Exploring Soft Prompt Tuning for Parameter-Efficient Fine-Tuning Across Domains in Lower Parametric LMs

Name
Center for Data Science
New York University
name@nyu.edu

Name
Center for Data Science
New York University
name@nyu.edu

Name
Center for Data Science
New York University
name@nyu.edu

Abstract

This work explores soft prompt tuning techniques, such as prefix and prompt tuning, for parameter-efficient fine-tuning in lower-parametric language models (LMs). We apply these techniques to the T5-Large model, achieving over 90% accuracy on cross-domain sentiment analysis tasks, including IMDB, Financial Phrasebank, and Twitter, with less than 1% of the model's parameters compared to fully fine-tuned models like Flan-T5. Our findings demonstrate that soft prompting can provide scalable and effective performance across domains, highlighting its potential as a parameter-efficient alternative to full model fine-tuning. Future work will focus on improving the scalability of soft prompting and comparing its performance in few-shot learning settings.

1 Introduction

In-Context Learning (ICL) has become an influential approach in NLP, allowing language models (LMs) to perform tasks with limited examples directly within prompts (https://arxiv.org/abs/2005.14165). While ICL has proven effective in few-shot learning and instruction-based tuning for zero-shot scenarios, these gains are often achieved in high-parametric models or fully instruction-fine-tuned models—both requiring substantial parameter counts and computational resources.

For many cross-domain tasks, soft prompting offers a parameter-efficient alternative. Techniques like Prompt Tuning and Prefix Tuning enable models to achieve strong task performance without full fine-tuning, making them ideal for settings with resource constraints. Fully fine-tuned models like Flan-T5 (https://arxiv.org/abs/2210.11416) perform robustly across various tasks, but their high parameter requirements limit scalability. In this project, we explore soft prompting on T5-large (https://arxiv.org/abs/1910.10683) for cross-domain sentiment analysis, training on the SST-2 (NEED CITATION) dataset and testing on diverse domains such as finance, movie reviews, and social media. Soft prompting yields comparable accuracy (over 90%) to fully fine-tuned models, using less than 1% of the parameters, underscoring the scalability and adaptability of parameter-efficient fine-tuning methods across domains.

2 Related Work

Soft prompting techniques, such as Prompt Tuning (https://arxiv.org/abs/2104.08691) and Prefix Tuning (https://arxiv.org/abs/2101.00190), have emerged as effective parameter-efficient methods to adapt language models with minimal updates. Instead of fine-tuning the entire model, these approaches add trainable prompt tokens or prefixes to the input, guiding the model's behavior while

preserving most of its parameters. This approach has shown strong performance on specific tasks while significantly reducing computational costs.

In contrast, FLAN (Fine-tuned Language Net) represents a fully instruction-tuned model that excels in zero-shot prompting. By training on a broad set of task instructions, FLAN generalizes effectively across tasks without needing additional fine-tuning, making it highly adaptable in zero-shot settings. However, FLAN's full fine-tuning requires more parameters, limiting its efficiency for constrained setups.

Our work leverages these advances, comparing the cross-domain performance of Prefix Tuning on T5-large with FLAN's fully fine-tuned model, aiming to match accuracy across domains while using only a fraction of the parameters.

3 Approach

Our approach centers on parameter-efficient fine-tuning through Prefix Tuning on the T5-large model. We evaluated both Prompt Tuning and Prefix Tuning, and found Prefix Tuning to be more effective. In Prefix Tuning, trainable prompt tokens, or "prefixes," are prepended to the model's input layer, guiding the model's response to specific tasks without updating the full set of model parameters. This method enables efficient adaptation, using only a small fraction of the parameters needed for full fine-tuning.

Cross-Domain Evaluation: We trained the Prefix-Tuned T5-large model on the SST-2 (Stanford Sentiment Treebank) dataset, then evaluated it on diverse sentiment datasets, including Financial Phrasebank, IMDB, and Twitter. The model achieved greater than 90% accuracy across these domains, demonstrating the strong cross-domain adaptability of soft prompts.

Comparison with Fully Fine-Tuned Models: To assess the efficiency and effectiveness of soft prompting, we compared our Prefix-Tuned T5-large model with the fully fine-tuned Flan-T5-large model, an instruction-tuned variant, across all test datasets, including SST-2. We did zero shot prompting on FLAN T5-Large with a simple prompt asking to classify the given text into positive or negative. The results showed comparable accuracy, with the Prefix-Tuned model even outperforming FLAN in certain cases. This highlights how fine-tuning with less than 1% of parameters through soft prompting can achieve results similar to a fully fine-tuned model.

4 Experiments

This section contains the following.

4.1 Data

For training, we utilized the Stanford Sentiment Treebank (SST-2) dataset, which consists of 67.3K labeled samples in its training split, providing a solid foundation for fine-tuning our model on sentiment analysis. To evaluate cross-domain performance, we tested both the prefix-tuned model and the fully fine-tuned comparison model on a diverse set of datasets. These included 872 samples from the SST-2 validation set, 1,000 samples from the IMDB Movie Sentiment Test Dataset (NEED CITATION), 1,500 samples from the Financial Phrasebank dataset (NEED CITATION), and 1,000 samples from the Twitter sentiment dataset (NEED CITATION). This variety allowed us to comprehensively assess and compare the cross-domain adaptability of soft prompting and full fine-tuning approaches across multiple sentiment analysis domains.

4.2 Evaluation method

For evaluation, we used accuracy score from scikit learn as the primary metric, given that the task was sentiment analysis, a binary classification problem. Accuracy was calculated on all test datasets, including SST-2, IMDB, Financial Phrasebank, and Twitter sentiment datasets. This metric allowed us to directly compare the performance of the prefix-tuned model against the fully fine-tuned Flan-T5 model using zero shot prompting, providing a clear measure of their effectiveness across different domains.

4.3 Experimental details

Parameter	Value
Model	T5-large (used for both Prefix Tuning and full fine-tuning)
Model Configuration	757M parameters
Batch Size	16
Number of Train Samples	67.3K
Learning Rate	1e-4
Number of Epochs	5
Optimizer	AdamW
Trainable Parameters for Prefix Fine-tuning	12M
Training Time	2 hours on a single A100 GPU

Table 1: Model and Training Configuration

4.4 Results

Dataset	Prefix Tuned T5-Large	Flan-T5 Zero-shot
SST-2 (Validation)	92.66	**94.72**
IMDB Movie Sentiment	**96.4**	92.4
Financial Phrasebank	**92.91**	89.19
Twitter Sentiment	91.66	**93.2**

Table 2: Model Performance Comparison

5 Future Work

In future work, we aim to investigate the limitations of soft prompting techniques, particularly their scalability across multiple tasks. One key challenge is improving their reliability when applied to diverse domains and tasks. Our goal is to develop a novel, parameter-efficient fine-tuning approach that addresses these limitations, ensuring robust performance across both domains and tasks. Additionally, we plan to compare this approach's performance in few-shot learning scenarios, contrasting it with In-Context Learning to highlight its effectiveness.