On Evaluating the Robustness of Language Models with Tuning

Colin Wang **Lechuan Wang** **Yutong Luo**
Halıcıoğlu Data Science Institute
University of California, San Diego
{ziw029, l6wang, y5luo}@ucsd.edu

Abstract

Prompt tuning and prefix tuning are two effective mechanisms to leverage frozen language models to perform downstream tasks. Robustness reflects models' resilience of output under a change or noise in the input. In this paper, we analyze the robustness of natural language models using various tuning methods with respect to a domain shift (i.e. training on a domain but evaluating on out-of-domain data). We apply both prompt tuning and prefix tuning on T5 models for reading comprehension (i.e. question-answering) and GPT-2 models for table-to-text generation.

1 Introduction

NLP models have recently achieved outstanding performances and are thus gained prevalent applications in real world. With this popularity, it is important to make sure these models could adapt well in the dynamic circumstances. More specifically, robustness with respect to domain shifts is supposed to be considered when developing models. Because the same large pretrained language models are often applied to different tasks or fields. It would be inefficient and impractical if we train the model with corresponding inputs every time we apply them to a different domain. We want large models can be easily reused. Improvement on models to ensure they are robust against change of inputs has been a hot topic for study.

With the advance of NLP, a wide range of mechanisms have been developed to adjust large pretrained language models to downstream tasks. To avoid the update and storage of language model parameters, Li and Liang (2021) developed prefix tuning, which freezes the parameters of language model, and only optimizes the small continuous task-specific vector (i.e. the prefix). They apply prefix tuning on GPT-2 models, and find great model performances under different data settings.

Prompt tuning (Lester et al., 2021) is proposed as a further simplification of prefix tuning. Similar to prefix tuning, the pretrained language model is frozon, but prompt tuning only allows the pre-1 of k soft prompt to the input data. With the end-to-end employment of prompt tokens, prompt tuning achieves outperforming results and efficient model reuse.

2 Experimental Setup

2.1 Datasets and Metrics

For GPT-2 model, we investigate the robustness respect to domain shift on Table-to-Text generations. We train the model on WebNLG (Colin et al., 2016), and test on DART (Radev et al., 2020). DART is more complex and has larger size than WebNLG. DART is open-domain while WebNLG has only 14 domains. We evaluate the performance using BLEU (Papineni et al., 2002) score, which is reported by the official evaluation script for WebNLG and DART. We will also include METEOR (Satanjeev and Lavie, 2005) and TER (Snover and Dorr, 2006) score, which measures the translation accuracy.

The WebNLG (Colin et al., 2016) corpus comprises of 25,298 (data, text) pairs and 9,674 sets of triplets(subject, property, object) describing facts (entities and relations between them) and the corresponding facts in form of natural language texts. The test set is split into two parts: on one hand, it contains DBpedia categories that were seen in the training data; and on the other hand, it consists of inputs from 5 unseen categories.

DART (Radev et al., 2020) is a large dataset for open-domain structured data record to text generation. It has a similar input format to WebNLG but richer and more diverse predicates than WebNLG. DART consists of 82,191 examples across different domains with hierarchical inputs based on a tree ontology that transforms a flat table into a tree

structure.

For T5 models, we investigate the robustness using question answering(QA) tasks. In our experiments, We train on the SQuAD (Rajpurkar et al., 2016) dataset, and test on the DuoRC (Saha et al., 2018) dataset. The evaluation metric for T5 is EM/F1 score, derived from the script provided by the MRQA challenge by Fisch et al. (2019). Later in this project, we may test the same model on more reading comprehension related datasets and report their evaluation metrics, as they are available from the MRQA challenge. We may also propose novel evaluation metrics that better demonstrate how a model has leaned toward an out-of-domain distribution.

SQuAD (Rajpurkar et al., 2016) is a reading comprehension dataset, containing 107,785 question-answer pairs. Questions in this dataset are posed by crowdworkers from Wikipedia articles, and the answer to every question is a segment of text from the corresponding reading passage, meaning the system will select the answer from all possible spans. Even though span-based answers are more constrained, SQuAD dataset still provides us with diverse questions and answer types.

DuoRC (Saha et al., 2018) is another dataset for reading comprehension dataset. DuoRC contains 186,089 (question,answer) pairs generated from a collection of 7680 pairs of movie plots. Every pair in the collection reflects two versions of the same movie since they are written by two different groups of crowdworkers. This makes the answers less overlapping, different in levels of plot details and higher requirements for reasoning process.

