AF Adapter: Continual Pretraining for Building Chinese Biomedical Language Model

Yongyu Yan1, Kui Xue2, Xiaoming Shi2, Qi Ye1,*, Jingping Liu1, Tong Ruan1,*

1 School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China
2 Shanghai Artificial Intelligence Laboratory, Shanghai, China

Abstract—Continual pretraining is a popular way of building a domain-specific pretrained language model from a general-domain language model. In spite of its high efficiency, continual pretraining suffers from catastrophic forgetting, which may harm the model’s performance in downstream tasks. To alleviate the issue, in this paper, we propose a continual pretraining method for the BERT-based model, named Attention-FFN Adapter. Its main idea is to introduce a small number of attention heads and hidden units inside each self-attention layer and feed-forward network. Furthermore, we train a domain-specific language model named AF Adapter based RoBERTa for the Chinese biomedical domain. In experiments, models are applied to downstream tasks for evaluation. The results demonstrate that with only about 17% of model parameters trained, AF Adapter achieves 0.6%, 2% gain in performance on average, compared to strong baselines. Further experimental results show that our method alleviates the catastrophic forgetting problem by 11% compared to the fine-tuning method. Code is available at https://github.com/yanyongyu/AF-Adapter.

Index Terms—Continual pretraining, Chinese biomedical natural language processing, Adapter tuning

I. INTRODUCTION

Currently, a vast volume of Chinese biomedical literature emerges, including medical diagnosis dialogues on Chinese medical communities and medical knowledge in Chinese encyclopedias. As an example, DXY1 (a Chinese medical community) contains millions of medical dialogues between 5.5 million patient users and 2 million doctor users, while Baidu Encyclopedia2 (a Chinese encyclopedia) includes over 30 million articles, covering a wide range of topics related to health, disease, and medical treatments. As a result, there is a growing demand for precise Chinese biomedical text mining tools to effectively extract valuable information from Chinese medical literature.

Natural language processing (NLP) methods greatly improve the automatic text mining from Chinese literature accurately. Among these algorithms, recently, pretrained language models (PLMs) [1] [2] have become increasingly popular, as they demonstrate impressive performance. Specifically, PLMs are trained on a vast amount of text data to learn better representations of natural language, providing a strong foundation for downstream NLP tasks.

However, directly applying PLMs on general domains to biomedical text mining suffers from the word distribution bias between general and biomedical texts, which harms performances. To alleviate the issue, PLMs in the biomedical domain are introduced. Methods of PLMs on specific domains are divided into two ways, pretraining from scratch [3] and continual pretraining [4] based on general-domain language models. In this work, a continual pretraining manner is adopted.

In spite of the high training efficiency, continual pretraining suffers from catastrophic forgetting [5] [6] [7]. Catastrophic forgetting means the model trained on new corpora tends to forget the knowledge of previous data. To alleviate the issue, there are three main approaches. The first one is training-based methods, including discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing [8]. The second one is parameter reserve [9], which keeps randomly selected pretrained parameters. The third one is adapter-based tuning methods [10], which inserts additional layers after specific layers of pretrained models. Adapter-based tuning methods better mitigate the forgetting problem than other methods, thus becoming popular recently.

Despite the good performance, the adding-layer manner of the current adapter-based tuning methods adds depth to the networks. The input needs to be fed forward to more layers and thus is more likely to be forgotten. To alleviate the issue, we propose a layer-extending continual pretraining method, named Attention-FFN Adapter (AF Adapter). The main contributions of this paper are summarized as follows:

- We propose a layer-extending continual pretraining method called AF Adapter, which aims to extend attention heads and hidden units for the BERT-based model. The method further alleviates catastrophic forgetting compared with layer-adding methods.
- We propose a Chinese biomedical-domain pretrained language model called AF Adapter based RoBERTa, which is trained using AF Adapter and biomedical corpora. The model contributes greatly to downstream medical tasks.
- Our method achieves 0.6%, 2% gain in performance on average, compared to strong baselines on the CBLUE benchmark. Besides, AF Adapter alleviates the catastrophic forgetting problem by 11% compared to the fine-tuning method.

