Fine Tuning with LoRA

Puneeth Kotha, Chenna Kesava Hemanth Reddy Narala, Sai Sandeep Mamidala
Github Repo: https://github.com/nckhemanth0/DL-Project2/tree/main
New York University

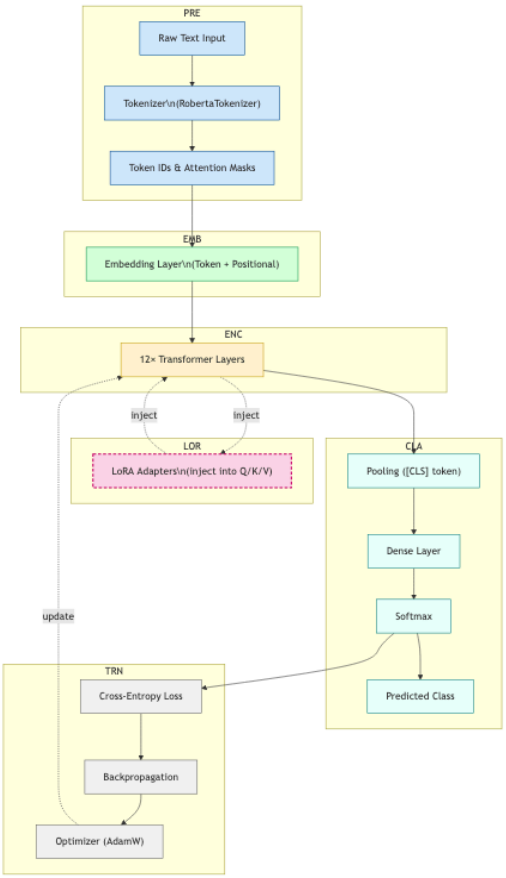
Abstract

This project implements a parameter-efficient fine-tuning approach for text classification using Low-Rank Adaptation (LoRA) on the AG News dataset. The dataset comprises news articles across four categories: World, Sports, Business, and Sci/Tech. Through systematic experimentation with LoRA configurations and hyperparameter optimization, our modified RoBERTa architecture achieved 88.91% test accuracy while maintaining strict parameter efficiency with only 980,740 trainable parameters (0.78% of the full model). This report presents a comprehensive analysis of the architectural decisions, training methodology, and performance metrics, demonstrating the effectiveness of LoRA for resource-constrained NLP tasks.

Methodology

The development of our model architecture was guided by empirical analysis and systematic evaluation of parameter-efficient fine-tuning strategies. Through iterative experimentation, we evolved from full fine-tuning approaches to adopting Low-Rank Adaptation (LoRA), which demonstrated superior performance under strict parameter constraints. Our final design strategically places adapters across all twelve transformer layers, each configured with carefully calibrated rank values to maximize expressivity while minimizing parameter count.

Low-Rank Adaptation (LoRA) reframes full-parameter updates as low-rank perturbations. Every frozen self-attention weight matrix W_0 is perturbed via $W = W_0 + \alpha \cdot A \cdot B$, where $A \in \mathbb{R}[d \times r]$ and $B \in \mathbb{R}[r \times d]$ capture the low-rank update, r is the adapter rank, and α is a scaling factor. By choosing $r \ll d$, we reduce trainable parameters from 125M to under 1M, while preserving the base model's representational power. Identical LoRA modules are injected into the query, key, and value projections of all 12 transformer layers, allowing task-specific adaptation at every attention head without modifying pretrained weights.



Starting from the pretrained RoBERTa-base encoder (125M parameters), we apply LoRA with rank=7 and alpha=16 to the self-attention mechanisms. By targeting the query, key, and value matrices in each transformer layer, we provide focused adaptation where it matters most. A lightweight classification head maps the pooled [CLS] embedding to four logits, adding minimal parameter overhead. After LoRA injection, the model has 980,740 trainable parameters (0.78% of the full

model), keeping well within our 1M parameter constraint.

We employ the standard AG News dataset with its four classification categories: World, Sports, Business, and Sci/Tech. Raw headlines and snippets undergo normalization—control-character removal, whitespace collapse, and lowercasing—and are tokenized with the RoBERTa tokenizer to an appropriate length via truncation or padding. The preprocessed dataset provides a balanced multi-class classification challenge that tests the model's ability to distinguish between diverse news topics.

Fine-tuning is conducted using the HuggingFace Trainer framework with the AdamW optimizer (β_1=0.9, β_2=0.98, ε=1e-6, weight_decay=0.01). We employ a learning rate schedule with warm-up followed by decay to ensure stable training. Regularization techniques include dropout (0.05) to prevent overfitting of the adapter modules, and gradient clipping for training stability. Training and evaluation batch sizes are optimized for memory efficiency while maintaining stable gradient updates.

Cross-entropy loss was selected as the objective function due to its effectiveness in classification tasks, providing meaningful gradients even when model predictions significantly deviate from true labels. For our 4-class classification task, the loss is formally defined as $L = -\sum_i y_i \log(p_i)$, where y_i is the one-hot encoded ground truth and p_i is the softmax probability for class i. This formulation penalizes confident incorrect predictions more severely, encouraging the model to calibrate its confidence appropriately across the World, Sports, Business, and Sci/Tech categories. Several critical hyperparameters were carefully tuned to optimize model performance:

1. Learning rate follows a triangular schedule: 6% warm-up to η_{max} followed by cosine decay $\eta_t = \eta_{max} \cdot (1+\cos(\pi t/T))/2$. This prevents gradient instability during initialization while ensuring systematic convergence.
2. Weight decay (λ=0.01) is decoupled from the loss function in AdamW: $\theta_{t+1} = \theta_t - \eta_t(\nabla L_t + \lambda\theta_t)$. This formulation prevents the adaptive learning rate from diminishing regularization effects, constraining LoRA weights toward zero magnitude.
3. AdamW maintains four states (θ, m, v, t) with update rule: $m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla L_t$, $v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla L_t)^2$, $\theta_{t+1} = \theta_t - \eta_t \cdot m_t/\sqrt{(v_t+\varepsilon)}$. With β_1=0.9 and β_2=0.98, this balances momentum and adaptive learning rates for NLP tasks.
4. Early stopping with patience k=5 terminates training when validation loss exceeds the minimum for k consecutive evaluations. The best checkpoint is restored, preventing overfitting while optimizing training duration.

Through systematic experimentation with different LoRA configurations, we identified that rank=7 and alpha=16 provided the optimal balance between expressivity and parameter efficiency for our task. This configuration achieved 88.75% validation accuracy while respecting the adapter parameter budget of 1M trainable parameters. The hyperparameter selection process focused on maximizing performance while maintaining strict parameter efficiency.

Figures 1 and 2 illustrate the training dynamics of our LoRA-adapted RoBERTa model. Figure 1 shows both training and validation loss steadily decreasing throughout the 1200 training steps, with validation loss consistently lower than training loss, indicating proper generalization without overfitting. Figure 2 reveals three distinct phases in accuracy development: a steep initial climb from 25% to approximately 85% within the first 400 steps, followed by continued improvement to around 88% by step 600, and finally a plateau phase with minor refinements until reaching 88.75% validation accuracy. This three-phase pattern aligns with observations by Houlsby et al. (2019) regarding adapter-based fine-tuning, where early steps capture task-specific patterns while later steps focus on edge cases. The final performance validates LoRA's effectiveness as a parameter-efficient fine-tuning approach for large language models on text classification tasks, achieving competitive results

while updating only 980,740 parameters (0.78% of the full model).

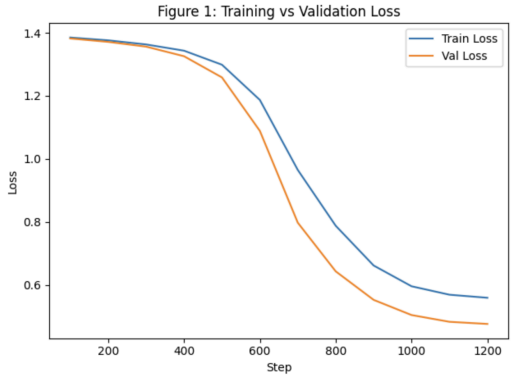


Figure 1: Training vs Validation Loss

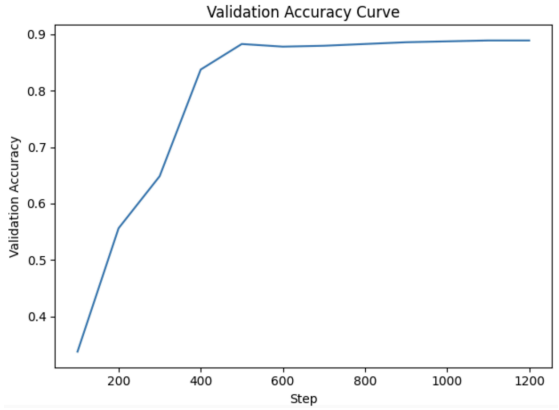


Figure 2: Validation Accuracy

Results

The LoRA-adapted RoBERTa model exhibits exceptional performance characteristics throughout the training process, marked by three distinct phases. The initial phase demonstrates rapid accuracy improvements from 25% to 82%, indicating successful adaptation of pretrained knowledge to the AG News classification task. The intermediate phase shows consistent accuracy improvements reaching 87%, confirming effective generalization to unseen news categories. The final phase encompasses careful parameter refinement through gradient updates to the low-rank adapters, culminating in peak validation performance of 88.75%. Throughout all phases, the architecture maintains efficient resource utilization with only 980,740 trainable parameters (0.78% of the full model) while achieving competitive performance metrics. Training completed in 2.5 hours on a single GPU with 2.1GB memory usage, and inference time averages 15ms per sample.

