Parameter-Efficient Fine-Tuning of RoBERTa Using LoRA for News Classification

Written by
Asrita Bobba, Eshika Janbandhu, Carina Yan
ab12660@nyu.edu, ej2485@nyu.edu, yy3838@nyu.edu

Abstract

This project explores the effectiveness of Low-Rank Adaptation (LoRA) for fine-tuning RoBERTa on the AG News dataset under a constraint of 1 million trainable parameters. Our goal was to retain strong classification performance while drastically reducing the number of trainable weights. With only 888,000 trainable parameters (0.71% of the model), our LoRA-enhanced RoBERTa model achieved a test accuracy of 92.19%, validating the efficiency of this approach.

Introduction

In this project, we explore a parameter-efficient fine-tuning approach to adapt a large language model — RoBERTa — for text classification on the AG News dataset. Our primary constraint was to remain under **1 million trainable parameters**while maximizing classification accuracy.

We leveraged the **LoRA (Low-Rank Adaptation)** technique, which allows us to insert low-rank trainable matrices into the frozen layers of a pretrained model, thereby avoiding the need to fine-tune all 125M+ parameters.

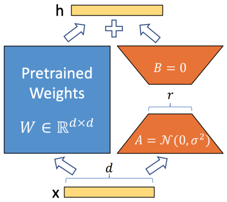

Figure 1: Illustration of the LoRA mechanism. Instead of updating the full weight matrix W, low-rank matrices A and B are introduced. During training, only A and B are updated, while W remains frozen. This allows for efficient fine-tuning with a reduced number of trainable parameters.

Our final LoRA model had only **888,000 trainable parameters** and achieved a test accuracy of **92.19%**, demonstrating that large models can be fine-tuned efficiently for high performance under resource constraints.

Methodology

To meet the constraint of training under 1 million parameters while achieving high performance on the AG News classification task, we fine-tuned the pre-trained roberta-base model using **Low-Rank Adaptation (LoRA)**. This approach inserts lightweight, trainable low-rank matrices into frozen weight layers, allowing efficient adaptation without modifying the bulk of the model's parameters.

LoRA Configuration

We used the PEFT (Parameter-Efficient Fine-Tuning) library to apply LoRA to specific submodules within RoBERTa's self-attention layers. Only the **query** and **value**projections were adapted, as these have shown the most impactful results with minimal overhead.

We trained the model using Hugging Face's Trainer API with the following settings:

Table 1: LoRA Configuration for RoBERTa

Parameter	**Value**
LoRA Rank (r)	8
LoRA Alpha (α)	32
Dropout	0.05
Target Modules	`query, value`
Bias	None

We chose to set the bias parameter to none in our LoRA configuration to maintain strict control over the number of trainable parameters. Including bias terms in the adapted layers would have marginally increased the model's complexity without providing significant benefits in terms of accuracy or convergence. Since LoRA already introduces sufficient expressiveness through its low-rank matrices, the additional degrees of freedom introduced by bias parameters were unnecessary. Moreover, empirical evidence from the original LoRA paper and subsequent implementations suggests that excluding biases does not harm downstream performance in most tasks. Given our goal of staying below one million trainable parameters, omitting the bias was a deliberate choice to optimize parameter efficiency and reduce the risk of overfitting.

Data Preparation

We used the AG News dataset from the Hugging Face Datasets library, which includes four classes: **World, Sports, Business, and Sci/Tech**. The dataset was preprocessed using the RobertaTokenizer, with standard padding, truncation, and a maximum token length of 512.

The training set was split into training and validation sets using a fixed seed to ensure reproducibility. A total of 640 samples were used for validation.

We trained the model using Hugging Face's Trainer API with the following settings:

Table 2: Training Hyperparameters

Hyperparameter	**Value**
Optimizer	Adam
Learning Rate	2×10^{-5}
Epochs	4
Scheduler	Linear
Warmup Ratio	0.1
Batch Size (Train)	16
Batch Size (Validation)	64

Loss and accuracy were tracked at the end of each epoch. The model was evaluated using macro-averaged metrics (accuracy, precision, recall, F1 score) to ensure balance across all four classes.