2.2 Methods & Hyperparameters

In our work, we will apply both prompt and prefix tuning on T5 and GPT-2 models. Our experimental design spans two dimensions for each model and tuning method. First, we measure the robustness of tuning with respect to different model sizes, given the same prompt length. Second, we measure the robustness of tuning with respect to different prompt lengths, given the same model size. We train both T5 and GPT-2 models with sizes range from small, base and large, and with prompt lengths from 1, 5, 10, 20 and 50. The prompts and prefixes are initialized from vocabulary.

For the T5 model, we followed one of the current de-facto ways of training the model. In particular, we trained it with AdaFactor with a learning rate of 0.001 and no scheduler. In terms of the optimizer, we disabled scaling the parameter and the relative step. We used a clip threshold of 1.0, and we did not have any warm up steps during training. We run 4 epochs through all the training data in our experiments. This applies to both prompt tuning and prefix tuning. We don't want to spend much time playing with the hyperparameters (since it's not for publication), and we hope that our setting will give a more realistic performance of the model.

For the GPT-2 model, we followed the optimized parameters provided by Prefix-tuning (Li and Liang, 2021). In particular, we trained it with AdamW optimizer (Loshchilov and Hutter, 2019) and a linear learning rate scheduler according to the HuggingFace default setup. The learning rate is $5 \cdot 10^{-5}$. At decoding time, we use beam search with a beam size of 5 for the DART dataset.

Configuration		**In-Domain**		**Out-of-Domain**	
Size	**# Tkns**	**EM**	**F1**	**EM**	**F1**
Small	1	−1	−1	−1	−1
	5	−1	−1	−1	−1
	10	−1	−1	−1	−1
	20	−1	−1	−1	−1
	50	−1	−1	−1	−1
Base	1	55.29	**79.84**	**30.71**	**49.74**
	5	47.70	72.44	18.79	36.13
	10	50.09	73.32	21.99	39.44
	20	**55.73**	75.95	25.98	42.38
	50	49.29	74.23	16.06	38.11
Large	1	−1	−1	−1	−1
	5	−1	−1	−1	−1
	10	−1	−1	−1	−1
	20	−1	−1	−1	−1
	50	−1	−1	−1	−1

Table 1: T5 results on question-answering task with prompt tuning. Here, SQuAD dataset was used as the training set to train the model. In-Domain evaluation metrics are reported based on the validation set of SQuAD dataset, while Out-of-Domain evaluation metrics are reported based on the test set of DuoRC dataset. All data came from the MRQA dataset bundle. Entries marked with **Run** indicates that the model is in training and the evaluation metrics will come out as soon as they finish training. Entries marked with −1 indicates that they are in queue.

3 Results

Although we largely haven't run through the experiments, we are getting some preliminary results on

prompt tuning with T5 models. Figure 1 reports the evaluation metrics of some T5 models trained on SQuAD dataset and evaluated on the DuoRC dataset. In the near future, we are expecting to release more results on 1. evaluation metrics of T5 models with different configurations tested on other out-of-domain datasets with prompt tuning in question-answering; 2. evaluation metrics of GPT2 models with different configurations tested on other out-of-domain datasets with prompt tuning in table-to-text generation.

4 Discussion

In this section, we will discuss the advantage of prefix/prompt tuning. We also want to address some limitations in this study.

4.1 Advantages

Prefix/prompt tuning will only train on a small subset of parameters and freeze other parameters, which significantly reduces training costs. Suppose we have many individual tasks but share the same model structure. In that case, prefix/prompt tuning could maintain modularity and save time/space by only adding and deleting prefix/prompt for each task. Beyond that, the inference is more efficient with prefix/prompt settings. Instead of having different models and calling forward multiple times, we can do a single forward pass with batches.

4.2 Limitations

Because of time limitations, we do not perform ablation tests to examine the internal representation of prefix/prompt tokens. However, this is another exciting topic we want to explore in the future. For example, if we could find some patterns in the space of prefix/prompt tokens, we could directly add a prefix/prompt to a pretrained model when a new task comes. This would allow us to obtain a model which has comparable performance to fine-tuned models, but with no extra costs.

5 Conclusion

We will make a conclusion once we get more experimental results.

Acknowledgements

This paper is a derivative of the data science capstone project at the University of California, San Diego. We thank our mentor Prof. Zhiting Hu for providing feedback and directions for the project. The computational resources are provided by DSMLP at the University of California, San Diego.

References

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Dragomir R. Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Nazneen Fatema Rajani, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Murori Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, and Richard Socher. 2020. DART: open-domain structured data record to text generation. *CoRR*, abs/2007.02871.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. *CoRR*, abs/1804.07927.

Banerjee Satanjeev and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with ...

Matthew Snover and Bonnie Dorr. 2006. A study of translation edit rate with targeted human annotation.

A Appendix

A.1 Links

Link to Project Proposal

A.2 Training Loss

The below figure shows the training loss on T5 with prompt tuning for base models: