PROMPT-DISTILLER: FEW-SHOT KNOWLEDGE DISTILLATION FOR PROMPT-BASED LANGUAGE LEARNERS WITH DUAL CONTRASTIVE LEARNING

Boyu Hou[1], Chengyu Wang[2], Xiaoqing Chen[1], Minghui Qiu[2,†], Liang Feng[1,†], Jun Huang[2]

[1] College of Computer Science, Chongqing University, Chongqing, China
[2] Alibaba Group, Hangzhou, China

ABSTRACT

Prompt-based learning has improved the few-shot learning performance of large-scale Pre-trained Language Models (PLMs). Yet, it is challenging to deploy large-scale PLMs in resource-constrained environments for online applications. Knowledge Distillation (KD) is a promising approach for PLM compression. However, distilling prompt-tuned PLMs in the few-shot learning setting is a non-trivial problem due to the lack of task-specific training data and KD techniques for the new prompting paradigm. We propose Prompt-Distiller, the first few-shot KD algorithm for prompt-tuned PLMs, which forces the student model to learn from both its pre-trained and prompt-tuned teacher models to alleviate the model overfitting problem. We further design a contrastive learning technique to learn higher-order dependencies from intermediate-layer representations of teacher models, considering different knowledge capacities of teacher and student models. Extensive experiments over various datasets show that Prompt-Distiller consistently outperforms baselines by a large margin.

Index Terms— knowledge distillation, few-shot learning, prompt-based learning, pre-trained language model

1. INTRODUCTION

Large-scale Pre-trained Language Models (PLMs) have significantly improved the performance of various downstream NLP tasks with the two-stage "pre-training and fine-tuning" paradigm [**?**, **?**]. Yet, a sufficiently large task-specific training set is required for model fine-tuning, which limits the applicability of PLMs.

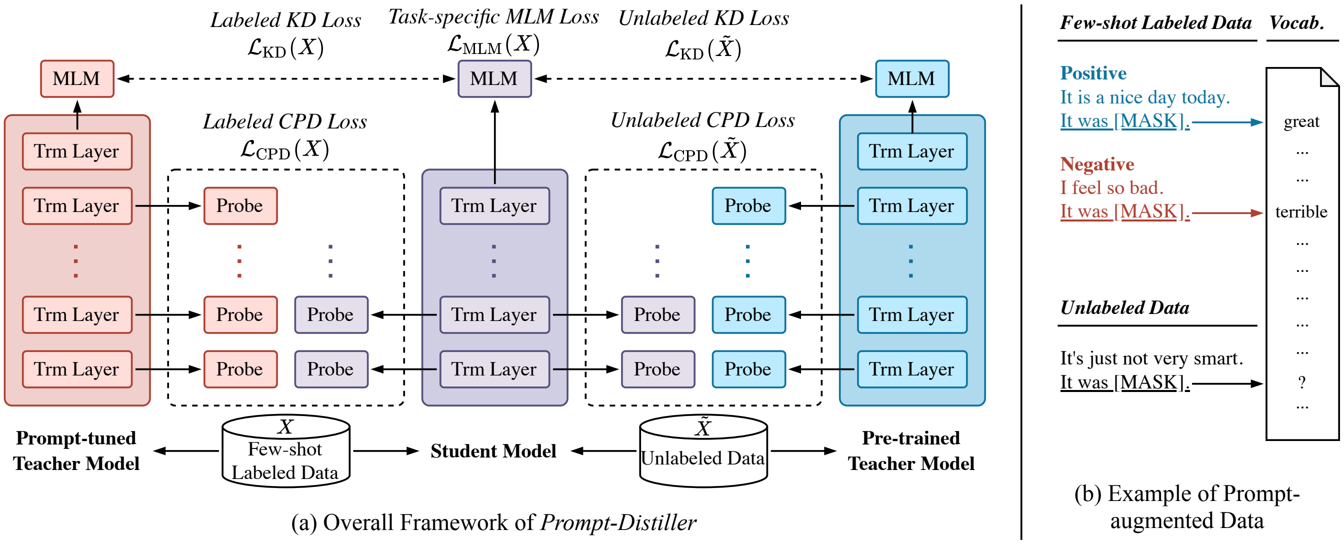
Recently, prompt-based learning has been proposed to reformulate different NLP tasks uniformly as cloze questions and to provide additional task guidance by task-specific prompts [**?**]. Notable works include PET [**?**], LM-BFF [**?**], AutoPrompt [**?**], WARP [**?**], P-tuning [**?**] and many others. As prompting requires rich pre-training knowledge "stored" in PLMs by the Masked Language Modeling (MLM) task, larger PLMs typically have better few-shot performance. Especially, ultra-large PLMs such as GPT-3 [**?**], Yuan 1.0 [**?**] and FLAN [**?**] even have good zero-shot performance on previously unseen NLP tasks.