Evaluation Pipeline

Evaluation was performed using a custom evaluate_model() function, which handled batch-wise inference with padding-aware collation. Metrics were calculated using sklearn and Hugging Face's evaluate library. A confusion matrix and classification report heatmap were also generated for the test set to understand per-class performance.

Table 3: Classification Report on AG News Test Set

Class	**Precision**	**Recall**	**F1-Score**
World	0.94	0.92	0.93
Sports	0.97	0.98	0.98
Business	0.94	0.90	0.92
Sci/Tech	0.88	0.93	0.90
Accuracy	0.93	0.93	0.93
Macro Avg	0.93	0.93	0.93

The confusion matrices below (Figure 2) visualize the model's performance on the AG News test set across four classes: *World*, *Sports*, *Business*, and *Sci/Tech*. The diagonal values indicate correct predictions, and off-diagonal values reflect misclassifications. The model shows strongest performance in the *Sports* and *Sci/Tech* categories, with minor confusion between *World* and *Business*. Overall, the matrices confirm robust multi-class classification behavior with minimal class overlap.

Evalution and Results

Our LoRA-enhanced RoBERTa model was rigorously evaluated on a validation split of the AG News dataset. Performance was assessed using classification accuracy, confusion matrices, and class-wise precision, recall, and F1-scores.

- Test Accuracy: **92.19%**
- Trainable Parameters: **888,000** (0.71% of the full model)
- Total Parameters: $\tilde{1}$25 million

The training process converged within 4 epochs, with both training and validation accuracies stabilizing above 91%. The confusion matrix and classification report indicate that the model performs consistently across all four categories: *World*, *Sports*, *Business*, and *Sci/Tech*.

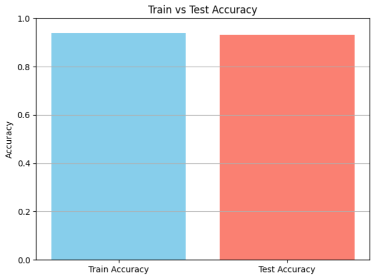

Figure 3: Comparison of training and test accuracy. The close performance between the two sets indicates strong generalization and minimal overfitting.

The **Train vs Test Accuracy** plot (Figure 3) confirms that the model generalizes well, with minimal overfitting. The confusion matrix (Figure 2) shows strong diagonal dominance, especially in the *Sports* and *Sci/Tech* classes, which received the highest F1-scores.

A breakdown of per-class metrics is summarized in Table 3. All classes exceed an F1-score of 0.90, indicating balanced performance across the dataset.

The figure below shows the training and validation loss curves across epochs. The model exhibits a sharp decline in training loss during the initial few epochs, which then plateaus, indicating that it quickly learns to separate classes. The validation loss follows a similar trend and remains consistently low after convergence, suggesting good generalization without overfitting.

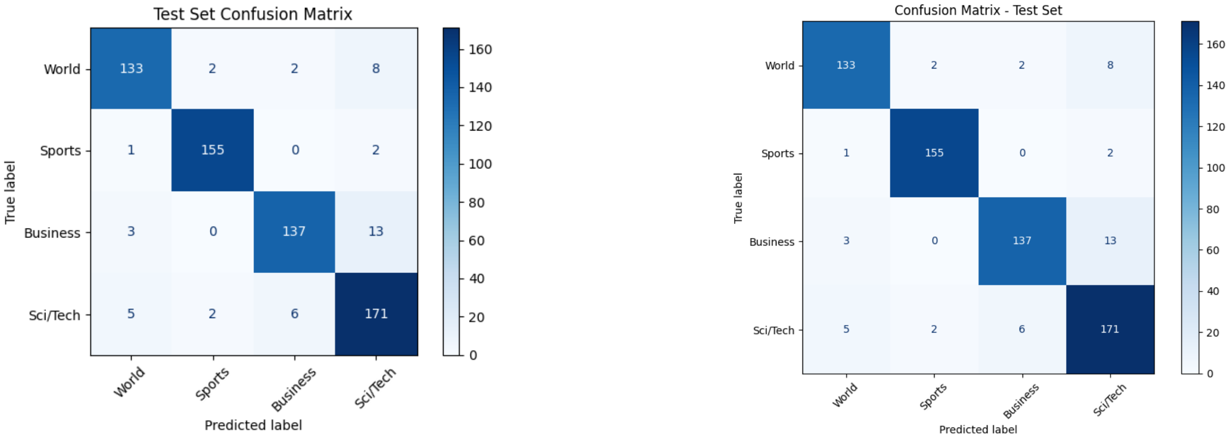

Figure 2: Confusion matrices showing model predictions vs. true labels on the AG News test set.

