NLP-final report: Can we teach a model twice?

2100012521 Shaoyang Cui

January 2024

1 Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks. A key aspect of their versatility lies in the potential for fine-tuning to enhance performance on specific downstream tasks. Our interest, however, extends to the relearning ability of LLMs, specifically the feasibility of tuning a model twice, each time for a different task. To explore this, we conducted experiments using the T5 model, applying various Parameter-Efficient Fine-Tuning (PEFT) methods. Our goal was to assess whether an LLM could undergo dual tuning for distinct tasks and still maintain stable performance.

2 Introduction

The field of Natural Language Processing (NLP) has seen remarkable progress since the advent of transformers(Vaswani et al. [10]). Currently, Large Language Models (LLMs) stand as some of the most sophisticated tools in AI. As the pre-train and fine-tune paradigm firstly demonstrated its impressive potential in Computer Vision field(Girshick [3], Krizhevsky et al. [5], Simonyan and Zisserman[9]), it's the central to maximizing the effectiveness of LLMs nowadays as seen in works by Radford et al. [7], Devlin et al. [2], Liu et al. [6], Brown et al. [1], and more.

Given the importance of generalization in AI, a key question emerges: Can a LLM be effectively fine-tuned more than once? To investigate this, we propose experiments with the exciting T5 model [8] and its PEFT version using LoRa [4]. Our approach involves initially fine-tuning these models on a summarization task, followed by a subsequent re-tuning on both machine translation and a different summarization task. The results of these experiments will be critically analyzed to provide deeper insights and stimulate further discussion in the field.

3 Method

3.1 Models

- **T5**: The T5(Transfer Text-to-Text Transformer), is an innovative model in the NLP domain that treats all tasks as text-to-text problems. By unifying various tasks under a single framework, T5 simplifies the complexities typically associated with specialized t, offering a more efficient and versatile solution in the field of language understanding and generation.

- **LoRa**: Low-Rank Adaptation is a PEFT(Parameter-Efficient Fine-Tuning) method designed for adapting large pre-trained models like Transformers with minimal additional parameters. It was introduced to address the challenge of fine-tuning large models, which can be computationally expensive and resource-intensive.

Table 1 shows some information about our model.(We only use the version version of T5 due to limitation of computing resources).

	Model Size	Trainable Parameters	Pretrain Datset
T5-small	60.5M	60,506,624	C4
T5-small(LoRa)	60.5M	1,179,648	C4

Table 1: Caption

3.2 Dataset

- **CNN**: The CNN dataset consists of news articles paired with bullet-point summaries from the CNN website. It's a high quality dataset and widely used for training text summarization models.
- **Xsum**: The XSum (Extreme Summarization) dataset is also a dataset for text summarization tasks. Its articals are from BBC and each accompanied by a single-sentence summary.
- **WMT-16**: WMT-16(**Workshop on Statistical Machine Translation 2016**). It includes various language pairs for machine translation tasks, the one we're using is its **German-English** subset.

For all datasets above, instead of using the whole dataset, we used only a split of them(shown in Table 2).

	Task type	DataSource	Data size(used)	Split[train, validation]
CNN	Summarization	CNN	28711	[10%, 10%]
Xsum	Summarization	BBC	20404	[10%, 10%]
WMT-16 (Ge-En)	Machine Translation	Multi source	136467	[3%, 2%]

Table 2: Datasets

4 Experiments

In our experiment, we initially fine-tuned both the 'vanilla' T5 and T5 with LoRa models on the CNN dataset, specifically focusing on summarization tasks. Subsequently, we re-tuned these models using the XSum and WMT-16 (German-English) datasets, trying to check their performance in both summarization and machine translation. This dual-phase tuning process was designed to investigate two critical aspects:

