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

Nishant Sharma[1], **Rahul Mallidi**[2], **Anushka Garg**[3]

NYU Tandon School of Engineering
ns6287@nyu.edu, rm7020@nyu.edu, ag9012@nyu.edu

Abstract

In this study, a refined RoBERTa model optimized using Low-Rank Adaptation (LoRA) is used to create an effective text classification model on the AG News dataset. Maximizing inference accuracy while adhering to a limitation of one million trainable parameters was the aim. We conducted knowledge distillation with both fine-tuned and non-fine-tuned instructor models, assessed many LoRA configurations, and specifically targeted particular layers in RoBERTa for adaptation. Our findings demonstrate the performance improvements of PEFT techniques, especially LoRA, with our top model attaining an inference accuracy of 88.575%.

Supporting Material

- **Code Repository:** Please refer to the attached link here for relevant codebase to the project.
- **Trained Model:** Please refer to the url attached here for our trained model submitted for the project.

Introduction

Text classification is a fundamental task in natural language processing (NLP) that involves assigning predefined categories to textual data such as news articles, reviews, or documents. Applications such as topic labeling, sentiment analysis, and spam detection depend heavily on it. Transformer-based models are now the most popular architecture for resolving a variety of NLP tasks, including text categorization, thanks to the development of deep learning.

A transformer-based language model based on BERT (2), RoBERTa (A Robustly Optimized BERT Pretraining Approach) (1) is one of these models. RoBERTa enhances BERT's pretraining by incorporating a larger training corpus, removing the Next Sentence Prediction objective, and using dynamic masking, which improves performance across several benchmarks. RoBERTa has consistently outperformed BERT in various NLP benchmarks and has become a standard baseline for many downstream tasks.

However, it takes a lot of resources to fully fine-tune models like RoBERTa. It entails changing every model parameter, which for base models may number more than 100 million. This raises the cost of computation and complicates deployment in systems with limited memory. Consequently, Parameter-Efficient Fine-Tuning (PEFT) has become a viable and affordable option. PEFT methods aim to adapt large models efficiently by tuning only a subset of parameters. LoRA achieves this by injecting compact, learnable updates into frozen layers, reducing computational demands while maintaining performance.

Low-Rank Adaptation (LoRA) (3) is one such PEFT technique. LoRA adds trainable low-rank matrices to the weight update process while freezing the initial pretrained weights. In order to make $\Delta W \approx AB$, LoRA breaks down the entire gradient updates ΔW into two low-rank matrices, $A \in R^{d \times r}$ and $B \in R^{r \times k}$, where r is substantially lower than d or k. This enables effective adaption of big models and drastically lowers the number of trainable parameters.

In this study, we use the AG News dataset to refine a RoBERTa-based model for a text classification job using LoRA. Achieving excellent classification accuracy while keeping the total number of trainable parameters under one million is our goal. We also experiment with knowledge distillation from a teacher model and selectively apply LoRA to the most influential layers in the RoBERTa architecture in order to further enhance model performance under this restriction. Our results show that LoRA allows for extremely effective and efficient fine-tuning for practical applications such as news classification.

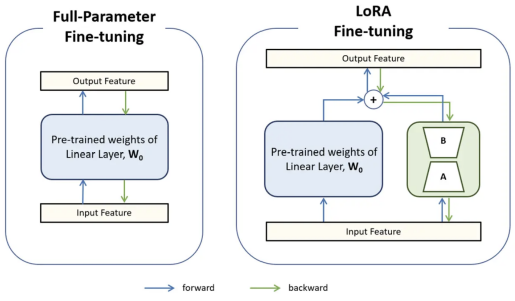

Figure 1: Comparison between full-parameter fine-tuning and LoRA-based fine-tuning, highlighting how LoRA adds low-rank updates (A and B) without modifying the pre-trained weights (W_0).

Existing Work

By enabling models to simultaneously attend to every segment of the input sequence through self-attention mechanisms, the Transformer architecture, first presented by Vaswani et al. in 2017 (4), transformed natural language processing. Since then, a number of variations have shown state-of-the-art performance on a variety of tasks, including BERT, RoBERTa, GPT, and others. However, issues with model fine-tuning, storage, and inference efficiency have arisen due to their increasing size.

To address these issues, several Parameter-Efficient Fine-Tuning methods have been proposed. These include:

- **Adapter Tuning** (5): Inserts small bottleneck layers between existing transformer layers and updates only the adapters during fine-tuning, leaving the rest of the model frozen.
- **Prefix Tuning** (6): Prepends a set of trainable continuous vectors (prefixes) to the input of each transformer layer, enabling the model to adapt to new tasks while keeping the model weights fixed.
- **Prompt Tuning**: Learns task-specific prompts in the embedding space to guide the pretrained model without altering its internal architecture or parameters.