Unfortunately, the good performance of large models also means that it is challenging to deploy them online in resource-constrained environments, which introduce inference delays and large computation costs. A promising approach to address this issue is Knowledge Distillation (KD) [**?**], which compresses a large model (i.e., the teacher model) into a small model (i.e., the student model), preserving the model performance while reducing the number of parameters. However, distilling such models in few-shot learning settings is non-trivial for two reasons. i) Existing KD algorithms for BERT-like models such as vanilla KD [**?**], Patient KD (PKD) [**?**] and TinyBERT [**?**] mostly focus on distilling the "dark knolwedge" (classification logits) and/or "hints" (intermediate representations) from the teacher model, which are not specifically designed for prompt-based PLMs. ii) In few-shot learning, the extreme lack of training data and the large parameter size of PLMs easily lead to severe overfitting of the student model. Hence, a natural question arises: how can we distill a prompt-based fine-tuned PLM to smaller models with few labeled training instances available?

In this paper, we propose Prompt-Distiller, the first few-shot KD algorithm for prompt-based language learners. As the high performance of few-shot learners mostly result from the exploitation of pre-trained knowledge to downstream tasks, we force the MLM head of the student model to learn from both its pre-trained and prompt-tuned teacher models for the acquisition of task-specific and universal knowledge. Specifically, the KD process of universal, pre-trained knowledge does not require any manually labeled data, which is suitable for few-shot learning. We further design the probe-based contrastive learning technique to indirectly learn higher-order dependencies from intermediate-layer representations, where teacher and student models have significantly different capacities. We conduct extensive experiments over eight public datasets, associated with a variety of NLP tasks. Experiments show that Prompt-Distiller consistently outperforms baselines by a large margin.

Boyu Hou and Chengyu Wang contributed equally to this work.

† Corresponding authors: Liang Feng (`liangf@cqu.edu.cn`) and Minghui Qiu (`minghui.qmh@alibaba-inc.com`).



(a) Overall Framework of *Prompt-Distiller*

(b) Example of Prompt-augmented Data

Fig. 1. An illustration of the Prompt-Distiller framework with examples of prompt-augmented data.

2. PROMPT-DISTILLER: PROPOSED APPROACH

We start with a brief overview of Prompt-Distiller. After that, we elaborate the technical details.

2.1. Overview

For illustration, Prompt-Distiller and data samples are presented in Figure 1. Formally, given an N-way-K-shot training set $X = \{(x_i, y_i)\}$, together with a prompt-tuned PLM parameterized by Θ_T, the goal of our work is to obtain a much smaller PLM parameterized by Θ_S such that the performance of Θ_S can be as close to Θ_T as possible. Here, y_i is the classification label of the input text x_i, where $\mathcal{Y}$ is the label set with $|\mathcal{Y}| = N$ and $y_i \in \mathcal{Y}$. During parameter tuning, following [**?**], we also have a few-shot development set that has the same size with X. As the size of X is extremely small (i.e., $N \times K$), we also assume that there is a larger, unlabeled dataset $\tilde{X} = \{x_i\}$ with $|\tilde{X}| = n \cdot |X|$, serving as the auxiliary dataset for KD. Without loss of generality, we use PET [**?**] to prompt-tune the teacher model on the training set X. Hence, the entire workflow of obtaining teacher and student models only requires $N \times K$ labeled training samples.