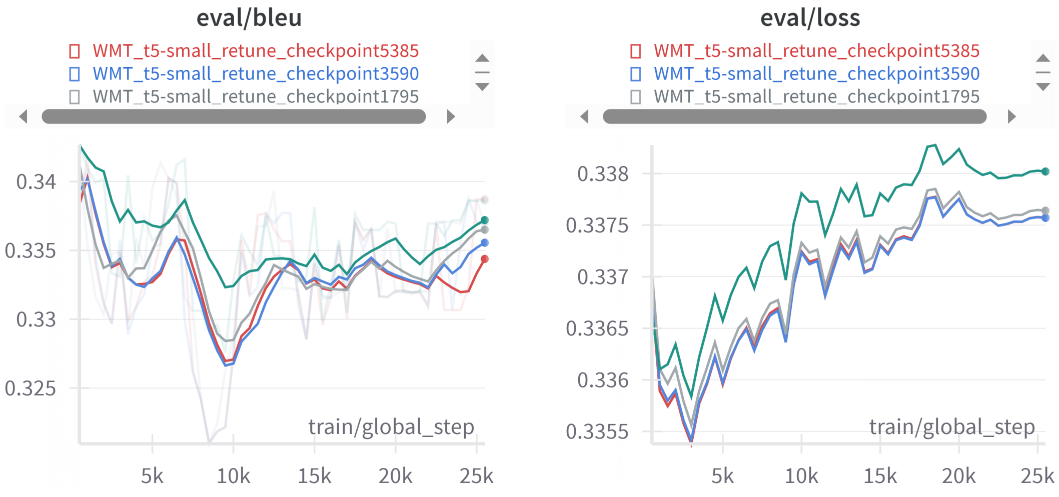
- The influence of the type of task on the re-tuning effectiveness.
- If tuning the model directly and tuning the model with PEFT method yeild the same result.

Here we offer our experimental configurations in table 3

epoch	batch size	lr	random seed	weight decay
3	16	1e-5	2023	0.01

Table 3: Configurations

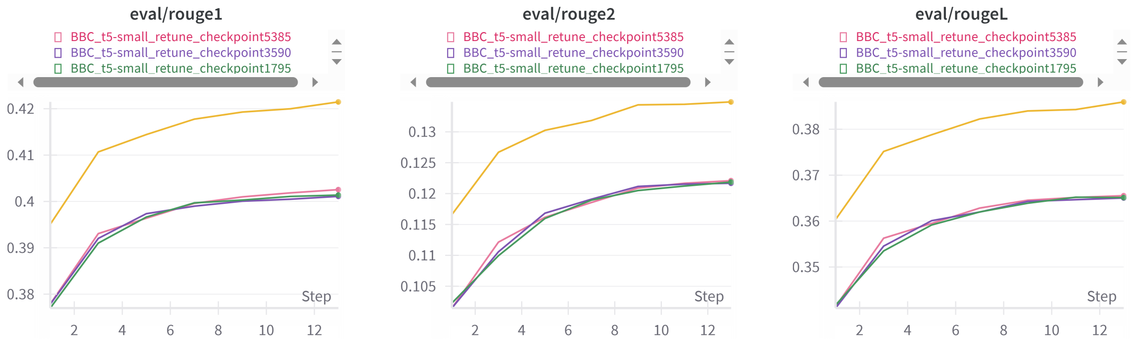
And the experiment results(shown in 1234), and in each experiment, we use 3 different checkpoints of the tuned T5 model for re-tuning.