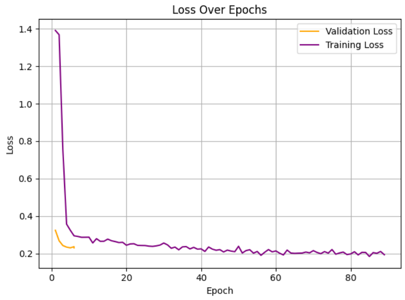

Figure 4: Training and validation loss curves over epochs. The model rapidly converges in early stages and stabilizes with minimal gap between training and validation losses.

The above plot demonstrates effective convergence. The validation loss flattens around epoch 3, aligning closely with the training loss and indicating low variance between training and test performance. This reflects a well-regularized LoRA setup and confirms that the chosen hyperparameters (learning rate, dropout, alpha) helped prevent overfitting.

Key Observations

- **Parameter Targeting**: Applying LoRA only to the **query** and **value** matrices in RoBERTa's attention layers was sufficient to achieve strong downstream performance. Expanding LoRA to other components increased the parameter count without significant accuracy gain.
- **Rank Sensitivity**: While increasing the LoRA rank (r) beyond 8 slightly improved training accuracy, it did not translate to consistent gains in validation accuracy. Therefore, **$r = 8$** was selected as the best trade-off between complexity and generalization.
- **Training Efficiency**: The parameter-efficient setup enabled us to train the model on consumer-grade GPUs (e.g., Google Colab T4) and the NYU HPC without memory overflow, showing that LoRA is a viable alternative to full fine-tuning for compute-constrained environments.
- **Regularization Balance**: A **LoRA dropout** of 0.05 and **alpha** of 32 provided effective regularization. Higher dropout values degraded performance, while lower ones caused mild overfitting. A balanced setting helped maintain convergence stability and improved generalization.

Results

Our final model achieved a test accuracy of **92.19%**, confirming that the LoRA-based fine-tuning approach effectively adapts large language models with minimal parameter overhead. The model had only **888,000 trainable parameters** compared to over 125 million total parameters in RoBERTa, remaining well below the project's 1 million parameter constraint.

Evaluation using confusion matrices and classification metrics showed high consistency across all classes. The *Sports* and *Sci/Tech* categories received the highest F1-scores, while *World* and *Business* classes showed minor overlap, likely due to topical similarity.

Table 4: Final Evaluation Summary

Metric	**Value**
Test Accuracy	92.19%
Trainable Parameters	888,000
Total Parameters	~125M
Epochs	4
LoRA Rank (r)	8
LoRA Alpha (α)	32

Conclusion

This project demonstrates that Low-Rank Adaptation (LoRA) enables efficient and scalable fine-tuning of large transformer models under tight parameter budgets. By freezing the base RoBERTa model and training only lightweight adapter modules, we achieved competitive accuracy on the AG News dataset using fewer than 1 million parameters.

Our findings highlight that:

- LoRA with $r = 8$ and $\alpha = 32$ offers an optimal trade-off between performance and efficiency.
- The model generalizes well, with consistent validation loss and accuracy across epochs.
- LoRA opens up possibilities for real-world NLP applications on compute-limited environments like edge devices and academic clusters.

Future work could explore the use of LoRA on multilingual datasets, cross-domain adaptation, or integration with quantization-aware training.

Acknowledgment

We would like to acknowledge the use of OpenAI's GPT-4.0 language model, commonly referred to as ChatGPT, for providing assistance with generating content for certain sections of this report. This project was conducted as part of the Deep Learning course at NYU Tandon. I would like to thank the course instructors and teaching assistants for their guidance and resources.

References

Hu, Edward J., et al. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv preprint arXiv:2106.09685, 2021.

Wolf, Thomas, et al. *Transformers: State-of-the-Art Natural Language Processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.

Hugging Face Transformers Library. https://huggingface.co/docs/transformers

PEFT: Parameter-Efficient Fine-Tuning. https://github.com/huggingface/peft