# Pre-Trained Prompt-Tuning Based on Adversarial Regularization for Text Classification

Xiaoying Huang[1], Baihui Tang[2] and Sanxing Cao[3]
[1]*Communication and Information System, Communication University of China, Beijing, China*
[2]*New Media Research Institute, Communication University of China, Beijing, China*
[3]*Internet Information Research Institute, Communication University of China, Beijing, China*

Keywords: Prompt-Tuning, Adversarial Regularization, Text Classification.

Abstract: The advent of large-scale pre-trained models has greatly promoted the development of natural language processing. Many natural language processing tasks choose to fit the gap between downstream tasks and pre-training tasks through fine-tuning. However, the existing pre-trained model has a large number of parameters, and it also needs a lot of data to fine-tuning. To adapt to the training of large-scale pre-trained models, researchers proposed to replace fine-tuning with prompt-tuning to reduce the demand for supervised data. However, the performance of prompt-tuning is not stable enough. This paper proposes a method of adding adversarial regularization training based on prompt-tuning, adding disturbance in word embedding, and continuously updating the disturbance in a small range, to increase the robustness of the model and make the model obtain higher accuracy under less supervised data.

## 1 INTRODUCTION

The success of text classification technology depends on a large number of supervised data. However, obtaining a large amount of marker data is very expensive. To solve this problem, researchers have proposed transfer learning.

The goal of transfer learning is to apply the knowledge learned in a certain field to different but related fields. It consists of two stages: The first stage is the pre-training stage, which trains a high-capacity model for high resource-related tasks outside the target domain, that is, the pre-trained model. The second stage is fine-tuning, which uses the target task supervised data to fine tune the parameters of the pre-trained model so that the pre-trained model can adapt to the target task. Wudao2.0, the largest language model to date, has about 1750 billion parameters. Table 1-1 shows the comparison of pre-trained model parameters. Because the pre-trained model is trained by a large corpus, it has strong generalization. In the training of downstream tasks, the weight of the pre-trained model is extracted for the initialization of the downstream task model, which can make the model fit faster and better. This fine-tuning approach solves the problem of insufficient resources in the target domain and achieve the best results in many NLP tasks. (Devlin et al., 2019)

Table1-1: Comparison of parameters of each pre-trained model.

| Model | Parameters |
|---|---|
| Bert-base | 110M |
| Bert-large | 335M |
| Roberta-large | 355M |
| T5 | 110000M |
| GPT-3 | 1750000M |
| WuDao2.0 | 17500000M |

However, due to the high complexity of large-scale pre-trained models and the limited supervised data from downstream tasks, when the supervised data cannot meet the model fine-tuning requirements, the problem of over-fitting will occur. At this time, the generalization ability of the model will decline and the performance of the data outside the training set will be poor. In order to solve the problem that limited supervised data can not meet the fine-tuning requirements, researchers changed from traditional fine-tuning to prompt-tuning.

Prompts can be divided into the hard prompt and the soft prompt. The hard prompt is to set a pair of manual prompts and verbalizers. By setting the

Huang, X., Tang, B. and Cao, S.
Pre-Trained Prompt-Tuning Based on Adversarial Regularization for Text Classification.
DOI: 10.5220/0011922200003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 286-291
ISBN: 978-989-758-622-4; ISSN: 2975-9463

mapping relationship between prompt and verbalizer, the difference between pre-training tasks and downstream tasks can be reduced by predicting [mask]. LAMA (Petroni et al., 2019) proposed using cloze to obtain the knowledge of the pre-trained model without fine-tuning.

Soft prompt replaces fine-tuning by training a new word vector and achieves the effect comparable to fine-tuning with greatly reduced calculation. (Xiao Liu et al., 2021) proposed to enhance the natural language understanding ability of the pre-trained model by automatically searching for a better prompt in the semantic space. (Xiang Lisa Li et al., 2021) proposed to replace fine-tuning with prompt and prefix adjustment. Only 0.1% of the parameters need to be trained, to get performance comparable to fine-tuning. (Xiao Liu et al., 2021) further applied prompt-tuning to complex natural language understanding tasks.