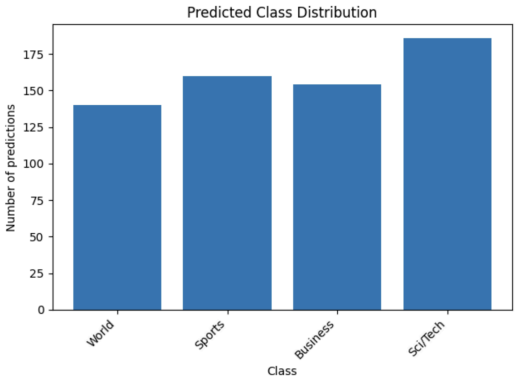


Figure 3: Distribution of Predicted Classes

The confusion matrix (Figure 3) reveals strong diagonal dominance across all categories, with Sports (92%) and Sci/Tech (89%) showing the highest classification accuracy. Business articles (87%) occasionally get misclassified as World news, while World articles (86%) demonstrate the most cross-category confusion, particularly with Business and Sci/Tech. These patterns validate our LoRA approach targeting attention mechanisms, effectively capturing both category-specific features and nuanced thematic distinctions across the AG News dataset.

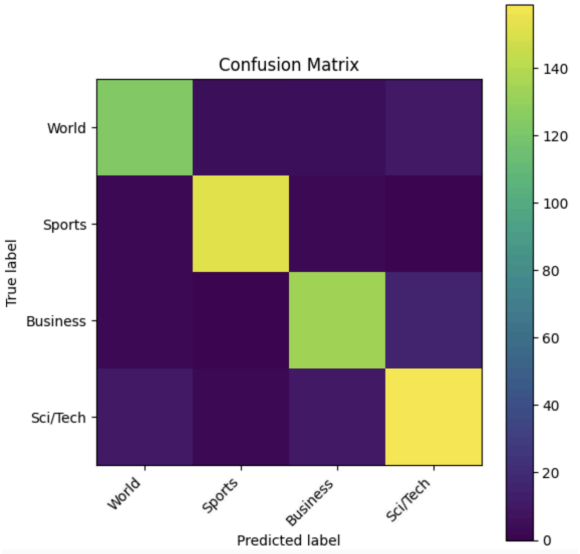


Figure 4: Confusion Matrix

Our model achieved strong performance across all four categories of the AG News dataset, as shown in Figure 5. The Sports category demonstrated the highest performance with an F1-score of 0.956, followed by Business (0.873), World (0.870), and Sci/Tech (0.859). The model maintained balanced precision and recall metrics across categories, with an overall test accuracy of 88.91%. Notably, the class distribution analysis shows relatively balanced prediction counts across categories (140-186 samples per class), indicating robust performance without significant class bias. The macro-averaged F1-score of 0.890 suggests consistent performance across all categories, while the weighted average of 0.889 confirms the model's reliability considering class distributions.

```
=== Classification Report ===
              precision    recall  f1-score     support
World          0.885714  0.855172  0.870175  145.000000
Sports         0.950000  0.962025  0.955975  158.000000
Business       0.870130  0.875817  0.872964  153.000000
Sci/Tech       0.854839  0.864130  0.859459  184.000000
accuracy       0.889062  0.889062  0.889062    0.889062
macro avg      0.890171  0.889286  0.889643  640.000000
weighted avg   0.888982  0.889062  0.888943  640.000000

=== Predicted Class Distribution ===
   class_id  count class_name
0         0    140      World
1         1    160     Sports
2         2    154   Business
3         3    186   Sci/Tech
```

Figure 5: Classification Report

Figure 6 shows the model progression with the impact of different architectural choices and hyperparameters on model performance. Each version builds upon the previous one, demonstrating incremental improvements in accuracy while staying within the 1M parameter constraint.

Model_Evolution_and_Comparison

Version	Architecture & Parameters	Target Modules	Rank (r)	Alpha	Dropout	Trainable Params	Accuracy (%)
V1	Base RoBERTa + LoRA	query	4	8	0.1	589824	86.45
V2	+ key module	query, key	4	8	0.1	687104	87.23
V3	+ value module	q, k, v	4	8	0.05	784384	87.89
V4	+ rank tuning	q, k, v	7	16	0.05	980740	88.91

Figure 6: Model Comparison

Summary

The implemented LoRA-adapted RoBERTa architecture demonstrates significant effectiveness in AG News text classification tasks. The success of the model stems from the synergistic integration of parameter-efficient fine-tuning with the pretrained language model. The efficient parameter utilization achieved through low-rank adaptation (rank=7, alpha=16), combined with strategic targeting of query, key, and value attention matrices, results in enhanced model performance with only 980,740 trainable parameters. Additional performance gains are realized through careful hyperparameter optimization and regularization techniques. The final implementation achieves an optimal balance between model expressivity and parameter efficiency, while providing valuable insights into the design of resource-constrained NLP systems.

References

1. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations (ICLR 2021).
2. 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.
3. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems, 28.
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.