*Corresponding author.
1https://portal.dxy.cn/
2https://baike.baidu.com/
II. PRELIMINARIES

In this section, we provide a brief overview of the existing approaches in the fields of domain-specific pretraining and two key components in the BERT architecture: the multi-head self-attention mechanism and the feed-forward network.

A. Domain-Specific Pretraining

General-domain pretrained language models face difficulties in specialized fields such as biomedicine. The word distributions differ significantly between general-domain text and biomedical-domain text, including the presence of biomedical terms. To address these challenges, researchers have proposed studies focusing on biomedical domain-specific pretraining, which can be categorized into two main types.

Domain-specific pretraining from scratch is based on large-scale biomedical corpora. One notable advantage of this approach is the ability to customize the model’s vocabulary to the domain. This allows for more appropriate tokenization of medical terms, avoiding fragmented subword representations.

Continual pretraining of a general-domain pretrained model is a common approach. This approach involves initializing with a standard model pretrained on general-domain corpora and then continuing the pretraining process using biomedical corpora. Continual pretraining benefits from the knowledge acquired during previous pretraining.

B. BERT

The BERT model architecture is built upon a multi-layer bidirectional Transformer encoder structure [11]. Each encoder consists of a stack of identical blocks that incorporate multi-head self-attention and feed-forward networks. The multi-head self-attention layer can be formulated as follows:

\[
Q(x) = xW_Q + b_Q, \quad K(x) = xW_K + b_K, \quad V(x) = xW_V + b_V, \quad (1)
\]

where \(W_Q \in \mathbb{R}^{d_{model} \times h d_k}, W_K \in \mathbb{R}^{d_{model} \times h d_k}, W_V \in \mathbb{R}^{d_{model} \times h d_v}, b_Q \in \mathbb{R}^{h d_k}, b_K \in \mathbb{R}^{h d_k} \) and \(b_V \in \mathbb{R}^{h d_v} \) represent the weight matrices and biases. \(d_k = d_v = \frac{d_{model}}{h} \) is the model dimension, and \(h \) is the number of attention heads.

The \(Q, K, \) and \(V \) are then split into \(h \) parts to calculate the multi-head attention. This can be expressed as:

\[
head_j = \text{Attention}(Q_j, K_j, V_j), \quad \text{MultiHead}(x) = [head_1, \ldots, head_h]W_O, \quad (2)
\]

where \(Q = [Q_1, \ldots, Q_h], K = [K_1, \ldots, K_h], V = [V_1, \ldots, V_h] \) and \(W_O \in \mathbb{R}^{hd_k \times d_{model}} \). The operation “:” stands for the column-wise concatenation.

After the multi-head attention, the feed-forward network is employed, consisting of two linear transformations with a GeLU activation function in between:

\[
\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2, \quad (3)
\]

where \(W_1 \in \mathbb{R}^{d_{model} \times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff} \times d_{model}}, b_1 \in \mathbb{R}^{d_{ff}} \) and \(b_2 \in \mathbb{R}^{d_{model}} \) represent the weight matrices and biases applied to the input \(x \) in the feed-forward network.

III. METHODOLOGY

In this section, AF Adapter is introduced to train domain-specific PLMs based on general-domain PLMs. The overall architecture of AF Adapter is illustrated in Fig. 1.

A. Model Architecture

Following adapter tuning [10] [12], additional trainable domain-specific parameters are inserted into the extension of the attention and FFN layer, while parameters in the original model are fixed to preserve the knowledge of the general domain. The operations “:” and “:” used below stand for the column- and row-wise concatenation, respectively.

