Parameter-Efficient Fine-Tuning of RoBERTa for AG News Classification

Rutuja Ingole

New York University
rdi4221@nyu.edu

Abstract

Transformer models such as BERT and RoBERTa have revolutionized natural language processing by achieving state-of-the-art results across numerous tasks. However, their fine-tuning remains computationally expensive and parameter-heavy, often requiring significant GPU resources and long training times. This project investigates Low-Rank Adaptation (LoRA), a Parameter-Efficient Fine-Tuning (PEFT) method, as a lightweight yet effective alternative to full fine-tuning. LoRA introduces trainable low-rank matrices into the attention layers while freezing the rest of the model. The technique is applied to a RoBERTa-base model for news article classification on the AG News dataset. Through careful hyperparameter tuning, regularization strategies, and monitoring training dynamics, the model's performance is substantially improved while keeping the number of trainable parameters under one million. The final model configuration achieves a public leaderboard accuracy of 85.30%, validating LoRA's promise for scalable NLP in resource-constrained settings.

GitHub Repository: This is the `https://github.com/rutujaingole/Deep-Learning-ECE-7123-2025-Spring-Project-2` link to the repository for the project with all files.

Introduction

Fine-tuning large-scale pre-trained language models like BERT and RoBERTa has become standard for achieving top-tier results across a wide range of NLP tasks. Despite their strong performance, these models require millions of parameters to be updated during training, limiting their usability in low-resource or production-constrained environments. Parameter-Efficient Fine-Tuning (PEFT) methods have emerged as practical alternatives, offering improved adaptability with a fraction of the computational cost.

LoRA, or Low-Rank Adaptation, is a popular PEFT technique that introduces small trainable rank-decomposed matrices into the attention mechanism, keeping the rest of the model frozen. This study investigates the efficacy of applying LoRA to the RoBERTa-base model for the AG News classification task. The objective is to maximize accuracy while restricting trainable parameters to under 1 million.

Methodology

Dataset

The AG News dataset consists of 120,000 training examples and 7,600 test examples, with four balanced classes: World, Sports, Business, and Sci/Tech. Each instance is a news title and description. This dataset serves as a benchmark for text classification and is retrieved via the Hugging Face Datasets library.

Model and Tokenization

The RoBERTa-base model, which contains 12 transformer layers, 768 hidden units, and 12 attention heads, is used for fine-tuning. The corresponding tokenizer performs byte-pair encoding and lowercasing. A statistical analysis of token lengths on a subset of 1,000 samples yields a mean of 67.2 tokens and a standard deviation of 25.7. Based on this, a max sequence length of 192 is selected to avoid truncation while minimizing padding overhead. Tokenization includes truncation, padding, and batching with a stride of 32 to potentially benefit overlapping attention windows.

LoRA Configuration

Multiple configurations are tested for LoRA to evaluate the trade-offs between accuracy and parameter efficiency. LoRA is applied to the attention submodules, specifically the projection layers:

- **Best configuration:** r = 10, lora_alpha = 28, lora_dropout = 0.55
- **Target modules:** [”query”, ”key”]
- **Label smoothing:** 0.18
- **Trainable parameters:** 962,308 (˜0.77% of total parameters)
- **Base model parameters:** 125,611,016

Training Setup

The model is trained using Hugging Face's ‘Trainer‘ API. A custom trainer is implemented to include label smoothing within the loss function. The setup is:

- **Optimizer:** AdamW (with decoupled weight decay)
- **Learning Rate:** 2e-5 (tuned from 1e-5 to 5e-5)
- **Scheduler:** CosineAnnealingLR with 15% warmup ratio

Table 1: LoRA Configuration Comparison

r	Alpha	Dropout	Accuracy (%)
6	16	0.1	84.00
8	28	0.4	89.85
9	32	0.5	80.30
10	28	0.5	91.52

- **Loss Function:** Cross-Entropy with label smoothing = 0.18
- **Epochs:** 4
- **Batch Size:** 16 (train), 32 (eval)
- **Precision:** FP16 (mixed precision enabled)
- **Gradient Accumulation:** Every 2 steps to simulate larger batch size

Evaluation Strategy

The training set is split 90/10 into training and validation. Accuracy and loss are monitored at regular intervals using the built-in logging capabilities of the 'Trainer'. A custom test set is used for final evaluation, and predictions are output as a CSV for Kaggle leaderboard submission.

