Prompt Tuning for Discriminative Pre-trained Language Models

Yuan Yao[1], **Bowen Dong**[1], **Ao Zhang**[2], **Zhengyan Zhang**[1],
Ruobing Xie[3], **Zhiyuan Liu**[1,4,5,6†], **Leyu Lin**[3], **Maosong Sun**[1,4,5,6†], **Jianyong Wang**[1]
[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
[2]Department of Computer Science, National University of Singapore, Singapore
[3]WeChat Search Application Department, Tencent, China
[4]Institute for Artificial Intelligence, Tsinghua University, Beijing, China
[5]Institute Guo Qiang, Tsinghua University, Beijing, China
[6]International Innovation Center of Tsinghua University, Shanghai, China
yaoyuanthu@163.com dongbw18@mails.tsinghua.edu.cn

Abstract

Recent works have shown promising results of prompt tuning in stimulating pre-trained language models (PLMs) for natural language processing (NLP) tasks. However, to the best of our knowledge, existing works focus on prompt-tuning generative PLMs that are pre-trained to generate target tokens, such as BERT (Devlin et al., 2019). It is still unknown whether and how discriminative PLMs, e.g., ELECTRA (Clark et al., 2020), can be effectively prompt-tuned. In this work, we present DPT, the first prompt tuning framework for discriminative PLMs, which reformulates NLP tasks into a discriminative language modeling problem. Comprehensive experiments on text classification and question answering show that, compared with vanilla fine-tuning, DPT achieves significantly higher performance, and also prevents the unstable problem in tuning large PLMs in both full-set and low-resource settings. The source code and experiment details of this paper can be obtained from https://github.com/thunlp/DPT.

1 Introduction

Recent years have witnessed the great success of the *pre-training-then-fine-tuning* paradigm in natural language processing (NLP) (Devlin et al., 2019; Yang et al., 2019; Clark et al., 2020; Lan et al., 2020; Raffel et al., 2020). Typically, language models are first pre-trained on large-scale corpora via self-supervised generative or discriminative tasks to learn universal text representations, and then fine-tuned to adapt to downstream tasks (Qiu et al., 2020; Xu et al., 2021). However, the significant gap between the objective forms of model pre-training and fine-tuning hinders taking full advantage of PLMs in downstream tasks (Liu et al., 2021).

Prompt tuning has recently shown its effectiveness in stimulating the capability of PLMs by transforming downstream tasks into the same form as pre-training (Petroni et al., 2019; Brown et al., 2020; Schick and Schütze, 2021; Gao et al., 2021; Liu et al., 2021). However, to the best of our knowledge, existing works focus on prompt-tuning generative PLMs (i.e., PLMs pre-trained by generating target textual tokens from the context, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020)). It is still unknown whether and how discriminative PLMs can be effectively prompt-tuned (i.e., PLMs pre-trained by discriminating replaced tokens, such as ELECTRA (Clark et al., 2020) and WKLM (Xiong et al., 2020)). Since discriminative PLMs typically enjoy competitive performance and superior computational efficiency compared with their generative counterparts (Clark et al., 2020), it can be especially appealing to prompt-tuning discriminative PLMs.

In this work, we present DPT, the first prompt tuning framework for discriminative PLMs. DPT reformulates downstream tasks into a discriminative language modeling problem, maximally mitigating the gap between model pre-training and tuning. Specifically, as shown in Figure 1, models are asked to discriminate correct answer tokens (e.g., correct labels for text classification, or answer spans for question answering) from the input tokens based on the reused discriminative language modeling head, where the objective form is identical to pre-training.

To evaluate DPT, we conduct comprehensive experiments on text classification and question an-

[†] Corresponding authors: Z.Liu (liuzy@tsinghua.edu.cn), M.Sun (sms@tsinghua.edu.cn)

Findings of the Association for Computational Linguistics: ACL 2022, pages 3468 - 3473
May 22-27, 2022

