University *of Ljubljana*
Faculty *of Computer and Information Science*

Parameter-Efficient Fine-Tuning of Large Language Models

Ondřej Komín, Andrej Sušnik, Eileen Vu

Abstract

This is the abstract

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

Since the introduction of attention [1], using large language models (LLMs) such as BERT [2] and GPT [3] has become inevitable to various applications across natural language processing (NLP) domains. These models have demonstrated remarkable capabilities in understanding and generating human-like text. However, to achieve optimal performance in specific tasks, fine-tuning these pre-trained models on task-specific data is often necessary.

This research paper focuses on presenting and comparing various parameter-efficient fine tuning (PEFT) methods in the context of optimizing natural language processing tasks. Through empirical experiments, we aim at exploring the effectiveness of different PEFT approaches in achieving desirable trade-offs between model complexity, computational efficiency, and task performance. Similarly to the work done in [4], we begin our research by reviewing and categorizing popular PEFT methods. We continue by presenting and discussing the theoretical foundations of three specific methodologies and applying these to five different datasets that cover various natural language understanding skills. Lastly, the fine-tuned models will be evaluated based on appropriate performance metrics, computational resources required, and ease of adaptation to different tasks. Through this exploration, we aim at facilitating more resource-conscious approaches to model optimization, accelerating progress in the field of NLP and its applications.

Related Work

Fine-tuning LLMs plays a crucial role in adapting these models to domain-specific tasks. As LLMs are pre-trained on vast amounts of text data, they capture linguistic patterns and semantic information. However, for tasks with specific requirements, such as sentiment analysis or named entity recognition, fine-tuning allows these models to tailor their representations to better suit the task at hand.

Traditional methods of full fine-tuning involve updating all parameters of the pre-trained LLM. The popular GPT-4 model released in early 2024 contains 1.76 trillion weights [5]. Fine-tuning (and storing) this amount of parameters whenever one wants to apply the model to a specific use case however not only requires significant computational resources but also poses a risk of overfitting, especially in scenarios with limited task-specific data.

In order to avoid these problems, PEFT methods have emerged as a solution to the drawbacks of full fine-tuning. These methods aim to optimize neural networks with fewer parameters while maintaining comparable performance to traditional fine-tuning approaches. By reducing the number of parameters updated during fine-tuning, PEFT not only mitigates computational costs but also helps alleviate overfitting concerns. As described in [4], PEFT techniques can be divided into five main categories: additive fine-tuning, partial fine-tuning, reparameterized fine-tuning, hybrid fine-tuning and lastly unified fine-tuning. While some of these methods aim at introducing new trainable parameters for use-case-specific fine-tuning, others reduce the number of trainable parameters by transforming the weights into lower dimensions.

In this paper, we focus on presenting and comparing three PEFT methods in the context of different NLP tasks. Specifically, we investigate the performance of the following methodologies: low rank adaptation (LoRA), soft prompt-based fine-tuning, and partial fine-tuning. **LoRA** [6] has become very popular in the last years due to its ability to reduce the number of parameters without introducing additional latency, unlike for example adapter methods. This lowers training computational requirements while improving performance for specific

NLP tasks. **Soft-prompting** [7] is a machine learning technique that offers subtle guidance to models during training, aiding in learning without the need for explicit labels. This approach is valuable as it allows for more flexible decision-making while still achieving desired outcomes, especially in scenarios where labeled data may be scarce or costly to obtain. Lastly, we will investigate the partial fine-tuning method **BitFit**[8]. This method only fine-tunes the bias term of the layers while freezing the rest of the network. This technique, which trains less than 0.1% of the total number of weights, was proven to achieve comparable performance than full fine-tuning.

Benchmark	**NLP Task**
CommonsenseQA [9]	Commonsense Reasoning
CoNLL-2012 [10]	Coreference Resolution
XSum [11]	Text Summarization
SST5 [12]	Sentiment Analysis
Slovene SuperGLUE [13]	General NLP Evaluation (Slovene)

Table 1. Chosen benchmarks for performance evaluation

We will provide an empirical comparison of these three methodologies based on five different NLP tasks. The benchmarks we have chosen for this each represent distinct natural language understanding skills allowing us to provide a comprehensive overview of the advantages and disadvantages of all techniques.

Methods

Results

Discussion

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training, 2018.

[4] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023.

[5] OpenAI. Gpt-4 technical report, 2024.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[7] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[8] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

[9] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019.

[10] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40, 2012.

[11] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018.

[12] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[13] Aleš Žagar and Marko Robnik-Šikonja. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France, June 2022. European Language Resources Association.