Minimal Effort, Maximum Effect? An Evaluation of Parameter-Efficient Fine-Tuning in Varying Data Regimes

Amitesh badkul
Department of Computer Science, Graduate Center
City University of New York, New York
abadkul@gradcenter.cuny.edu

Abstract

This research aims to investigate several Parameter-Efficient Fine-Tuning (PEFT) techniques—specifically We will implement Low-Rank Adaptation (LoRA), Dynamic LoRA (DoRA), BitFit, Prefix Tuning, and Adapter Tuning for fine-tuning pre-trained large language models (LLMs) for the important downstream natural language processing (NLP) tasks, these involve - Hate speech detection, using the ethos dataset, and paraphrase detection, utilizing the Quora Question Pairs (QQP). We use the pretrained bidirectional encoder representations from transformer (BERT) as our LLM to provide each PEFT's performance in both the tasks as well as the computational efficiency, scalability, resource utilization. The expected outcome is a comprehensive analysis of which PEFT techniques yield the best trade-off between performance and efficiency for BERT, guiding future NLP model adaptation in resource-constrained settings.

1 Introduction

With the rapid growth of attention mechanisms, transformers, and LLMs such as GPT, BERT, and others, our ability to tackle various NLP tasks has significantly improved, setting a new standard for interpreting human language. However, since these LLMs are pre-trained on large corpora of general language, as their internal architectures become larger and the number of transformer layers increases, fine-tuning them—which involves retraining all model parameters—becomes increasingly difficult and computationally expensive, as it requires updating LLM's millions of parameters, particularly for low-resource systems. To address this, we focus on PEFT methods, which adjust only additional parameters while keeping most of the pretrained model frozen (Chen et al., 2022). We examine two key tasks: hate speech detection and paraphrase detection task are important NLP tasks with broad applications. Hate speech detection is an important NLP task with broad social implications. It involves classifying whether a piece of text contains hateful or offensive content. The paraphrase detection task is the task of determining whether two questions have the same intent. Recent PEFT techniques – including LoRA, DoRA, BitFit, Prefix Tuning, and Adapter Tuning – enable faster and efficient adaptation with far fewer trainable parameters.

2 Literature Review

Several PEFT approaches have been proposed in recent years. Below we review five methods relevant to this project and their significance:

1. Adapter Tuning: Introduced by Houlsby et al. (Houlsby et al., 2019) adapter tuning inserts small trainable adapter modules (dense layers with a bottleneck) at each Transformer layer and fine-tunes only these adapters while freezing the rest of the model. Adapter tuning has been successfully extended to enable multi-task composition, such as AdapterFusion (Pfeiffer et al., 2021), showing strong results on NLP transfer learning tasks like GLUE.

2. Prefix Tuning: is a way to adapt LLMs to a new task without changing the model's original weights (Li and Liang, 2021). Instead, it adds a small set of new task-specific vectors (called "prefixes") to the beginning of the model's input at each layer. These prefixes help guide the model's attention toward what's important for the new task. Prefix tuning has been demonstrated to achieve competitive performance to full fine-tuning on text generation tasks and language understanding tasks across T5 models (Li and Liang, 2021; Lester et al., 2021).

3. LoRA (Low-Rank Adaptation): LoRA (Hu et al., 2022) fine-tunes a model by adding trainable low-rank matrices to the existing weight matrices of the model, in doing so it is able to reduce the trainable parameters significantly as well as combat catastrophic forgetting which occurs during fine-tuning. Recent studies have demonstrated the strong effectiveness of LoRA and its variants across diverse NLP tasks: Whitehouse et al. (2024) showed that LoRA performs competitively in both high- and low-resource multilingual summarization settings, while Liu et al. (2024) proposed ALoRA, which dynamically adjusts rank allocation to further enhance adaptation performance (Whitehouse et al., 2024; Liu et al., 2024).

4. DoRA (Dynamic Low-Rank Adaptation): DoRA (Mao et al., 2024) is a recent extension of LoRA that dynamically adjusts the low-rank updates during training. Instead of using a fixed rank, DoRA decomposes a high-rank LoRA update into multiple rank-1 components and employs a dynamic rank allocation to have a higher performance, while maintaining the computational load.

5. BitFit: This method takes a unique approach by fine-tuning only the bias terms of the model's neurons and leaving all weight matrices unchanged (Ben Zaken et al., 2022). This "bias-only" tuning drastically reduces the number of trainable parameters (to well under 1% of the original model).

3 Data

We will use two well-established datasets to evaluate the PEFT approaches on the target tasks:

1. **Ethos**: The ETHOS dataset (Mollas et al., 2022) consists of 998 short-form comments annotated for the presence of hate or offensive speech. Each example is labeled with a binary target: 1 if the content is considered hateful or offensive, and 0 otherwise. The dataset spans a variety of hate categories, including but not limited to gender-based, religion-based, and race-based hate.

2. **Quora Question Pairs (QQP)**: A dataset of sentence pairs from the Quora Q&A platform, labeled 1 if the two questions are paraphrases (duplicate questions) and 0 otherwise (Sharma et al., 2019). It contains over 400,000 question pairs.

These datasets not only differ in task structure—single-sentence classification (ETHOS) versus sentence-pair classification (QQP)—but also represent contrasting data regimes: ETHOS provides a very low-resource setting ($\sim$ 998 examples), while QQP offers a high-resource setting ($\sim 364k$ labeled pairs). This allows us to evaluate how different PEFT methods perform under both data scarcity and abundance, providing a more comprehensive understanding of their efficiency.

4 Experiments

We plan to fine-tune a pretrained BERT model on each task using each of the five PEFT strategies. All experiments will be implemented with the HuggingFace library and it's PEFT library. Key details of our experimental plan include:

1. **Fine-Tuning Setup:** We will keep the majority of BERT's parameters frozen and insert the PEFT modules as required by each method. For LoRA/DoRA, this means adding low-rank update matrices to BERT's attention and/or feed-forward layers. For adapters, we will insert adapter layers at each transformer layer. Prefix tuning will involve prepending a sequence of learnable prefix tokens to each layer's input. BitFit will simply mark all bias terms in BERT as trainable and freeze everything else. We plan to use the same training hyperparameters across all methods to ensure fairness.

2. **Evaluation Metrics:** Model performance will be primarily evaluated using classification accuracy on the test set for both tasks. If time permits, the significance tests or standard deviation will be reported after running the various methods across different seeds.

3. **Efficiency Analysis:** A major aspect of our experiments is measuring fine-tuning efficiency. For each method, we will record the number of trainable parameters introduced. We will also monitor training time per epoch and peak memory usage for each method under the same hardware conditions.

References

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. 2022. Revisiting parameter-efficient tuning: Are we really there yet? *arXiv preprint arXiv:2202.07962*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zequan Liu, Jiawen Lyn, Wei Zhu, Xing Tian, and Yvette Graham. 2024. ALoRA: Allocating low-rank adaptation for fine-tuning large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 622–641, Mexico City, Mexico. Association for Computational Linguistics.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. DoRA: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11662–11675, Bangkok, Thailand. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228, Mexico City, Mexico. Association for Computational Linguistics.