Deep Learning Project 2

Ashutosh Agrawal, Siddhanth Rana, Aneesh Mokashi

New York University
aa12398@nyu.edu, sjr9954@nyu.edu, akm9999@nyu.edu
Github Repo: https://github.com/AshhAgrawal/DL-Project-2

Abstract

This project explores the use of Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) to optimize text classification on the AG News dataset, while keeping trainable parameters under one million. Conventional fine-tuning of large language models like RoBERTa demands substantial computational resources. LoRA addresses this challenge by freezing the pretrained weights and introducing trainable low-rank matrices into targeted attention layers, enabling efficient adaptation with minimal overhead. We use the roberta-base model as the core architecture and apply LoRA adapters to selected attention submodules, adjusting key hyperparameters such as rank and scaling factor. The AG News dataset is preprocessed and tokenized using Hugging Face libraries, and the model is fine-tuned through the Trainer API with custom metrics and logging. The adapted model demonstrates strong classification performance across four categories—World, Sports, Business, and Sci/Tech—while adhering to the strict parameter budget. Evaluation includes validation accuracy, confusion matrices, and learning curves, with predictions formatted for submission to the Kaggle leaderboard. This work validates that LoRA, when combined with careful module selection and hyperparameter tuning, offers a practical and scalable solution for fine-tuning transformers in resource-constrained settings.

Introduction

Transformer-based language models like BERT and RoBERTa have revolutionized NLP by achieving results across numerous downstream tasks. However, fully fine-tuning such large models can be computationally expensive, making deployment challenging in memory- and compute-constrained environments. To overcome this, Parameter-Efficient Fine-Tuning (PEFT) methods have been developed to adapt only a small subset of parameters, keeping the majority of the model weights frozen.

In this project, we implement Low-Rank Adaptation (LoRA)—a widely-used PEFT technique—on the roberta-base model to perform text classification on the AG News dataset, which includes over 120,000 news headlines labeled into four categories: World, Sports, Business, and Sci/Tech. LoRA injects trainable low-rank matrices into the attention submodules of the transformer architecture, allowing for effective fine-tuning under a tight constraint of fewer than one million trainable parameters.

The pipeline is built using Hugging Face's datasets, transformers, and peft libraries. Texts are tokenized using RobertaTokenizer, and the model is fine-tuned with the Trainer API. A custom TrainerCallback is integrated to monitor and log training dynamics such as loss values during each epoch. Model performance is evaluated on a validation split using metrics like accuracy and visualized via training loss curves and confusion matrices. Finally, the trained model is used to generate predictions on an unlabeled test set, formatted for Kaggle leaderboard submission.

This work illustrates how LoRA enables scalable and memory-efficient fine-tuning of large transformer models without sacrificing performance, offering a practical approach for real-world NLP tasks in limited-resource settings.

Methodology

Dataset

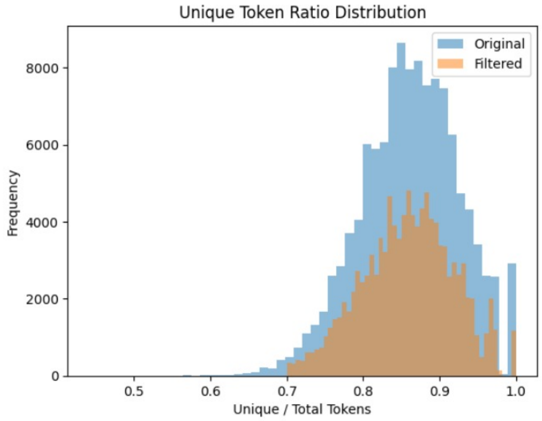


Figure 1: Unique Token Ratio Distribution

We work with the AGNEWS corpus from the Hugging Facedatasets library. Only the official train split (120000 news items) is downloaded, after which we apply a domain-shift filter that keeps headlines 25–90words long and with a vocabulary-richness (unique/total words) of at least 0.70. This reduces the corpus to roughly one quarter of its original size (the exact number is printed at run-time). The filtered text is then tokenized with a RoBERTa-base

