentation (2024)
2. PEFT: Parameter-Efficient Fine-Tuning Library (2024)
3. Maas et al. (2011). Learning Word Vectors for Sentiment Analysis. ACL.
4. Kaggle IMDb Sentiment Analysis Dataset.

Appendix - Code Repository

The source code, fine-tuning scripts, and result files are available at:

https://github.com/xiuxiuface/dsa4213-assignment3.git