On the other hand, some researchers have added adversarial training in fine-tuning to improve the robustness of the model. (Chen Zhu et al., 2019) proposed to improve the robustness of the model by adding adversarial disturbance in word embedding and minimizing the adversarial risk generated in different areas around the input sample. (Haoming Jiang et al., 2020) proposed a framework of smooth regularization and Bregman near point optimization to prevent radical updates during model adversarial training.

In this paper, we propose a training method based on prompt-tuning with adversarial regularization:

1. An adversarial regularization algorithm is proposed. Add disturbance in word embedding, and increase the robustness of the model by updating the disturbance in a small range, so that the model can obtain higher accuracy under fewer supervision data:

2. In the process of prompt tuning, adversarial regularization is organically incorporated to improve the robustness of the model.

## 2 RELATED WORK

With the advent of large-scale pre-trained models, deep learning has rapidly moved closer to large-scale pre-trained models, changing the traditional mode of deep learning and becoming a new benchmark for various deep learning tasks. The more model parameters, the more knowledge learned, the better generalization ability, and the better performance in the training of downstream tasks.

On the other hand, the large-scale pre-trained model is extremely complex, and the limited supervised data can not pry hundreds of millions of parameters, resulting in poor transferability of the large-scale pre-trained model. In order to solve this problem, researchers propose a manual prompt, that is, to design a prompt template for task data so that the downstream task is as close as possible to the pre-trained task. This method greatly reduces the requirements of downstream tasks on the amount of supervised data and allows small parameters to pry the large model. (Shengding Hu et al., 2021) proposed using an external knowledge base to expand the mapped tag language space, which greatly improves the accuracy of short text classification. However, this method relies heavily on prompt template and validation set data, and its performance is not stable. LAMA (Petroni et al., 2019) showed the cases in the following Table 2-1 in the knowledge inquiry. It can be seen that the change of one of the words will lead to a huge difference in the results.

Table 2-1: Effect of different prompt templates on accuracy.

| Prompt | Accuracy |
|---|---|
| [X] is located in [Y]. | 31.29 |
| [X] is located in which country or state? [Y]. | 19.37 |
| [X] is located in which country? [Y]. | 31.40 |
| [X] is located in which country? In [Y] | 51.08 |

After the discrete manual prompt, researchers have proposed a continuous automatic prompt, that is, freezing the parameters of the pre-trained model and only fine-tuning the continuous prompt. (Xiang Lisa Li et al., 2021) proposed that by adding prefixes, 0.1% of the parameters can be trained to obtain performance matching with fine-tuning, which proved that GPT was equally excellent in natural language understanding tasks. (Brian Lester et al. 2021) proposed to add trainable continuous embedding (also known as continuous prompts) to word embedding in the original sequence, freeze the pre-trained model parameters during training, and only update the continuous prompts to complete downstream tasks. With the continuous development of prompt-tuning, it has achieved the same effect as fine-tuning. (Yuxian Gu et al., 2021) added prompt in the pre-training stage to pre-train the prompt, so as to obtain better initialization of prompt in the downstream task and achieve better performance than fine-tuning in the classification task. (Brian Lester et al., 2021) pointed out that the effect of prompt tuning is positively correlated with the size of the pre-trained