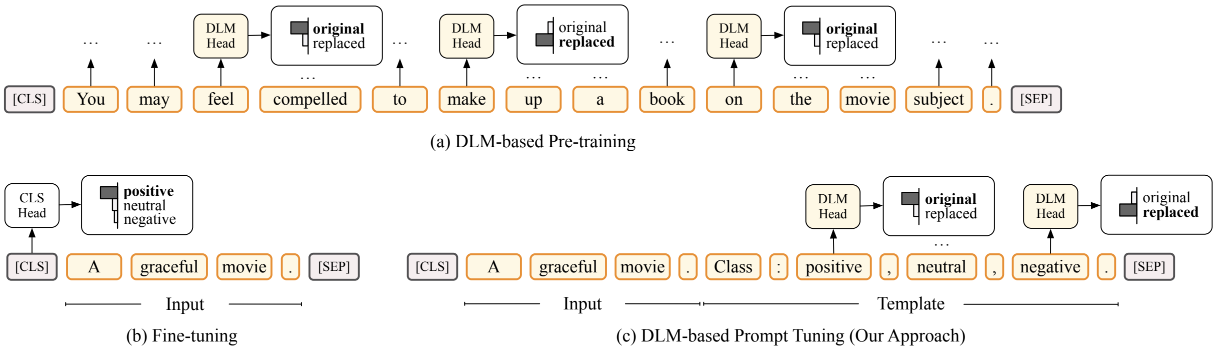


Figure 1: Illustration of (a) discriminative language modeling (DLM) based pre-training with the DLM head, (b) vanilla fine-tuning with a new classification (CLS) head, and (c) our DPT prompt tuning approach that reformulates NLP tasks into a discriminative language modeling problem. DPT fills the input text into the template containing answer candidates, and discriminates whether each answer candidate is correct (i.e., original), or incorrect (i.e., replaced) based on the reused DLM head.

swering in both full-set and low-resource settings. Experimental results show that despite its simplicity, DPT significantly outperforms vanilla fine-tuning (e.g., 4.1% accuracy improvement in the low-resource SST-5 evaluation). Moreover, previous works have shown that fine-tuning large PLMs can be highly unstable and even produce divergent results (Devlin et al., 2019; Dodge et al., 2020), which undermines the practicality of large PLMs. We show that DPT also addresses the unstable problem in tuning large discriminative PLMs.

The contributions of our work are summarized as follows: (1) We present the first prompt tuning framework for discriminative PLMs. (2) Comprehensive experimental results on text classification and question answering demonstrate the effectiveness of the proposed prompt tuning framework.

2 Preliminary

In this work, without loss of generality, we take ELECTRA (Clark et al., 2020) as a representative example of discriminative PLMs, while applying DPT to other discriminative PLMs is also applicable. Here we introduce the main procedure of pre-training and fine-tuning, and we refer readers to the paper (Clark et al., 2020) for more details.

Pre-training. During pre-training, a generator first corrupts the text via token replacement. Then the discriminator is asked to detect the replaced tokens, by classifying each token into binary categories, i.e., $\{\text{original}, \text{replaced}\}$, as shown in Figure 1. Finally, the generator is discarded and the discriminator is fine-tuned on downstream tasks.

Vanilla Fine-tuning. (1) During fine-tuning, to perform *text classification*, a new classification head is typically introduced to classify the hidden representation of the `[CLS]` token in the last layer (Clark et al., 2020). (2) For general *multi-span question answering*, the answer could be multiple spans from the input text (Dasigi et al., 2019; Dua et al., 2019). State-of-the-art fine-tuning approaches formulate the task as a sequence-labeling problem, and classify each input token into binary labels based on a new classification head, indicating whether the token belongs to the answer or not (Segal et al., 2020; Ye et al., 2020).

Note that the classification head typically introduces new parameters, and learning the parameters from scratch usually requires a large amount of labeled data. Moreover, previous works have shown that fine-tuning large PLMs can be highly unstable, and even produce divergent results (Devlin et al., 2019; Dodge et al., 2020). As a result, multiple fine-tuning trials are usually needed to find a good random seed that leads to a stably fine-tuned PLM, which undermines the practicality of large PLMs.

3 Methodology

In this section, we introduce the framework of DPT for prompt-tuning discriminative PLMs. We first introduce DPT using text classification as the running example, and then illustrate its application in question answering.

DLM-based Reformulation. DPT reformulates NLP tasks into a dscriminative language modeling problem, maximally mitigating the gap between pre-training and tuning. Specifically, as shown in Figure 1 (c), for a text classification task with class set $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, DPT defines a template that contains all answer candidates $\mathcal{T}(\cdot; \mathcal{C})$. Given an input text x (e.g., "*A graceful movie.*"), DPT fills the input text into the template as follows:

$$\mathcal{T}(x;\mathcal{C}) = \texttt{[CLS]}\, x \text{ Class: } c_1, c_2, \ldots, c_n.\texttt{[SEP]} \quad (1)$$

Intuitively, $\mathcal{T}(x;\mathcal{C})$ can be understood as creating a virtual context that assumes all candidate classes are correct for the input text x. It is then straightforward for discriminative PLMs to decide whether each class candidate token is proper in the context, by classifying the tokens into original (i.e., correct), or replaced (i.e., incorrect) based on the reused DLM head. In our experiments, we find that the order of classes in template has minimal influence on the performance, and a random order can produce good prompt-tuning results.