LoRA has become especially well-liked among these due to its ease of use and efficiency. LoRA, which was first presented by Hu et al. (2021), supplements the weight matrices in feedforward or attention layers with trainable rank decomposition matrices. With only a few new parameters to learn during training, it preserves the effectiveness of full model inference.

A recent study by Xu et al. (2023) offers a thorough analysis of PEFT techniques and compares how well they perform on various tasks, such as question answering, categorization, and summarization (7). They discover that the optimal compromise between performance and efficiency is achieved by LoRA and adapter tuning.

Other studies have looked into how LoRA can be used with distillation and quantization (such as QLoRA) to further reduce the memory footprint of models without significantly affecting performance. By applying LoRA to the AG News dataset and further enhancing performance through targeted module modification and knowledge distillation approaches, our effort expands on these findings.

Our goal is to demonstrate through these studies that state-of-the-art performance for classification tasks may be attained without the need for computationally costly full fine-tuning, provided that careful hyperparameter tuning and architecture choices are made.

Dataset

The AG News dataset is a well-known benchmark used for evaluating text classification models. It consists of 120,000 training samples and 7,600 test samples, evenly distributed across four classes:

- World
- Sports
- Business
- Sci/Tech

Each data point includes a short news title and a corresponding description. These short texts require the model to effectively understand both context and topic from limited input. The dataset is balanced and diverse, making it suitable for evaluating general-purpose classifiers under parameter constraints.

Methodology

Our methodology was based on three core experiments:

- **LoRA Tuning:** We varied LoRA's rank (r) and scaling factor (α) across attention and feedforward layers in RoBERTa while monitoring trainable parameter count, testing ranks between 4 and 12, and alpha values between 16 and 32.
- **Knowledge Distillation:** Both fine-tuned and frozen teacher models were used to teach the student model using soft labels. The optimal distillation setup used a fine-tuned teacher model with a rank of 8 and alpha of 32, paired with a student model configured at rank 16 and alpha 8.
- **Targeted Adaptation:** Instead of applying LoRA uniformly, we targeted specific layers such as query/key projections in the self-attention modules and select dense layers (e.g., layers 0, 1, 5, 10, and 11 in RoBERTa's encoder) to enhance model efficiency and accuracy.

We used HuggingFace's 'transformers' and 'peft' libraries, and optimized with AdamW using cosine learning rate scheduling.

The table below presents the epoch-wise training and validation metrics for our best-performing model configuration, identified by achieving the highest validation accuracy. These metrics illustrate the model's learning progression, convergence behavior, and consistency across epochs, highlighting the point of optimal performance.

Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
0.2750	0.2975	0.9016	0.9014	0.9020	0.9016
0.2581	0.2380	0.9234	0.9235	0.9237	0.9234
0.2159	0.2293	**0.9297**	**0.9296**	**0.9297**	**0.9297**
0.2177	0.2252	0.9281	0.9281	0.9282	0.9281
0.2201	0.2247	0.9281	0.9281	0.9282	0.9281

Table 1: Epoch-wise training and validation metrics for our model with highest Validation Accuracy

We created a custom wrapper for RoBERTa that applies a 20% dropout layer to the final pooled output before to classification in order to further minimize overfitting. Because it enables the model to acquire more robust representations without expanding the number of trainable parameters, this targeted regularization is particularly helpful when fine-tuning using LoRA.

We chose an alpha of 32 and a rank of 12 for the LoRA configuration, with a dropout of 0.05 within the adapter modules. Early layers (0–1) for base understanding, a mid layer (5) for intermediate representation, and end layers

(10–11) for task-specific feature adaptation were the important locations for adapters in the self-attention layers. Our method's parameter effectiveness was demonstrated by the fact that just roughly 796K parameters, or 0.63% of the entire model, were trained.

With the AdamW optimizer and cosine learning rate scheduling, training was carried out over five epochs at a learning rate of 3e-5. To stabilize training, other methods were used, including gradient accumulation (steps=2), learning rate warmup (15%), and weight decay (0.1). Each epoch involved evaluation and model checkpointing, from which the optimal model was chosen based on validation accuracy.