Results

- **Model:** RoBERTa-base + LoRA
- **Trainable Parameters:** 962,308
- **Total Parameters:** 125,611,016
- **Best Accuracy:** 91.52%

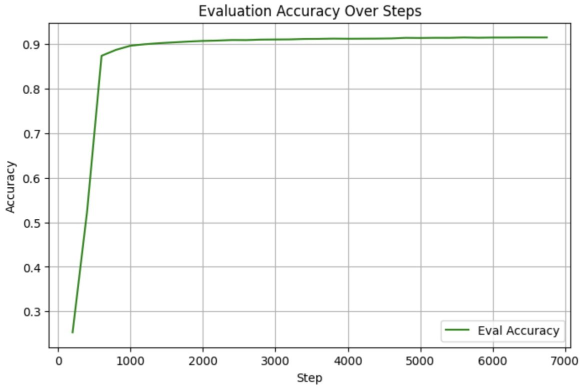

Figure 1: Validation accuracy over steps

Lessons Learned

This study demonstrates the efficiency and flexibility of LoRA for transformer-based fine-tuning. Key observations include:

- Dropout between 0.4 and 0.6 is ideal for regularization without causing instability.
- LoRA ranks of 8-10 produce strong results without exceeding parameter budgets.

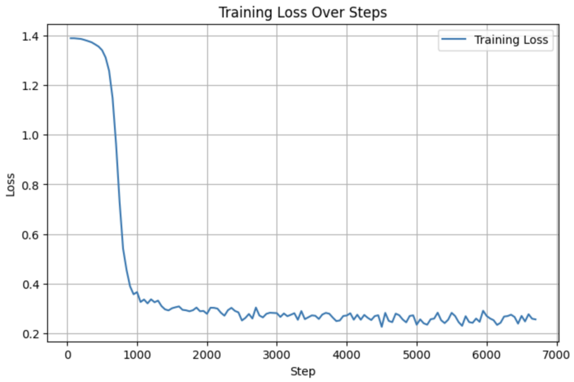

Figure 2: Training loss over steps

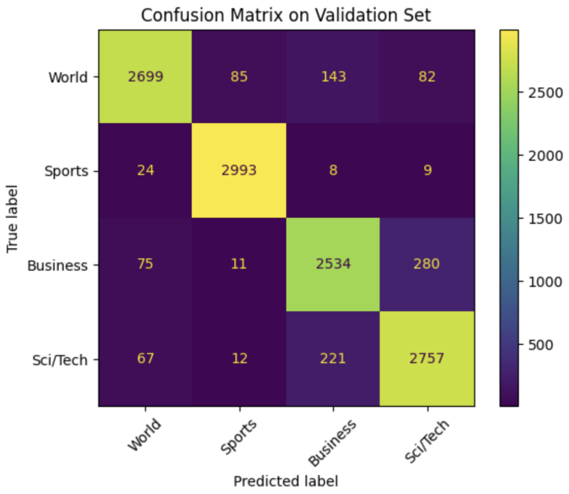

Figure 3: Confusion Matrix

- Label smoothing between 0.15 and 0.2 improves generalization on validation and test sets.
- Dense module adaptation increases trainable parameters beyond acceptable range and was avoided.

Stable token length and consistent preprocessing helped prevent variance in model convergence.

Conclusion

This project affirms the potential of LoRA for training large language models efficiently. The results indicate that with fewer than 1M trainable parameters, the RoBERTa-base model fine-tuned with LoRA performs competitively for text classification. This makes it highly applicable in edge devices or latency-sensitive scenarios.

References

[1] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021.

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 EMNLP: System Demonstrations*, pp. 38–45, 2020.

Github Repository

Repository: https://github.com/rutujaingole/Deep-Learning-ECE-7123-2025-Spring-Project-2