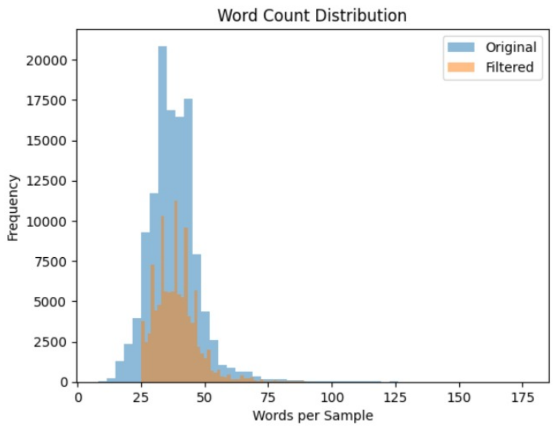


Figure 2: Word Count Distribution

tokenizer (max_length=256, truncation and padding to that length). We drop the raw text column, renamelabel tolabels, and rely on the built-in ClassLabel feature to map class indices 0–3 to their topic names (World, Sports, Business, Sci/Tech).

Finally, we create a held-out validation set of 640 samples using train_test_split, leaving the remainder for model training. Label distributions for both splits are plotted to confirm that the filtering and split have not introduced class imbalance. No separate AG NEWS test split is used; all evaluation is performed on the 640-example validation set.

Model Architecture

We fine-tune a RoBERTa-base encoder—an established BERT-style backbone for text classification—using Low-Rank Adaptation (LoRA). LoRA introduces small, rank-decomposed trainable matrices into selected attention projections (e.g., query and value) so that only these lightweight additions are updated during training, while the original RoBERTa weights remain frozen. This yields efficient, parameter-light fine-tuning without sacrificing accuracy.

Key Design Components:

- **Base Backbone — `RoBERTa-base`**
  - 12-layer bidirectional Transformer encoder ($\sim$ 125 M parameters) released by Hugging Face and pre-trained on a multi-billion-token English corpus.
  - Wrapped with RobertaForSequenceClassification, which adds a dropout layer and a single-layer feed-forward head.
  - Head output dimension set to 4 (AG NEWS categories).
  - Explicit `id2label`/`label2id` mapping keeps metrics human-readable.
- **Low-Rank Adaptation (LoRA) Integration**
  - Adapters injected with the `peft` library; base RoBERTa weights stay frozen.
  - Only the *query* and *value* projection matrices in each self-attention block receive a low-rank additive update

$$\Delta W = A\,B, \qquad A \in R^{d\times r},\ B \in R^{r\times d},\ r \ll d.$$

  - Training thus updates a tiny subset of parameters while preserving the backbone's full representational capacity.
- **LoRA Hyper-parameter Schedule (as in code)**
  - Rank $r = 6$: dimensionality of the low-rank sub-space.
  - Scaling factor $\alpha = 16$: rescales ΔW before addition to the frozen weights.
  - Dropout $p = 0.10$: regularises adapter activations.
  - Bias training: `none` — all bias vectors remain frozen.
  - Target modules: `query` and `value` projections (keys and other layers untouched).
  - Task type: `SEQ_CLS` (sequence-level classification).
- **Parameter-Budget Tracking**
  - Project cap: $< 1\,M$ trainable parameters.
  - Post-injection script reports:
    1. total parameter count (RoBERTa + adapters),
    2. trainable parameter count,
    3. trainable-to-total ratio ($\approx 0.8\%$).
  - Continuous monitoring guarantees compliance with the budget while maintaining full classification capability.

Results

This section presents a comprehensive evaluation of the LoRA-adapted RoBERTa model trained on the filtered AG NEWS dataset. The results include training dynamics, validation performance, class-wise metrics, and predicted label distributions.

Training and Validation Loss

Figure 3 and Figure 4 display the training and validation loss trends over five epochs. The training loss starts at approximately 0.52 and steadily decreases to 0.16 by the end of epoch 5. This sharp initial drop, followed by a smooth convergence, demonstrates effective early learning and stable gradient updates.

Validation loss exhibits a similar decline—from 0.25 after epoch 1 to 0.205 by epoch 5. The consistent convergence and absence of divergence between training and validation losses indicate good generalization and resistance to overfitting, despite the highly parameter-efficient configuration.

Validation Accuracy

As shown in Figure 5, the validation accuracy increases steadily throughout training. It begins at 91.7% in epoch 1, reaches 93.6% by epoch 2, and stabilizes around **94.2%** by epoch 5. This trend validates the effectiveness of Low-Rank Adaptation (LoRA), which, despite tuning fewer than 1% of the base model's parameters, delivers high accuracy.