(a) bleu

(b) eval loss

Figure 1: Comparing Tune-Retune results on WMT : Tune model directly



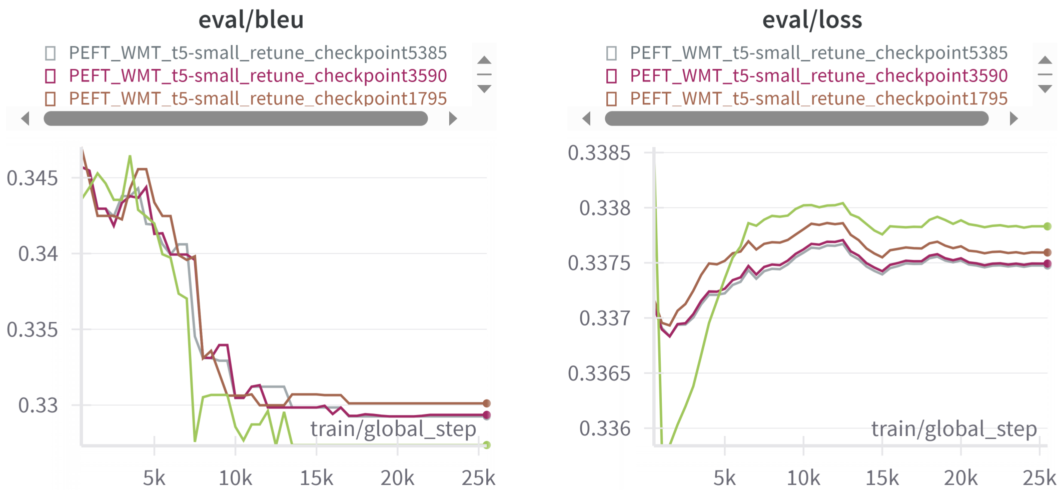
(a) ROUGE-1

(b) ROUGE-2

(c) ROUGE-L

Figure 2: Comparing Tune-Retune results on BBC: Tune model directly

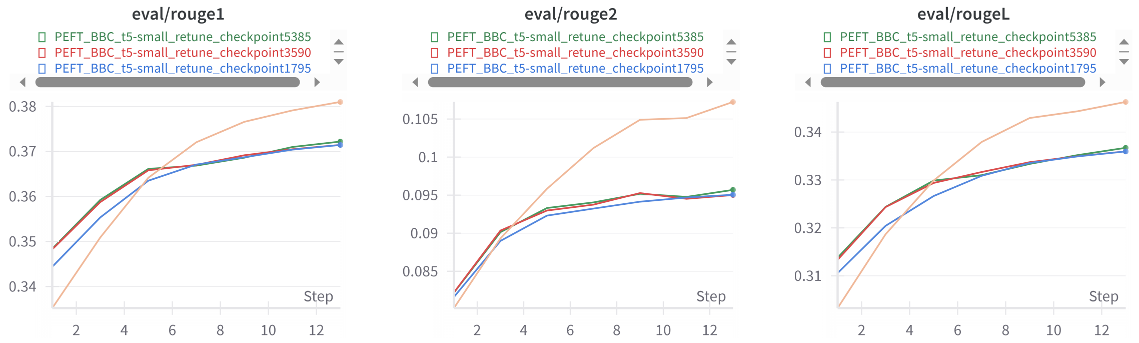
Figure 1, 2 shows the re-tuning results on both Xsum dataset using directly tuned T5 model(on CNN dataset). We found that for both two re-tune tasks, the untuned model beat the tuned model(Green curve in Figure 1 and yellow curve in 2). And also, we observed that for WMT task, a earlier checkpoint model(which means its not that sufficiently tuned on CNN) performs comparatively better. This observation can be an evidence of the model couldn't be tuned twice for different task type.



(a) bleu

(b) eval loss

Figure 3: Comparing Tune-Retune results on WMT : Tune model with PEFT LoRa



(a) ROUGE-1

(b) ROUGE-2

(c) ROUGE-L

Figure 4: Comparing Tune-Retune results on BBC: Tune model with PEFT LoRa

Figures 4 and 3 illustrate the re-tuning results of the T5 model using the PEFT method, LoRa. We observed that a previously tuned model generally starts at a higher initial point but eventually converges to a lower value compared to an untuned model across all three metrics. This could be attributed to the tuned model becoming trapped in local minima during the second tuning process.

However, an interesting trend emerged when tuning the model with LoRa for the WMT task. We noted that the BLEU metric consistently decreased. Consequently, a model that had been previously tuned did not experience as significant a drop in performance during re-tuning, leading to a higher convergence BLEU score.

5 Conclusion

Through comparative experimental results, we found that for a T5 model that has been fine-tuned on the CNN dataset (summarization task), the effect of tuning the model a second time, whether on the Xsum dataset or the WMT dataset (regardless of whether the second tuning task is related to the first tuning task), is not as good as directly tuning the original model.

Additionally, this phenomenon occurs both when directly tuning the model and when using the PEFT method to tune the model. A notable exception is observed when using PEFT to experiment with the WMT dataset; the final convergence value of the BLEU score for the model tuned twice is better than that of the model tuned once, which does not occur when the PEFT method is not used.

Despite this exception, based on the overall experimental results, we tend to conclude that (at least for T5 model) without using special methods (such as data integration, modifying model structures, etc.), the performance achieved by using a model already tuned for a certain task is not as good as directly tuning a new model for a task.

6 Discussion

It is important to acknowledge that our conclusions on this matter cannot be definitively drawn from our limited experiments alone. For a more comprehensive understanding, several aspects warrant further exploration:

\

1. The inclusion of a wider variety of task types beyond just Summarization and Machine Translation is essential. Moreover, our investigation should encompass not just Seq2Seq tasks but a broad spectrum of task categories.

2. Expanded experimentation across a diverse range of models is necessary (our study was confined to the T5 model). An important question to consider is the impact of the size of Large Language Models (LLMs) on the outcomes.

3. A broader range of PEFT methods should be evaluated to ensure a more holistic analysis.

4. Our research should extend to a greater number of datasets, rather than being limited to the three we used.

The question of whether a model can be effectively taught twice presents an intriguing area of study, ripe with opportunities for discovery and innovation. This inquiry opens up a vibrant avenue for future research, beckoning us to unravel the complexities of model learning and adaptation in the ever-evolving landscape of machine learning.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

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

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.

[4] 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.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

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

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.