Extended Attention Layer. A key component of BERT is the multi-head scaled dot-product self-attention mechanism. In this work, the self-attention layer is extended to alleviate catastrophic forgetting. Specifically, a number of “domain-specific heads” in each attention layer are added. Note that added domain-specific heads in different layers are independent of each other. To extend the self-attention mechanism, \(i \) additional attention heads are added. Formally, the input into the multi-head self-attention layer is denoted as \(x \), \(x \) is firstly fed into linear layers, and mapped into \(Q', K', V' \). The extended weight matrices are denoted as \([W_Q : W_Q'], [W_K : W_K'], [W_V : W_V'] \). \(W_Q, W_K, \) and \(W_V \) are matrices...
from the original pretrained model, as described in Equation 1. Then, \( Q'(x) \in \mathbb{R}^{d_{model} \times (h+i)d_k} \), \( K'(x) \in \mathbb{R}^{d_{model} \times (h+i)d_k} \), \( V'(x) \in \mathbb{R}^{d_{model} \times (h+i)d_k} \) are obtained,
\[
\begin{align*}
Q'(x) &= [W_Q; W_Q'] + [b_Q; b_Q'], \\
K'(x) &= [W_K; W_K'] + [b_K; b_K'], \\
V'(x) &= [W_V; W_V'] + [b_V; b_V'],
\end{align*}
\]
where the projections are parameter matrices \( W_Q' \in \mathbb{R}^{d_{model} \times d_k} \), \( W_K' \in \mathbb{R}^{d_{model} \times d_k} \), \( W_V' \in \mathbb{R}^{d_{model} \times d_k} \), \( b_Q' \in \mathbb{R}^{d_k} \) and \( b_V' \in \mathbb{R}^{d_k} \). \( d_k, d_v, d_{model} \), \( h \) are hyperparameters described in Equation 1.

Then, the \( Q' \), \( K' \) and \( V' \) are divided into \( h+i \) parts for calculating the multi-head attention,
\[
\text{head}_j = \text{Attention}(Q'_j, K'_j, V'_j),
\]
where \( Q' = [Q_1; \cdots; Q_h; Q'_h; \cdots; Q'_i] \), \( K' = [K_1; \cdots; K_h; K'_h; \cdots; K'_i] \), \( V' = [V_1; \cdots; V_h; V'_h; \cdots; V'_i] \).

Finally, the outputs of these heads are concatenated and then fed to linear transformations,
\[
\text{MultiHead}(x) = [\text{head}_1; \ldots; \text{head}_h; \text{head}_h'; \ldots; \text{head}_i'][W_O; W_O'],
\]
where \( W_O \in \mathbb{R}^{(d_k \times i) \times d_{model}} \), and \( \text{head}_j \) is the \( j \)th head described in Equation 2.

**Extended Feed-Forward Network.** For FFN in the model, a multi-layer perceptron with one hidden layer, we add a number of hidden units in each FFN layer with GelU activation between two linear transformations. Formally, we add “domain-specific hidden units” of size \( a \) by extending \( W_1 \) and \( W_2 \) to \([W_1; W'_1] \) and \([W_2; W'_2] \),
\[
\text{FFN}(x) = \text{GeLU}(x[W_1; W'_1] + [b_1; b'_1])(W_2; W'_2] + [b_2; b'_2],
\]
where \( W_1' \in \mathbb{R}^{d_{model} \times a} \), \( b_1' \in \mathbb{R}^a \), \( W_2' \in \mathbb{R}^{a \times d_{model}} \), and \( b_2' \in \mathbb{R}^{d_{model}} \). \( W_1, W_2, b_1 \) and \( b_2 \) are matrices described in Equation 3.

**B. Pretraining Details**

**Pretraining Tasks.** To continually pretraining the BERT-based model, the masked language modeling (MLM) is utilized to pretrain the model. In MLM, a subset of input tokens is randomly replaced with a special token (e.g., `[MASK]`), and MLM is designed to predict these tokens. The training objective is the cross-entropy loss between the original tokens and the predicted ones.