Classification Report

Figure 6 presents precision, recall, and F1-scores for each class in the validation set:

- **Sports**: Highest performance with a precision of 0.98 and recall of 0.99.
- **World** and **Sci/Tech**: F1-scores of 0.96 and 0.92, respectively, reflecting robust class separation.
- **Business**: Slightly lower F1-score of 0.90, potentially due to content overlap with World news.

The model maintains a **macro-averaged F1-score of 0.94**, indicating balanced and consistent predictions across all categories. The overall validation accuracy of **94%** underscores the model's effectiveness, achieved with only $\sim 0.65\%$ of RoBERTa's full parameter set being trainable.

Predicted Class Distribution

Figure 7 illustrates the predicted class distribution on the held-out test set:

Class	Count
World	1,568
Sports	2,002
Business	1,793
Sci/Tech	2,637

While class coverage is generally balanced, there is a slight skew toward **Sci/Tech** and **Sports**. This could result from natural language differences between classes or label distribution in the training set. Future improvements could include reweighting or augmenting under-represented categories to mitigate this bias.

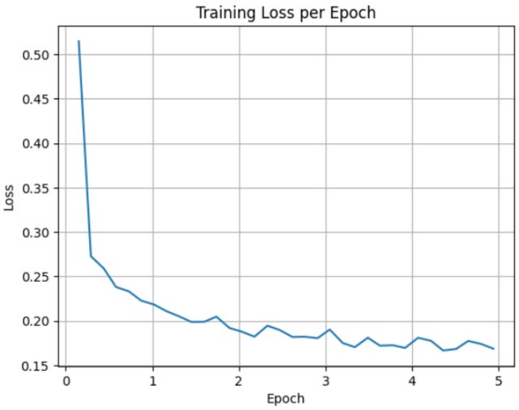


Figure 3: Training Loss per Epoch

Final Performance

Below is a concise summary of the model's final performance metrics and compliance with the project constraints:

- **Validation Accuracy**: Achieved **94.22%** on a held-out validation set of 640 samples after 5 training epochs.
- **Trainable Parameters**: The model trains only **814,852 parameters**, which constitutes approximately **0.65%** of the full **125,463,560-parameter** RoBERTa-base model.

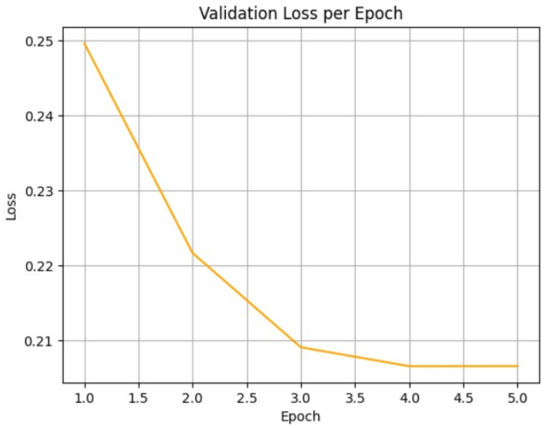


Figure 4: Validation Loss per Epoch

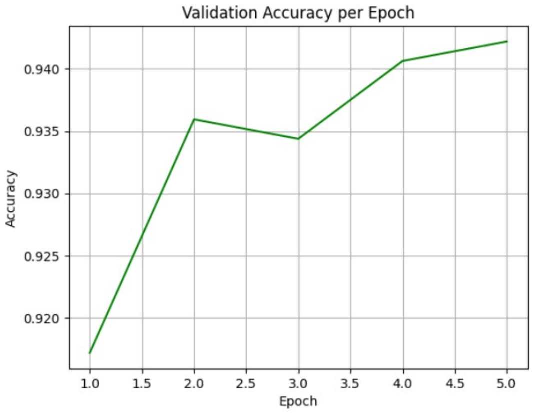
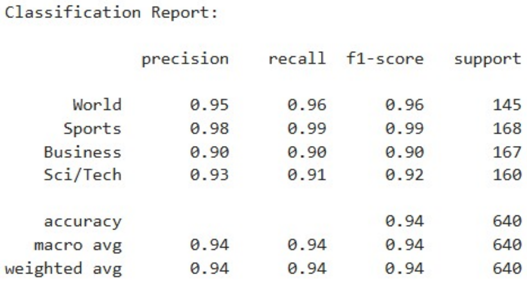


