Prefix-Tuning

Mac Turner, Michael Ngo, Eric Hu, Neeraj Parihar

May 10, 2025

1 Introduction

If we want to finetune LLMs for specific tasks, we would store the newly trained weights for every task we fine tune on. LLMs are also powerful enough where prompt engineering help LLMs perform better on specific tasks. *Prefix-Tuning: Optimizing Continuous Prompts for Generation* by Xiang Lisa Li and Percy Liang [1] introduces a method of fine-tuning that is memory efficient like prompt engineering and reaps the performance of full fine tuning.

They show that prefix tuning GPT-2, BART and T5 on table-to-text generation and text summarization performas better than fine tuning and achieves a training speed-up. It is more parameter efficient that other fine tuning methods.

2 Chosen Result

We chose to reproduce prefix-tuning for GPT-2 Medium on table-to-text generation and compare it against full fine tuning. This is the first two rows and the first column of Table 1 in the original paper.

We chose this people it's the first major result of the paper that prefix-tuning is comparable, if not better than fine tuning, and GPT-2 and the E2E table-to-text dataset used were the simplest models and datasets to set up.

3 Methodology

3.1 Prefix-Tuning

We fine-tune GPT-2 Medium. The model is a stack of attention blocks. For each attention block i, we learn the first L keys and values via a network, $\mathsf{MLP}_{\theta,i}$ and input matrix $P_{\theta,i}$. $P_{\theta,i}$ is of dimension $L \times 768$. The MLP is 2 layers. A linear layer from 768 to 800, followed by a tanh activation, and a linear layer from 800 to the embedding dimension of $2 \cdot 768$ (time 2 because of key and value).

To do prefix-tuning, we load a pretrained GPT-2 Medium model from HuggingFace and pass in the learned keys and values through the `past_key_values` keyword. Additionally, generation was done via beam search with beam length of 5.

3.2 Dataset

The dataset is the E2E Table-to-text generation dataset [2]. It is a dataset of tables conaining information about restaurants and the goal is to write a sentence that summaries the tabular

information. Generated or proposed sentences are compared to a list acceptable reference sentences and evaluated with a standard suite of metrics: BLEU, NIST, METEOR, ROUGE-L, and CIDEr. [Insert explanation of metrics.]

3.3 Training

Finally, prefix-tuning is trained by running 5 epochs over the E2E dataset, with learning rate of $8e-5$, batch size of 10, and prefix-length of 5. Full finetuning has similar parameters. We trained on an NVIDIA RTX 4080.

4 Results & Analysis

Model	BLEU	NIST	METEOR	ROUGE-L	CIDEr
Prefix-Tuning (0.1%)	68.5	8.71	44.3	70.4	2.34
Fine-Tuning	**70.3**	**8.94**	**46.2**	**72.2**	**2.47**
Prefix-Tuning (0.1%) [1]	69.7	8.81	46.1	71.4	2.49
Fine-Tuning [1]	68.2	8.62	**46.2**	71.0	**2.47**

Table 1: Results of prefix-tuning and full fine-tuning on the E2E table-to-text generation dataset.

Our prefix-tuning fails to break ahead of full finetuning. But it does do better in some metric than what the original authors found when they fine-tuned. We attribute this to the fact that they used a smaller learning rate for fine-tuning GPT whereas we stuck to $8e-5$ for both prefix-tuning and fine-tuning. We are not sure why we are not reproducing the exact results.

At least, our prefix-tuning shows almost comparable performance to fine tuning. It's also faster to train and memory efficient.

We had to implement beam search from scratch. Also understanding the input-output set up we had to guess a bit and as well as how we actually can implement inputting keys and values into the model.

5 Reflections

If there are little other resources, implement it from scratch. It's much more efifcent to just start to code something that might not work, then see it fail adn dfigure out how to iterate and make it work better htan to spend an eternity trying t figrurout exactly hwo to get it right the first time.

Future ideas for special task-specific tokens to prime our model for flexibility.

References

[1] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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,

August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/. 1, 2

[2] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. The E2E dataset: New challenges for end-to-end generation. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5525. URL https://aclanthology.org/W17-5525/. 1