Leveraging LoRA for Efficient Fine-Tuning of Large Models in NLP

Shreeman Agrawal
Indian Institute Of Science
shreemana@iisc.ac.in

Abstract

This report examines the application of pretrained large language models (LLMs) such as T5-Small, BART-Base, and PEGASUS to two key Natural Language Processing (NLP) tasks: document classification and text summarization. The models were fine-tuned using LoRA (Low-Rank Adaptation) to efficiently update their parameters with minimal computational cost. By conducting experiments on benchmark datasets, we demonstrate the effectiveness of LoRA fine-tuning in adapting these models to the specific needs of document classification and summarization. The results reveal task-specific performance and efficiency trade-offs, highlighting the potential of LLMs in real-world applications.

1 Introduction

Recent advancements in Natural Language Processing (NLP) have been driven by large language models (LLMs), such as T5 [2] , BART [3], and PEGASUS [1], which have demonstrated state-of-the-art performance across a variety of tasks. These models, pretrained on vast amounts of text data, have significantly advanced the field by providing generalizable solutions to complex language tasks such as document classification, question answering, natural language inference (NLI), and text summarization. The ability to fine-tune these models on specific tasks has made them highly adaptable, offering flexibility in a wide range of NLP applications.

Document classification and text summarization are two fundamental tasks in NLP. Document classification involves categorizing documents into predefined categories, enabling applications such as spam detection, sentiment analysis, and topic categorization. Text summarization, on the other hand, aims to generate concise summaries from larger bodies of text, which is crucial for applications in news aggregation, legal document analysis, and scientific research.

In this report, we explore the use of three widely adopted pretrained models: T5-Small, BART-Base, and PEGASUS, for the tasks of document classification and text summarization. Specifically, we apply Low-Rank Adaptation (LoRA)[**?**] fine-tuning techniques to efficiently adapt these models for our specific tasks. LoRA fine-tuning offers a computationally efficient method to update only a small number of parameters while retaining the pretrained model's generalization capabilities. Through experiments conducted on benchmark datasets, we evaluate the performance of these models on both tasks, providing insights into their strengths and weaknesses in real-world applications.

2 Problem Statement

Despite the impressive advancements in Natural Language Processing (NLP), several challenges remain in adapting large pretrained models for efficient use in specific tasks, especially under computational constraints. Pretrained models such as T5, BART, and PEGASUS have set new benchmarks in a wide range of NLP applications, but fine-tuning them for specific tasks like document

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

classification and summarization often involves substantial computational resources. This issue becomes even more prominent when working with large-scale models, as they require extensive memory and processing power.

In the context of document classification, the challenge lies in efficiently categorizing documents into multiple categories while retaining high accuracy. This becomes difficult when the dataset is large or highly diverse, leading to the need for models that can generalize well across different types of documents while being computationally efficient.

Similarly, text summarization is a critical task in various domains, such as news aggregation, legal document summarization, and scientific research. Traditional extractive summarization methods often fall short in generating coherent, human-like summaries, whereas abstractive methods, though more promising, are computationally expensive. Fine-tuning large models like BART or PEGASUS on summarization tasks poses significant computational challenges due to their large number of parameters.

The problem addressed in this report is twofold:

1. How to efficiently fine-tune large pretrained models like T5-Small, BART-Base, and Pegasus for document classification and summarization tasks, with minimal computational overhead.
2. How to achieve competitive performance on these tasks using Low-Rank Adaptation (LoRA) fine-tuning, which allows the models to be adapted for task-specific use without the need for extensive computational resources.

The goal is to explore the potential of LoRA fine-tuning in balancing computational efficiency with task performance, making it feasible to deploy state-of-the-art models in environments with limited computational resources, without compromising too much on accuracy.

3 Results

The results were obtained by fine-tuning pretrained models, including T5-Small, BART-Base, and PEGASUS-Small, on several datasets: CNN/Daily Mail for text summarization, Gigaword for headline generation, and SST-2 for sentiment classification. Fine-tuning was performed using the Hugging Face Transformers library, where LoRA (Low-Rank Adaptation) was applied to optimize the models' parameters with minimal computational cost. For the low-rank approximation, a rank of 8 was used for T5 and BART, while PEGASUS utilized a rank of 4. For summarization tasks, ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L) were used to evaluate model performance, while for document classification, accuracy, F1-score, and precision were the key metrics. The models were trained using the AdamW optimizer with a learning rate of 5×10^{-5}, and the training was conducted for 3 epochs with a batch size of 16 (8 for PEGASUS) on a single GPU. The evaluation results highlight T5-Small's superior performance in certain tasks, achieving competitive results with significantly reduced computational overhead compared to larger models. PEGASUS was trained for classification using seq-to-seq generation, with the accuracies reported for exact matches, and less than 1% of outputs were considered garbage. Link for code - github.com/shreeman1000/DLNLP_term_paper

Model	**CNN ROUGE-1**	**CNN ROUGE-2**	**CNN ROUGE-L**
T5-Small	27.2	15.6	26.1
BART-Base	30.2	14.1	28.5
Pegasus	**32.3**	**16.3**	**30.1**

Table 1: Performance of T5-Small, BART-Base, and PEGASUS on CNN/Daily datasets. ROUGE scores are used as metrics.