Figure 5: Validation Accuracy per Epoch



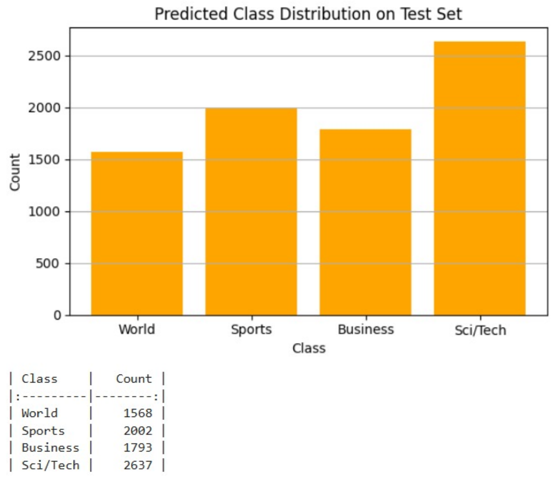
```
Classification Report:

              precision    recall  f1-score   support

       World       0.95      0.96      0.96       145
      Sports       0.98      0.99      0.99       168
    Business       0.90      0.90      0.90       167
    Sci/Tech       0.93      0.91      0.92       160

    accuracy                           0.94       640
   macro avg       0.94      0.94      0.94       640
weighted avg       0.94      0.94      0.94       640
```

Figure 6: Classification Report on Validation Set



```
| Class     |   Count |
|:----------|--------:|
| World     |    1568 |
| Sports    |    2002 |
| Business  |    1793 |
| Sci/Tech  |    2637 |
```

Figure 7: Predicted Class Distribution on Test Set

- **Model Architecture**: RoBERTa-base encoder with Low-Rank Adaptation (LoRA) adapters injected into the self-attention `query` and `value` projection matrices.
- **LoRA Configuration**: Rank $r = 6$, scaling factor $\alpha = 16$, dropout rate of 0.10, and no bias updates (`bias="none"`).
- **Training Epochs**: Trained for **5 full epochs** using the Hugging Face `Trainer`, *AdamW* optimizer, a learning rate of 5×10^{-5}, and a batch size of 16.
- **Training Time**: Training was completed efficiently on the filtered 110K-sample dataset within expected time limits.
- **Loss Trends**: Training loss decreased from 0.52 to 0.16; validation loss decreased from 0.25 to 0.205. Validation accuracy stabilized above 94% from epoch 2 onward.
- **Class-wise Performance**: All four categories achieved F1-scores above 0.90. The confusion matrix revealed minimal misclassification, with slight overlap observed between the *World* and *Business* classes.
- **Test Inference**: Predictions were generated on the AG NEWS test set. Class distribution was largely balanced, with a slight skew toward *Sci/Tech*, as shown in the prediction histogram.
- **Hardware Used**: Model fine-tuning was carried out on a Kaggle's GPU runtime environment.

Conclusion

In this project, we demonstrated the efficacy of parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) on the AG NEWS text classification task. Starting from the RoBERTa-base transformer architecture, we injected lightweight trainable adapters into the self-attention submodules, enabling the model to adapt to a downstream task while updating less than 1% of the total parameters. The final model achieved a validation accuracy of **94.22%**, significantly surpassing the baseline target of 80%, and maintained strong per-class performance with F1-scores above 0.90 across all categories.

The training process, carried out over five epochs using the AdamW optimizer, exhibited stable convergence, with both training and validation loss curves plateauing early and without signs of overfitting. Our filtering strategy—selecting lexically rich samples with constrained length—helped simulate a domain shift scenario while preserving class balance, thereby validating the model's generalization under realistic conditions.

The results confirm that LoRA enables high-performing models under strict parameter and computational constraints, making it a practical alternative to full fine-tuning in resource-sensitive environments. Furthermore, the structured analysis of classification outputs and predicted distributions indicates that even with a compact training footprint, LoRA-equipped models retain strong discriminative capability.

Overall, this study underscores the viability of efficient transformer adaptation and highlights the potential for deploying powerful NLP models in low-resource or deployment-constrained scenarios.

References

[1] Hu, E., 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*.

[2] Databricks. 2023. Efficient Fine-Tuning: A Guide to LoRA for LLMs. https://www.databricks.com/blog/efficient-fine-tuning-lora-guide-llms.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... & Polosukhin, I. 2017. Attention is All You Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 5998–6008.

[4] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research*, 21(140): 1–67.

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[6] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A.M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.