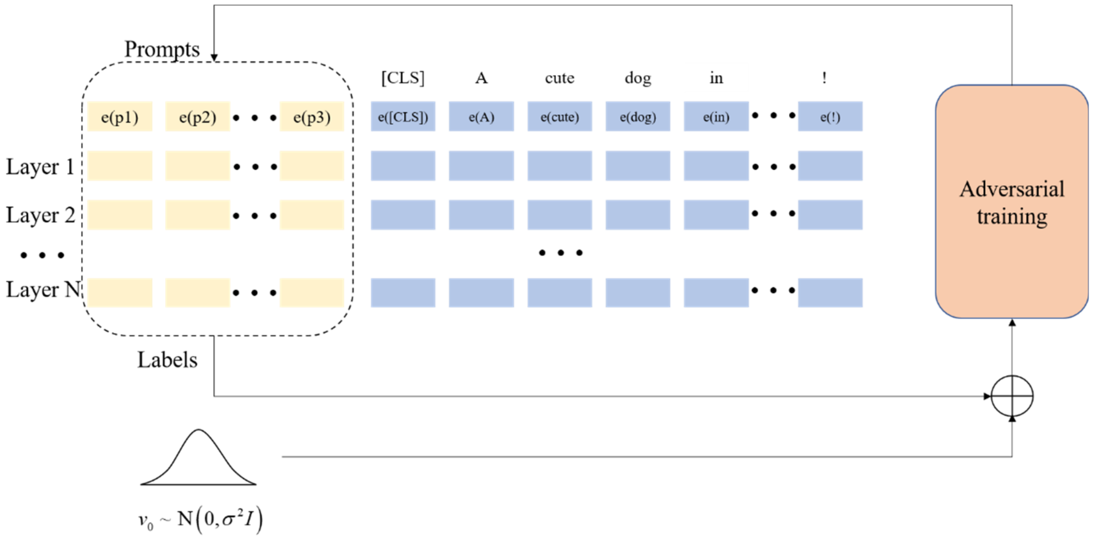

Figure 2-1: Model structure.

model and the robustness of prompt-tuning in domain transfer is better than fine-tuning. (Xiao Liu et al., 2021) proposed a multi-task training strategy to further improve the performance of prompt-tuning. However, this strategy could not work in the scenario with limited supervised data.

To make the model get better performance under the limited supervised data, this paper adds adversarial regularization to the prompt-tuning, so that it can obtain the performance of full data fine-tuning with a small amount of data training. We named this “Robust Prompt-tuning” method “RPT”.

# 3 ROBUST PROMPT-TUNING

## 3.1 Structural Design

The model structure proposed in this paper is shown in Figure 2-1.

Given a pre-trained language model M, use [CLS] to obtain the eigenvectors of the pre training model M. A discrete input token sequence $\mathrm{X}_{0:n} = \{x_0, x_1, \dots, x_n\}$ will be mapped to the input embedding $E_x = \{E_{x_0}, E_{x_1}, \dots, E_{x_n}\}$, concatenate $\mathrm{E_x}$ with $\mathrm{E_y}$. During training, Freeze the parameters of the pre-trained model M, and add adversarial regularization to the training of $\mathrm{E_y}$ to reduce the demand for data volume. The goal is to obtain the classification result of the input X through M.

## 3.2 Adversarial Regularization Training

As shown in algorithm 1, adversarial regularization is added in the training. Random noise $v_0$ is added to each input $x_i$ during training to obtain $\tilde{x}_i$, and the noise $v_0$ obeys normal distribution. The $\tilde{x}_i$ and the tag $y$ of the $x_i$ are fed into the model to obtain the cross-entropy loss and calculate its gradient $g_t$ to the noise $v_t$. The product of step size $\lambda$ and gradient $g_t$ is added to $v_t$, then the L2 norm is normalized to update $v_t$. Repeat this for $K$ times. Finally, the cross-entropy loss of the adversarial training is obtained, weighted and added it with the cross-entropy loss by feeding $x_i$ and its tag $y$ into the model to update the model parameters. In short, when the model learns that “you look beautiful” indicates positive emotion, it adds a small disturbance to the semantic space of “you look beautiful” to obtain “you look nice”, so that its prediction label is closer to the label of “you look beautiful”. The model increases the robustness of the model by resisting the continuous expansion of the positive/negative emotion semantic space after regularization.

Algorithm 1: Robust Prompt-tuning (RPT).

```
Require: Training samples $X = \{(x_i, y)\}$ ,
    adversarial rate $b$ , adversarial step size $l$
Initialize $q$
For epoch = 1, ..., N do
    For s = 1, ..., S do
```

```
        Sample a minibatch $B$ Ì $X$
          $L(f_q(x_i),y)$ ® $q_s$
          $v_0 \sim \mathrm{N}\left(0, s^2 I\right)$
          For t = 1, …, K do
            $Ñ_{v_{t-1}} L(f_q(x_i + v_{t-1}),y)$ ® $g_t$
            $\frac{l g_t + v_{t-1}}{\left\| l g_t + v_{t-1} \right\| + e}$ ® $v_t$
          End for
          $L(f_q(x_i + v_K),y)$ ® $q_{adv}$
          $A\,dam\,Update_{\mathrm{B}}\left(q_s + b q_{adv}\right)$ ® $q_{s+1}$
      End for
  End for
```