DPT Training. After template filling, $\mathcal{T}(x;\mathcal{C})$ is fed into PLMs to obtain the hidden representations $\{\mathbf{h}_{\texttt{[CLS]}}, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m, \mathbf{h}_{\texttt{[SEP]}}\}$. PLMs are then prompted to discriminate whether each class is correct. Specifically, we compute the score of class c_i based on the representation of the corresponding token t_i as:[1]

$$s(c_i) = 1 - \sigma(\mathbf{h}_{\text{DLM}}^{\top}\mathbf{h}_{t_i}), \quad (2)$$

where $\mathbf{h}_{\text{DLM}}$ is the reused DLM head, and $\sigma(\cdot)$ is the sigmoid activation. Note that in Equation 2, the computation of class scores is different from the vanilla fine-tuning approaches which encourage large inner products between the correct answer and classification head (Devlin et al., 2019; Clark et al., 2020). The rationale is that during pre-training, discriminative PLMs are typically required to produce large inner products for the replaced tokens (i.e., incorrect ones), and small inner products for the original tokens (i.e., correct ones) (Clark et al., 2020), and therefore Equation 2 better fits the semantics in pre-training. In our experiments, we find this simple operation can lead to significantly better results in prompt-tuning discriminative PLMs. After obtaining the class score, the model is optimized as:

$$\mathcal{L} = \sum_i [-y_i \log s(c_i) - (1-y_i)\log(1 - s(c_i))], \quad (3)$$

where $y_i \in \{0, 1\}$ indicates the ground-truth label. Since DPT tunes PLMs by reusing the pre-trained DLM head in the same objective form as pre-training, compared with vanilla fine-tuning, we expect DPT will lead to more sample efficient and stable tuning results.

DPT for Question Answering. Besides text classification, DPT can also be applied for the question answering task. Given a question and a paragraph, directly concatenating them without additional templates can already create a good prompting context. Then similar to text classification, we ask PLMs to discriminate whether each token in the paragraph is part of the answer (i.e., original), or not (i.e., replaced) based on the reused DLM head. During inference, we threshold the token scores to obtain multiple answer spans.

4 Experiments

In this section, we empirically evaluate DPT on the task of text classification and question answering.

Datasets. We evaluate DPT on four widely used text classification datasets, including SST-2, SST-5, TREC and AGNews. For question answering, we adopt the challenging QUOREF dataset, where for each question, there may exist multiple answer spans in the paragraph. We refer readers to Section B for more dataset details.

Evaluation Protocols. We evaluate the models under two settings, including (1) *full-set* setting, where the full training data is available, and (2) *low-resource* setting, where only 10% of the full training data for each dataset is available. We report the accuracy for text classification, and exact match (EM) and F1 score for question answering. To account for the unstable problem of baseline models, we report the average results from 3 best random seeds among 10 trials.

Baselines. We compare DPT with several strong baseline models, including vanilla fine-tuning of ELECTRA (Clark et al., 2020), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The fine-tuning of ELECTRA adopts the identical discriminative PLM to our model, and serves as the most direct baseline for comparison.

Main Results. We report the main results in Table 1 and Table 2, from which we observe that: (1) DPT significantly improves the performance of discriminative PLMs. The improvements are consistent across different tasks and datasets, as well as base and large models. (2) Previous works show that despite the significant improvements in low-resource setting, template-based prompt tuning typically can only approach fine-tuning performance in full-set

[1]If the class name consists of multiple tokens, the representation of the first token is used.