Model	**Gigaword ROUGE-1**	**Gigaword ROUGE-2**	**Gigaword ROUGE-L**
T5-Small	29.1	13.7	27.5
BART-Base	30.5	14.2	29.0
Pegasus	**33.1**	**15.6**	**31.5**

Table 2: Performance of T5-Small, BART-Base, and PEGASUS on the Gigaword dataset. ROUGE-1, ROUGE-2, and ROUGE-L scores are used as metrics.

Model	Accuracy (%)	F1-Score (%)
T5-Small	87.84	88.27
BART-Base	**92.20**	**92.46**
PEGASUS	90.00	91.00

Table 3: Performance of T5-Small, BART-Base, and PEGASUS on SST-2 News dataset for document classification. Accuracy and F1-score are used as metrics.

Model	Accuracy (%)	F1-Score (%)
T5-Small	**90.4**	**90.8**
BART-Base	86.5	86.2
PEGASUS	88.4	88.0

Table 4: Performance of T5-Small, BART-Base, and PEGASUS on AG News dataset for document classification. Accuracy, F1-score, and Precision are used as metrics.

4 Analysis and Discussion

The results from fine-tuning large models such as **T5-Small**, **BART-Base**, and **PEGASUS** for downstream tasks using **LoRA** show promising performance across multiple datasets, including **CNN/Daily Mail**, **Gigaword**, **SST-2**, and **AG News**. The use of **LoRA** for fine-tuning significantly reduces computational costs by introducing low-rank adaptation layers, which allows for efficient training of large pre-trained models without requiring massive computational resources. For the low-rank approximation, a rank of 8 was used for **T5-Small** and **BART-Base**, while **PEGASUS** utilized a rank of 4.

Below is a detailed analysis and discussion based on the results:

4.1 CNN/Daily Mail Dataset (ROUGE Scores)

PEGASUS consistently outperforms both **T5-Small** and **BART-Base** across all ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L). The high scores suggest that **PEGASUS** is particularly suited for extractive and abstractive summarization tasks, as expected from its design. **BART-Base** performs better than **T5-Small** in terms of ROUGE-1 and ROUGE-L, with **T5-Small** trailing slightly behind in most metrics. However, **T5-Small** remains a competitive model for summarization tasks, especially when computational efficiency is considered.

4.2 Gigaword Dataset (ROUGE Scores)

Similar to the CNN/Daily Mail dataset, **PEGASUS** leads in all ROUGE scores, further solidifying its capabilities in handling large-scale text summarization tasks. **BART-Base** again performs slightly better than **T5-Small**, particularly in ROUGE-1 and ROUGE-L, but both fall behind **PEGASUS**. These results indicate that **PEGASUS**'s ability to generate high-quality summaries is a key strength for tasks involving large datasets like Gigaword.

4.3 SST-2 Dataset (Accuracy and F1-Score)

BART-Base outperforms **T5-Small** and **PEGASUS** on the **SST-2** sentiment analysis task in terms of both **accuracy** and **F1-score**. This demonstrates the strong performance of **BART-Base** on text classification tasks. **T5-Small** shows a solid performance, while **PEGASUS** also delivers competitive results, although it slightly lags behind **BART-Base**. The **F1-score** values are quite close across the models, indicating that all three models are effectively capturing the sentiment of the text.

4.4 AG News Dataset (Accuracy and F1-Score)

T5-Small achieves the highest **accuracy** and **F1-score** on the **AG News** dataset, outperforming both **BART-Base** and **PEGASUS**. This suggests that **T5-Small** is particularly well-suited for multi-class document classification tasks like AG News. **BART-Base** performs well but is outperformed by both **T5-Small** and **PEGASUS**. It suggests that **T5-Small**'s architecture may be more appropriate for

handling the nuances in this type of classification task. **PEGASUS** shows competitive performance but lags behind **T5-Small** in this specific task.

4.5 General Observations

- **PEGASUS** tends to outperform both **T5-Small** and **BART-Base** in summarization tasks (CNN/Daily Mail, Gigaword), indicating that its architecture is specifically tailored for such tasks.
- **BART-Base** is the top performer for sentiment analysis (SST-2), demonstrating its effectiveness in capturing semantic nuances in text classification.
- **T5-Small** is a highly efficient model, performing strongly in document classification tasks (SST-2, AG News) while requiring less computational overhead compared to larger models like **BART-Base** and **PEGASUS**.
- **LoRA** proves to be an effective technique for fine-tuning large models, as evidenced by the high performance of all models across the various tasks. The low-rank adaptation layers allow fine-tuning on downstream tasks with reduced memory and computational cost, which is critical when working with large pre-trained models.

4.6 Conclusion

The performance of **T5-Small**, **BART-Base**, and **PEGASUS** on different datasets demonstrates the versatility of these models across various natural language processing tasks. While **PEGASUS** excels in summarization, **BART-Base** performs strongly in sentiment analysis, and **T5-Small** achieves competitive results in document classification. The use of **LoRA** for fine-tuning these large models has proven effective, allowing for high-quality results with reduced computational resource requirements, requiring only 8 to 10 GB of GPU memory for training, which is extremely low compared to fine-tuning the original models.

References

[1] J. Zhang, Y. Zhao, and Y. LeCun, "PEGASUS: Pretraining with Extracted Gap-sentences for Abstractive Summarization," in *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*.

[2] C. Raffel, R. Shinn, A. Roberts, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, 2020.

[3] M. Lewis, Y. Liu, N. Goyal, et al., "BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

[4] J. Hu, H. Shen, and E. Xie, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*.