# 4 EXPERIMENT

## 4.1 Dataset

The corpus adopts the emotion binary classification data set SST-2 in the public data set GLUE(Alex Wang et al., 2018), which is composed of 9612 sentences like Table 4-1:

Table 4-1: SST-2 composition.

| | |
|---|---|
| positive | 4649 |
| negative | 4963 |

All experimental data will be used in the first part of the experiment. The proportion of training set, verification set and test set is shown in Table 4-2 below.

Table 4-2: SST-2 full training data.

| | |
|---|---|
| Training set | 6920 |
| Validation set | 872 |
| Test set | 1820 |

In the second part of the experiment, the data of the training set was randomly reduced to 2000. The proportion of the training set, the verification set and the test set is shown in Table 4-3.

Table 4-3: SST-2 Small amount of training data.

| | |
|---|---|
| Training set | 2000 |
| Validation set | 872 |
| Test set | 1820 |

## 4.2 Models

As follows, the model proposed in this paper will be compared with other models.

Bert (Jacob Devlin et al., 2018): This is the Bert-base model, which is stacked by the bidirectional encoders of transformers, and has renewed natural language processing records on GLEU (Alex Wang et al., 2018), SQuAD (Pranav Rajpurkar et al., 2016), RACE (Guokun Lai et al., 2017) and XNLI (Alexis Conneau et al., 2018).

Bert FT: perform fine-tuning on the Bert model.

Bert PT: perform prompt-tuning on the Bert model.

Bert RPT: This is the model proposed in this paper. It is based on Bert pre-trained model and prompt-training, and on this basis, adversarial regularization is added.

Roberta (Liu et al., 2019): This is the Roberta-large model. Based on Bert, this model improves the static mask in Bert into a dynamic mask, cancels the training task of next sentence prediction, uses more training data, and achieves better performance than Bert in natural language processing tasks such as GLEU (Alex Wang et al., 2018), SQuAD (Pranav Rajpurkar et al., 2016) and RACE (Guokun Lai et al., 2017).

Roberta FT: perform fine-tuning on the Roberta large model.

Roberta PT: perform prompt-tuning on the Roberta large model.

Roberta RPT: This is the model proposed in this paper. It is based on the Roberta-large pre-trained model and prompt-training, and on this basis, the adversarial regularization is added.

## 4.3 Experiment Setting

The experiments mainly use the PyTorch tool and the hugging face platform. The experimental data use the SST2 fully supervised data set. In experiments, the learning rate of e-5 and the size of the batch are 64. The pre-trained model is trained with 20 epochs. The length of the prompt is set to 2.

### 4.4 Results

Table 4-4: Accuracy(ACC) of Bert and Roberta large models with fine-tuning (FT), prompt-tuning (PT), and robust prompt-tuning (RPT) methods under full data and a small amount of data.

| MODEL | FULL SST2 ACC | 2K SST2 ACC |
|---|---|---|
| **BERT FT** | 89.45 | 87.89 |
| **BERT PT** | 89.51 | 87.83 |
| **BERT RPT** | 90.01 | 88.28 |
| **ROBERTA-LARGE FT** | 93.52 | 91.96 |
| **ROBERTA-LARGE PT** | 94.20 | 93.42 |
| **ROBERTA-LARGE RPT** | 94.36 | 93.92 |

It can be seen from Table 4-4 that Bert and Roberta-large models have achieved higher accuracy under the RPT method compared with FT and PT methods under full data. For the Bert model, RPT improved the accuracy by +0.56 compared with FT and improved the accuracy by + 0.5 compared with PT. For the Roberta-large model, RPT improved the accuracy by + 0.84 compared with FT and improved the accuracy by + 0.16 compared with PT. It can be seen that the larger scale of the pre-trained model, the better effect of RPT.