	PLM	Tuning Approach	Full-set Setting SST-2	SST-5	TREC	AGNews	Low-resource Setting SST-2	SST-5	TREC	AGNews
Base	BERT	FT	91.32	53.41	95.93	93.68	86.91	42.46	86.73	90.23
	RoBERTa	FT	94.69	56.09	95.27	93.92	91.23	50.41	91.07	90.25
	ELECTRA	FT	94.38	56.60	94.87	93.70	91.68	49.40	88.40	89.17
	ELECTRA	DPT (Ours)	**95.26**	**58.34**	**96.27**	**94.22**	**93.83**	**53.48**	**93.93**	**90.60**
		Δ	**+0.88**	**+1.74**	**+1.40**	**+0.52**	**+2.15**	**+4.08**	**+5.53**	**+1.43**
Large	BERT	FT	93.32	54.10	96.73	94.89	90.77	50.89	94.73	92.93
	RoBERTa	FT	95.46	56.80	96.80	95.26	94.27	51.41	95.20	93.41
	ELECTRA	FT	95.72	58.27	97.13	94.80	93.74	53.65	94.00	92.33
	ELECTRA	DPT (Ours)	**96.58**	**60.69**	**98.07**	**95.38**	**96.09**	**57.00**	**95.67**	**93.58**
		Δ	**+0.86**	**+2.42**	**+0.94**	**+0.58**	**+2.35**	**+3.35**	**+1.67**	**+1.25**

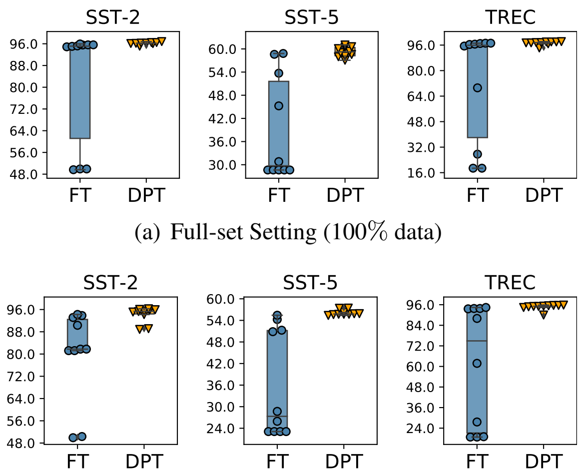
Table 1: Experimental results on text classification. Full-set setting: 100% data, Low-resource setting: 10% data. FT: fine-tuning, DPT: discriminative prompt tuning. Δ: Improvements of DPT over fine-tuning ELECTRA.

PLM	Tuning Approach	Full Set EM	F1	Low Resource EM	F1
BERT	FT	75.67	79.99	53.02	61.36
RoBERTa	FT	78.29	84.56	59.31	67.56
ELECTRA	FT	77.79	83.72	54.29	63.71
ELECTRA	DPT (Ours)	**79.66**	**86.03**	**63.65**	**73.09**
	Δ	**+1.87**	**+2.31**	**+9.36**	**+9.38**

Table 2: Experimental results of $\text{ELECTRA}_{\text{large}}$ on QUOREF multi-span question answering dataset.

Tuning Approach	SST-2	SST-5	TREC	AGNews
Fine-tuning	91.68	49.40	88.40	89.17
DPT (σ)	92.16	50.96	88.00	90.29
DPT ($1-\sigma$)	**93.83**	**53.48**	**93.93**	**90.60**

Table 3: Ablation on reuse forms of DLM head based on $\text{ELECTRA}_{\text{base}}$ in low-resource setting.



(a) Full-set Setting (100% data)

(b) Low-resource Setting (10% data)

Figure 2: Performance distribution of $\text{ELECTRA}_{\text{large}}$ using fine-tuning and DPT from 10 seeds.

setting (Gao et al., 2021). In comparison, we note that DPT can improve the performance in both low-resource and full-set settings. The reason is that DPT enables PLMs to jointly model the input text and class candidates for better text understanding. In summary, DPT is effective in improving the performance of discriminative PLM tuning.

Tuning Stability. Previous works have commonly observed the instability of fine-tuning large generative PLMs (Devlin et al., 2019; Dodge et al., 2020). Some works attempt to alleviate the problem by careful initialization and optimization (Zhang et al., 2021), or intermediate fine-tuning on other large-scale datasets (Phang et al., 2018). To investigate the tuning stability of discriminative PLMs, we tune $\text{ELECTRA}_{\text{large}}$ using fine-tuning and DPT from 10 random seeds. From the results in Figure 2, we observe that: (1) Similar to generative PLMs, fine-tuning large discriminative PLMs is also highly unstable, and can even frequently produce divergent results (e.g., nearly 20% accuracy for 5-way classification in SST-5 in low-resource setting). The problem is exacerbated by sparse data in low-resource setting, but remains even in full-set setting. (2) DPT achieves significantly more stable tuning results in both full-set and low-resource settings, where all tuning trials converged and closely approach the best performance. This is due to the reuse of DLM head parameters and identical objective forms to pre-training.

