PARAMETER-EFFICIENT FINE-TUNING IN READING COMPREHENSION

Rustam Abdumalikov[1]

[1]Institute of Computer Science, University of Tartu

UNIVERSITY OF TARTU

Abstract

Question Answering is an important task in Natural Language Processing. There are different approaches to answering questions, such as using the knowledge learned during pre-training or extracting an answer from a given context, which is commonly known as reading comprehension. One problem with the knowledge learned during pre-trained is that it can become outdated because we train it only once. Instead of replacing outdated information in the model, an alternative approach is to add updated information to the model input. However, there is a risk that the model may rely too much on its memorized knowledge and ignore new information, which can cause errors. Our study aims to analyze whether parameter-efficient fine-tuning methods could improve the model's ability to handle new information. We assess the effectiveness of these techniques in comparison to traditional fine-tuning for reading comprehension on an augmented NaturalQuestions dataset. Our findings indicate that parameter-efficient fine-tuning leads to a marginal improvement in performance compared to fine-tuning. Furthermore, we observed that data augmentations contributed the most substantial performance enhancements.

Example

Context: The nearby Spanish settlement of St. Augustine attacked Fort Caroline, and killed nearly all the French soldiers defending it. The Spanish renamed the fort San Mateo [...]

Original Question: What was Fort Caroline renamed to after the Spanish attack? San Mateo (0.98)

Fort Caroline → Robert Oppenheimer

Adversarial Question: What was Robert Oppenheimer renamed to after the Spanish attack? San Mateo (0.99)

This example adopted from [2].

Research Question

Can **Parameter-Efficient Fine-Tunings** give better generalization over Fine-Tuning in reading comprehension?

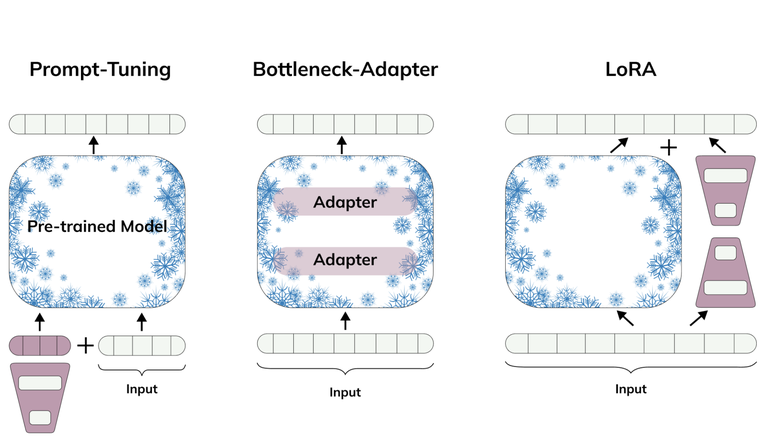
Dataset

In this study, we will utilize the NaturalQuestions datasets, which comprise questions that people asked using Google Search. The datasets consist of long answers that represent Wikipedia passages and short answers that are embedded within the long answers. Additionally, we will incorporate data augmentation techniques, including counterfactuals and answerability. Counterfactual augmentation involves substituting factual answers within a given context with random alternatives. Whereas, answerability augmentation focuses on randomizing the context for all the questions, and erasing context all together for all the questions.

	Factual	Counterfactual	Answerability
Train	85,540	30,653	85,540
Validation	21,386	7,698	21,386
Test	1,365	1,365	1,365

f	✓		
f+cf	✓	✓	
f+a	✓		✓
f+cf+a	✓	✓	✓

Parameter-Efficient Fine-Tuning Methods



Results

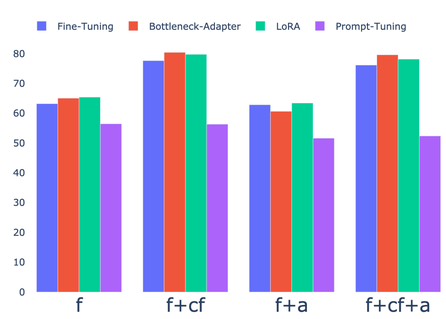


Fig. 2: Evaluation on counterfactual test split

- **Counterfactual test split**: parameter-efficient fine-tuning gave better performance. However, it was just marginal improvement.
- **Empty test split**: all the methods achieved perfect accuracy i.e. 100%.
- **Random test split**: all the methods except prompt-tuning achieved near-perfect accuracy i.e. ∼98%.