**Whole Word Masking [13].** In the original BERT, the text is split into subwords and tokenized into tokens. The whole word masking mitigates the drawback of masking only a part of the whole word. For the whole word masking in Chinese, the traditional Chinese word segmentation tool “Jieba” is utilized to split the sentence into several words.

**Training Strategy.** To alleviate the catastrophic forgetting problem [5], we first make progress in the model architecture and only train the domain-specific parameters added in the self-attention and FFN \((W_O, W_K', W_V', W_O', W_V', W_K')\). All parameters inherited from the general-domain model are frozen.

| Corpora Type | # of Sentences | # of Tokens |
|--------------|----------------|-------------|
| Biomedical Question Answering | 4,842k | 93M |
| Medical Encyclopedia | 793k | 33M |
| Electronic Medical Record | 1,762k | 45M |

| Task | Type | Train | Dev | Test | Metric |
|------|------|------|-----|------|--------|
| CMeEE | NER | 15,000 | 5,000 | 3,000 | Micro-F1 |
| CMeIE | RE | 14,339 | 3,585 | 4,482 | Micro-F1 |
| CHIP-CDN | NORM | 6,000 | 2,000 | 10,000 | Micro-F1 |
| CHIP-STS | STS | 16,000 | 4,000 | 10,000 | Micro-F1 |
| CHIP-CTC | TC | 22,962 | 7,682 | 10,192 | Micro-F1 |
| KUAKE-QIC | TC | 6,931 | 1,955 | 1,994 | Acc |
| KUAKE-QTR | NLI | 24,174 | 2,913 | 5,465 | Acc |
| KUAKE-QQR | NLI | 15,000 | 1,599 | 1,596 | Acc |

**IV. Experiments**

In this section, we provide the details of our experimental setup and present the results obtained from the evaluation. We pretrain the domain-specific models using collected datasets and then evaluate them on downstream tasks. Furthermore, we conduct an ablation study on different pretraining techniques and investigate convergence and stability. Additionally, we perform a catastrophic forgetting analysis to evaluate the models’ ability to retain previously learned knowledge.

**A. Pretraining Data and Settings**

**Datasets.** To evaluate the performance of our method, we collect a variety of Chinese biomedical corpora. These corpora include Chinese biomedical question answering, medical encyclopedia from Baidu Encyclopedia and Wikipedia, and electronic medical records from Fudan University Shanghai Cancer Center. Details of the pretraining corpora used for our experiments are presented in Table I.

**Backbone.** We perform experiments based on the Chinese RoBERTa-wwm-ext-base model [13], add one additional attention head to each layer and increase the intermediate size of the feed-forward network by 1,024.

**Pretraining Settings.** We pretrain our model using the original vocabulary provided with the Chinese RoBERTa model. We use the AdamW optimizer with warm-up and weight decay. Specifically, we begin with a learning rate of zero, which increases linearly to a peak rate of \( 4 \times 10^{-4} \) over the first 1000 steps of training. We train the model for 100,000 steps using a total batch size of 512, spread across two NVIDIA A100 (80G) GPUs with a batch size of 64 and gradient accumulation steps of 4.

**Baselines.** The baseline models used in our experiment are of similar sizes and have been widely used. For comparison of PLMs, the models include BERT-base [1], BERT-wwm-ext-base [13], RoBERTa-wwm-ext-base [13], PCL-MedBERT [3], and MacBERT-base [14]. For comparison of pretraining techniques, the baselines include Fine-Tuning, FL-Tuning [12],

3https://www.ihub.org.cn/html/2020/news_0824/47.html
and LoRA [15]. We compare the performance of our model against these baselines to evaluate its effectiveness.