Ablation Study. In DPT, different from conventional fine-tuning approaches, correct labels are encouraged to have small inner products with classifiers (as indicated by the $1-\sigma$ in Equation 2). We evaluate DPT using conventional score computation (i.e., σ), and report the results in Table 3. The significant drop in performance shows that a proper form of reusing DLM head is crucial to the results of prompt-tuning discriminative PLMs.

5 Conclusion and Future Work

In this work, we present a simple and effective prompt tuning approach for discriminative PLMs. We note directly performing large-scale classification (e.g., for hundreds of classes) with DPT may be computationally inefficient. In future, we plan to address the problem by classifying text following class hierarchies, where each hierarchical layer typically consists of a moderate number of classes.

6 Acknowledgement

This work is suooprted by the Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 61621136008 / DFC TRR-169, Institute Guo Qiang at Tsinghua University, and International Innovation Center of Tsinghua University, Shanghai, China.

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. QUOREF: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of EMNLP-IJCNLP*, pages 5924–5931.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pages 2368–2378.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of ACL-IJCNLP*, pages 3816–3830.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of ICLR*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pages 2463–2473.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTS: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *SCTS*, pages 1–26.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21:140:1–140:67.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of NAACL-HLT*, pages 2339–2352.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of EMNLP*, pages 3074–3080.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, page 1631–1642.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of SIGIR*, pages 200–207.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *Proceedings of ICLR*.

Han Xu, Zhang Zhengyan, et al. 2021. Pre-trained models: Past, present and future. *arXiv preprint arXiv:2106.07139*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of NeurIPS*, pages 5754–5764.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of EMNLP*, pages 7170–7186.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample BERT fine-tuning. In *Proceedings of ICLR*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of NIPS*, pages 649–657.

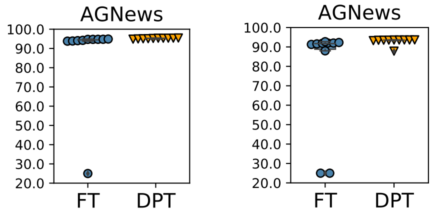
A Implementation Details

In this work, we take ELECTRA (Clark et al., 2020) as an representative example of discriminative PLMs, including (1) $\text{ELECTRA}_{\text{base}}$ with 768 dimensional hidden representations, 12 encoding layers and 110M parameters, and (2) $\text{ELECTRA}_{\text{large}}$ with 1, 024 dimensional hidden representations, 24 encoding layers and 340M parameters.

For text classification tasks, we follow the hyperparameters in Clark et al. (2020), and train the base models for 10 epochs with learning rate 2e-5 and batchsize 32 on 2 GeForce RTX 2080 Ti GPUs. And we train the large models for 10 epochs with learning rate 2e-5 and batchsize 8 on 2 GeForce RTX 2080 Ti GPUs. For question answering, we follow the hyparameters in Segal et al. (2020), and train the large models for 20 epochs with learning rate 5e-6 and batchsize 2 on 6 GeForce RTX 2080 Ti GPUs. During inference, a token is considered as part of the answer if its score is lower than 0.6.

B Dataset Details

We evaluate DPT on four popular text classification datasets, including SST-2 (Socher et al., 2013), SST-5 (Socher et al., 2013), TREC (Voorhees and Tice, 2000) and AGNews (Zhang et al., 2015). For question answering task, we adopt the challenging QUOREF dataset (Dasigi et al., 2019), where there may exist multiple answers in the paragraph for each question. Specifically, QUOREF contains 21, 817 questions and 4, 225 paragraphs, where each question has 1.15 answers on average. The average length for the questions and paragraphs are 15.49 and 325.68 respectively. We report the results on the validation set for QUOREF, since its test set is not publicly available, and report the results on the test set for the other datasets.



(a) Full Set. (b) Low Resource.

Figure 3: Performance distribution of $\text{ELECTRA}_{\text{large}}$ using fine-tuning and DPT from 10 seeds.

C Further Results of Tuning Stability

We report the performance distribution of AGNews in Figure 3. We observe that the unstable problem of fine-tuning large discriminative PLMs remains even for the large-scale AGNews dataset with 120K training samples. The results show the advantage of DPT in stably tuning discriminative PLMs.