Random context

Question: Panda is a national animal of which country?

Context: The fourth season began airing on October 10, 2017, and is set to run for 23 episodes on The CW until May 22, 2018.

ChatGPT context

Question: Panda is a national animal of which country?

Context: In the dusty streets of Beijing, a young girl named Mei-Ling gazed up at the towering skyscrapers, dreaming of one day traveling the world to see the Great Wall of China.

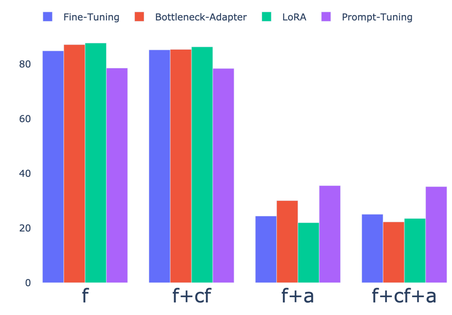


Fig. 3: Evaluation on ChatGPT context with factual answer

ChatGPT counterfactual context

Question: Panda is a national animal of which country?

Context: In the dusty streets of Beijing, a young girl named Mei-Ling gazed up at the towering skyscrapers, dreaming of one day traveling the world to see the Great Wall of Hawaii.

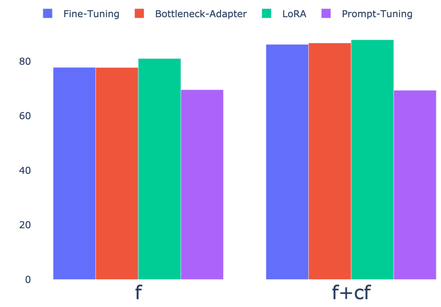


Fig. 4: Evaluation on ChatGPT context with counterfactual answer

Conclusions

The objective of this work was to investigate whether better generalization in reading comprehension could be achieved by using parameter-efficient fine-tuning (PEFT) methods. Several methods, such as LoRA, Bottleneck-Adapter, and prompt-tuning, were utilized and tested against the Natural Question dataset, which was augmented with counterfactual and answerability augmentations to force the model to pay more attention to the input. Our results showed that PEFT methods did not provide a significant performance advantage over fine-tuning due to several factors, including the use of a noisy dataset, a rigid strategy for answer validation, and model insensitivity to details.
Our study also found that while answerability augmentation improved model robustness, counterfactual augmentation led to a reduction in model robustness by inducing simplistic extractor behavior. Prompt-tuning demonstrated that these augmentations were not complementary, with improvements in the answer abstaining task leading to a degradation in the answer extraction task. Interestingly, prompt-tuning revealed that answer abstaining was the easiest task, and established the baseline performance of the pre-trained model on the answer extraction task.

Acknowledgements

I am deeply grateful to my supervisors, Yova Kementchedjhieva and Kairit Sirts, for their unwavering guidance, support, and generosity with their time throughout my academic journey. Their invaluable guidance has been instrumental in the completion of this work, especially during moments when I felt lost or overwhelmed. Their thought-provoking discussions and feedback have challenged me to think critically and creatively about my research and have helped me to develop a deeper understanding of the subject matter.

References

1. Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering, 2022.
2. Johannes Welbl, Pasquale Minervini, Max Bartolo, Pontus Stenetorp, and Sebastian Riedel. Undersensitivity in neural reading comprehension, 2020.