B. Evaluation Tasks

We compare models by applying them to the downstream NLP tasks, specifically the Chinese Biomedical Language Understanding Evaluation (CBLUE) [16] benchmark as shown in Table II. The evaluation process is similar to the CBLUE benchmark toolkit [16], and we leverage the hyperparameters provided by the CBLUE RoBERTa baseline.

C. Main Results

In this section, we present the comparative analysis of different PLMs and pretraining techniques.

1) Comparison of PLMs: For comparison, we use the CBLUE’s public results of baselines. Table III shows the performances of Chinese NLP tasks of the CBLUE benchmark.

For the NER task, our AF Adapter based RoBERTa achieves a score of 62.984%, outperforming the other models. In the RE task, we observe that AF Adapter based RoBERTa outperforms the backbone model by about 1.2%, surpassing the other models as well. This highlights its effectiveness in extracting biomedical entities and semantic relationships of a certain type, demonstrating its superior performance in biomedical information extraction. Regarding the NORM, TS, and TC tasks, AF Adapter based RoBERTa achieves similar performance compared to the backbone model. Notably, our model outperforms the other models in the QIC task, showcasing its advantage in accurately classifying medical intents. In the NLI tasks, our model achieves the best performance, surpassing the backbone model by approximately 1.9%.

Overall, our AF Adapter based RoBERTa demonstrates strong performance across multiple task types and performs the best among models on average. It outperforms the backbone model RoBERTa-wwm-ext-base and medically domain-specific model PCL-MedBERT by 0.6% and 2%, respectively. While achieving similar performance in NORM, TS, and TC tasks, our model stands out in the NER, RE, and NLI tasks. These results highlight the versatility and effectiveness of our AF Adapter based RoBERTa in addressing a wide range of biomedical language understanding challenges.

2) Comparison of Pretraining Techniques: To investigate the impact of pretraining techniques on domain-specific models, we continually pretrain the model using Fine-Tuning and other methods with the same corpus. To ensure a fair comparison, we adjust the hyperparameters of the techniques, excluding Fine-Tuning, to achieve a similar size of trainable parameters. The evaluation results of the pretraining techniques are presented in Table IV.

The experimental results highlight the superior performance of the pretrained model using AF Adapter. Compared to the Fine-Tuning approach, AF Adapter achieves remarkable results while training only about 17% of the model parameters. By adding additional parameters, AF Adapter consistently outperforms LoRA by an average margin of 5.5%. Notably, our experiments reveal that the Fine-Tuning method leads to suboptimal performance. These findings suggest that continuing pretraining without considering domain-specific knowledge and representations may fail to effectively balance domain-specific information with general-domain representations.

D. Detailed Analysis

Convergence and Stability Analysis. We examine the convergence and stability of pretraining techniques throughout the pretraining process and present the results in Fig. 2.

From the results, we observe that our AF Adapter exhibits a significant decrease in loss compared to LoRA. This slower convergence indicates that the model requires more iterations to reach an optimal solution. Alongside convergence, our AF Adapter demonstrates consistent and stable performance, maintaining a smooth and gradual decrease in loss. In contrast, FL-Tuning and Fine-Tuning exhibit more fluctuations and variations in loss values. This stability analysis further supports the robustness of our AF Adapter approach in learning domain-specific information.

Catastrophic Forgetting Analysis. To assess the forgetting problem caused by sequential task training, we randomly
choose 10k general-domain samples from the “WuDao” corpus [17]. These samples are preprocessed using a masking rate of 15%, ensuring consistent inputs across all compared models. We evaluate the accuracy between Chinese RoBERTa-wmm-ext-base, Fine-Tuning based, FL-Tuning based, LoRA based, and AF Adapter based RoBERTa trained in Section IV-C2. The experimental result, presented in Table V, demonstrates the models’ performance in terms of catastrophic forgetting. Our model achieves only -5.324% accuracy reduction compared to the original RoBERTa. Additionally, it is evident that AF Adapter based RoBERTa outperforms Fine-Tuning based RoBERTa by a significant margin, achieving an accuracy of 82.428 compared to 70.964. This indicates that our model better retains previously learned knowledge when trained on new tasks. The model trained using fine-tuning exhibits a substantial drop in accuracy, indicating a significant forgetting of general knowledge. This aligns with its overall poor performance in CBLUE, as shown in Table IV.