For the Bert model under a small amount of data, RPT improved the accuracy by +0.39 compared with FT and improved the accuracy by + 0.45 compared with PT. For the Roberta-large model under a small amount of data, RPT improved the accuracy by + 1.96 compared with FT and improved the accuracy by + 0.5 compared with PT. It can be seen that when training with a small amount of data, the large model contains more pre-training knowledge than the small model, and the effect of using the RPT is better.

When Roberta-large model training with a small amount of data, the accuracy of the model decreases by 1.56 by FT, but only decreases by 0.44 by RPT. It can be seen that the RPT method requires less supervised data than the FT method.

As shown in Figure 4-1, compared with the FT method, the performance of the model optimized by the RPT method has been further improved whether it is full data or a small amount of data. The performance improvement is more obvious in the large model.

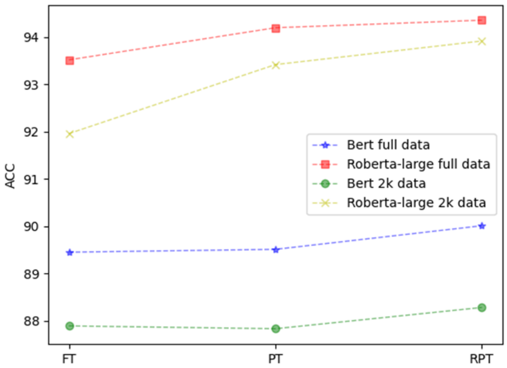

Figure 4-1: Results.

## 5 CONCLUSIONS

In this paper, we propose a method to replace fine-tuning with adversarial prompt-tuning. It improves the understanding ability and robustness of the language model by automatically searching for a suitable prompt in the continuous space and adding noise disturbance to the semantic space of the prompt. Compared with the fine-tuning method, this model relies less on large-scale data sets. In the public data set of SST-2, the adversarial prompt reduces the amount of computation and improves accuracy. Under the condition of a small amount of training data, this method is obviously superior to the fine-tuning method.

This method is only verified in the classification task for the time being. For other NLP tasks and other pre-trained models, the verification will continue in the future.

## REFERENCES

Moore, R., Lopes, J., 1999. Paper templates. In *TEMPLATE'06, 1st International Conference on Template Production*. SCITEPRESS.

Smith, J., 1998. *The book*, The publishing company. London, 2$^{nd}$ edition.

Jeremy Howard, & Sebastian Ruder (2018). Fine-tuned Language Models for Text Classification.. arXiv: Computation and Language.

Fabio Petroni, Tim Rocktäschel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, & Sebastian Riedel (2019). Language Models as Knowledge Bases empirical methods in natural language processing.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, & Jie Tang (2021). GPT Understands, Too arXiv: Computation and Language.

Xiang Lisa Li, & Percy Liang (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation meeting of the association for computational linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, & Jie Tang (2021). P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks arXiv: Computation and Language.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, & Jingjing Liu (2019). FreeLB: Enhanced Adversarial Training for Language Understanding.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, & Tuo Zhao (2020). SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization meeting of the association for computational linguistics.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, & Maosong Sun (2021). Knowledgeable Prompt-tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification arXiv: Computation and Language.

Fabio Petroni et al. “Language Models as Knowledge Bases” empirical methods in natural language processing (2019): n. pag.

Brian Lester, Rami Al-Rfou, & Noah Constant (2021). The Power of Scale for Parameter-Efficient Prompt Tuning arXiv: Computation and Language.

Yuxian Gu, Xu Han, Zhiyuan Liu, & Minlie Huang (2021). PPT: Pre-trained Prompt Tuning for Few-shot Learning arXiv: Computation and Language.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, & Samuel R. Bowman (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding Learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding north american chapter of the association for computational linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, & Percy Liang (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text empirical methods in natural language processing.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, & Eduard Hovy (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations empirical methods in natural language processing.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, & Veselin Stoyanov (2018). XNLI: Evaluating Cross-lingual Sentence Representations empirical methods in natural language processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Michael Lewis, Luke Zettlemoyer, & Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach arXiv: Computation and Language.