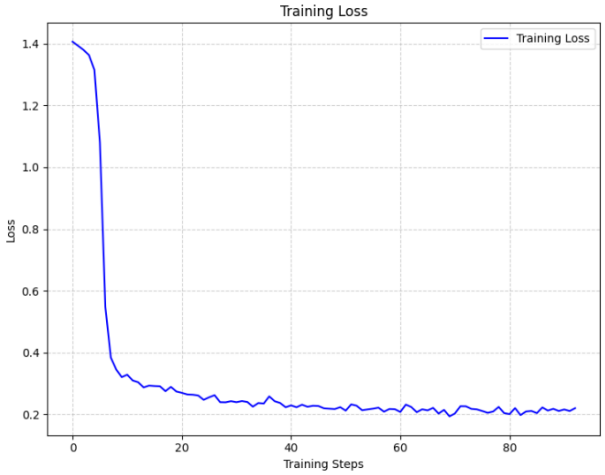

Figure 2: Training progression of the model showing loss reduction over epochs.

Results

Our best result of 85.575% inference accuracy was achieved with tuned hyperparameters and knowledge distillation. We observed that selectively applying LoRA led to higher accuracy at lower parameter cost.

Configuration	Params	Val. Acc.	Inf. Acc.
LoRA only	888,580	0.8437	0.8207
LoRA + Distillation	925,444	0.8890	0.8300
LoRA + Distillation (Reduced)	814,852	0.8843	0.8430
LoRA + Distillation (Tuned)	888,580	0.8922	0.8490
Targeted LoRA Modules	796,420	**0.9297**	**0.8453**

Table 2: Comparison of model variants under 1M parameter constraint.

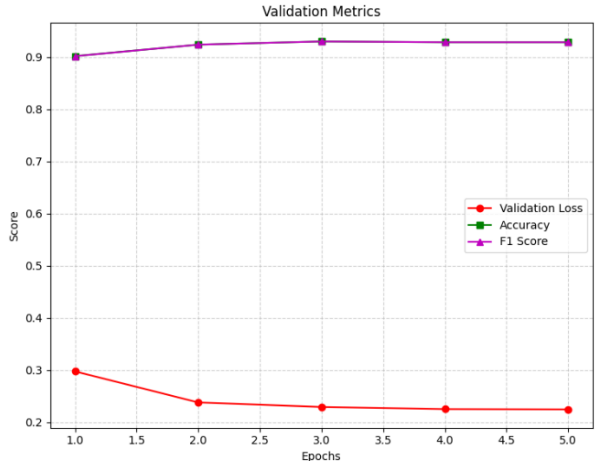

Figure 3: Validation metrics of the model on the training set.

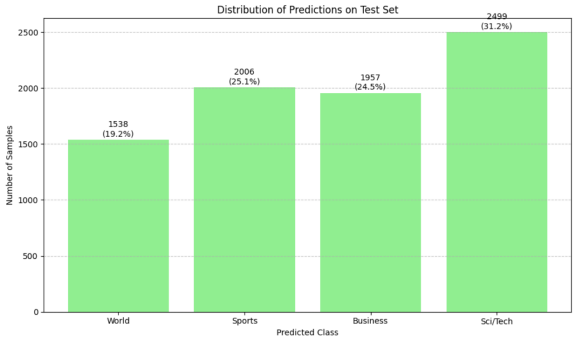

Figure 4: Predictions of the model on the test set.

Discussion

The tests demonstrated how important it is to strategically allocate parameters while adjusting RoBERTa models using LoRA. Selectively focusing on particular modules, including query/key attention layers and dense output layers, produced better accuracy with fewer parameters, even if applying LoRA widely across all layers enhanced performance. By acting as a regularization process, knowledge distillation further stabilized and enhanced performance, offering extra advantages. Accurately adjusting hyperparameters such as scaling factor (α) and rank (r) was crucial, showing quantifiable performance improvements, especially when parameter budgets were tight. To further maximize parameter efficiency and model performance, future research may look into a more thorough integration of targeted adaptation and improved distillation techniques.

Conclusion

According to this work, LoRA can successfully adjust RoBERTa for AG News classification while adhering to a 1M parameter cap. Our model uses knowledge distillation, educated hyperparameter selection, and modular LoRA tweaking to achieve high accuracy. The necessity for careful review across measures is highlighted by the crucial lesson

we learned: better validation accuracy does not always imply the optimal model for inference. More precise control over adapter placement, cross-layer attention sharing, and interaction with quantization methods like QLoRA will all be investigated in future research.

References

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

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT.

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

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is All You Need*. Advances in Neural Information Processing Systems.

[5] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gelly, S., Wang, W., & Metcalfe, P. (2019). *Parameter-efficient Transfer Learning for NLP*. Proceedings of the 36th International Conference on Machine Learning.

[6] Li, X., & Liang, P. (2021). *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.

[7] Xu, L., Xie, H., Qin, S., Tao, X., & Wang, F. (2023). *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. arXiv preprint arXiv:2312.12148.