V. CONCLUSION

In this paper, we propose a continual pretraining approach, named AF Adapter, for building domain-specific pretrained language models. By incorporating additional attention heads and hidden units within the BERT-based model, AF Adapter enables effective learning of domain-specific knowledge while leveraging the general language representations. We also present a Chinese biomedical-domain pretrained language model called AF Adapter based RoBERTa, which is trained using AF Adapter and biomedical corpora. The evaluation of AF Adapter based RoBERTa on the CBLUE benchmark shows its superior performance, outperforming other models of “similar size” by an average of 0.98%. Moreover, our analysis reveals that AF Adapter exhibits robust convergence and stability compared to other pretraining techniques. These findings highlight the effectiveness and potential of AF Adapter in domain-specific pretraining for language models.

ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Program of China (No. 2021YFC2701800, 2021YFC2701801), the National Natural Science Foundation of China (No. 62306112) and the Shanghai Science Foundation (No. 23YF1409400).

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.
[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
[3] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, “Domain-specific language model pretraining for biomedical natural language processing,” ACM Transactions on Computing for Healthcare (HEALTH), vol. 3, no. 1, pp. 1–23, 2021.
[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
[5] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” in Psychology of learning and motivation. Elsevier, 1989, vol. 24, pp. 109–165.
[6] R. Ratcliff, “Connectionist models of recognition memory: constraints imposed by learning and forgetting functions.” Psychological review, vol. 97, no. 2, p. 285, 1990.
[7] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, “An empirical investigation of catastrophic forgetting in gradient-based neural networks,” arXiv preprint arXiv:1312.6211, 2013.
[8] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” arXiv preprint arXiv:1801.06146, 2018.
[9] C. Lee, K. Cho, and W. Kang, “Mixout: Effective regularization to finetune large-scale pretrained language models,” arXiv preprint arXiv:1909.11299, 2019.
[10] N. Housby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attarjan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” in International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
[12] J. Liu, Y. Song, K. Xue, H. Sun, C. Wang, L. Chen, H. Jiang, J. Liang, and T. Ruan, “Fl-tuning: Layer tuning for feed-forward network in transformer,” arXiv preprint arXiv:2206.15312, 2022.
[13] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, “Pre-training with whole word masking for chinese bert,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3504–3514, 2021.
[14] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, “Revisiting pretrained models for Chinese natural language processing,” in Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, Nov, 2020, pp. 657–668.
[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” arXiv preprint arXiv:2106.09685, 2021.
[16] N. Zhang, M. Chen, Z. Bi, X. Liang, L. Li, X. Shang, K. Yin, C. Tan, J. Xu, F. Huang et al., “Cbblue: A chinese biomedical language understanding evaluation benchmark,” arXiv preprint arXiv:2106.08087, 2021.
[17] Y. Yuan, H. Zhao, Z. Du, M. Ding, X. Liu, Y. Cen, Z. Zou, Z. Yang, and J. Tang, “Wudaocorpora: A super large-scale chinese corpora for pre-training language models,” AI Open, vol. 2, pp. 65–68, 2021.

TABLE V
CATASTROPHIC FORGETTING ANALYSIS

| Model                  | Accuracy | Diff |
|------------------------|----------|------|
| RoBERTa                | 87.752   | -0   |
| Fine-Tuning based RoBERTa | 70.964  | -16.788 |
| FL-Tuning based RoBERTa | 83.555  | -4.197 |
| LoRA based RoBERTa     | 78.231   | -9.521 |
| AF Adapter based RoBERTa (ours) | 82.428 | -5.324 |