2.2. Learning from Dual Teachers

Following PET [**?**], let $l(y)$ be the label word for the class y, and $s_{\Theta_T}(l(y)|x_i)$ be the score of predicting $l(y)$ at the masked language token with the input x_i and the PLM Θ_T. The probability of x_i being assigned to the class y based on Θ_T is defined as follows:

$$p_T(y|x_i) = \frac{\exp\{s_{\Theta_T}(l(y)|x_i)\}}{\sum_{y' \in \mathcal{Y}} \exp\{s_{\Theta_S}(l(y')|x_i)\}}. \qquad (1)$$

We further denote $p_T(\vec{y}|x_i)$ as the probability vector across all N classes $\mathcal{Y}$. $\vec{y}_i$ is the corresponding N-dimensional one-hot ground-truth vector for x_i. It is straightforward to derive the classification loss for the student model as follows:

$$\mathcal{L}_{\text{MLM}}(X) = \frac{1}{|X|} \sum_{(x_i, y_i) \in X} \text{CE}(\vec{y}_i, p_S(\vec{y}|x_i)) \qquad (2)$$

where $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss between the two vectors. During the KD process, we wish the student model to also learn from the MLM prediction of the prompt-tuned teacher model. We define the labeled KD loss as follows:

$$\mathcal{L}_{\text{KD}}(X) = \frac{1}{|X|} \sum_{(x_i, y_i) \in X} \text{CE}\left(\frac{p_T(\vec{y}|x_i)}{\alpha}, p_S(\vec{y}|x_i)\right) \qquad (3)$$

where $\alpha > 0$ is the temperature factor.

A significant challenge of our work is the lack of training data, which makes the supervised signals rather limited. Here, we consider learning directly from the pre-trained teacher model without any prompt-tuning. Assume the teacher model Θ_T is prompt-tuned from its pre-trained initialization $\Theta_{T'}$. We further define the unlabeled KD loss based on $\Theta_{T'}$ and $\tilde{X}$ as follows:

$$\mathcal{L}_{\text{KD}}(\tilde{X}) = \frac{1}{|\tilde{X}|} \sum_{x_i \in \tilde{X}} \text{CE}\left(\frac{p_{T'}(\vec{y}|x_i)}{\alpha}, p_S(\vec{y}|x_i)\right). \qquad (4)$$

2.3. Learning with Contrastive Learning

Apart from the MLM head, intermediate representations in PLMs provide useful hints for KD. However, due to the gap in model capacities, directly minimizing their differences yield

Paradigm	Method	MNLI	MNLI-mm	MPQA	MR	RTE	SNLI	SST-2	TREC	Avg.
FT	*Teacher FT (Upper Bound)*	*44.3*	*46.3*	*66.3*	*75.0*	*50.5*	*48.5*	*79.7*	*85.3*	*61.9*
	Student FT (Lower Bound)	34.8	35.1	63.9	53.8	51.3	41.4	63.5	69.0	51.6
	Vanilla KD [?]	34.9	35.6	64.5	52.5	50.2	40.6	63.5	68.5	51.3
	BERT-PKD [?]	35.4	36.1	56.2	46.8	53.1	43.7	62.5	59.3	49.1
	TinyBERT [?]	35.6	36.4	64.7	48.8	53.3	42.8	63.5	63.2	51.0
PT	*Teacher PT (Upper Bound)*	*67.0*	*69.0*	*84.5*	*87.0*	*71.3*	*77.3*	*93.1*	*85.8*	*79.3*
	Student PT (Lower Bound)	35.4	36.2	64.0	55.8	51.8	38.9	62.5	70.5	51.8
	Prompt-KD	35.5	36.0	63.5	53.8	53.7	39.3	61.9	69.9	51.6
	Prompt-Distiller (Ours)	**43.3**	**45.0**	**75.5**	**71.9**	**55.6**	**48.0**	**78.4**	**71.3**	**61.1**

Table 1. Comparison between Prompt-Distiller and baselines in terms of accuracy for few-shot KD (%). "FT" and "PT" refer to traditional fine-tuning and prompt-tuning, respectively.

poor results.[1] We transfer the hidden knowledge in intermediate representations by knowledge probes. We first freeze the parameters of Θ_T and Θ_S, and train an MLM-based probe classifier for each transformer encoder layer (apart from the last layer) w.r.t. the ground-truth label words. In total, we have N_T probes for the teacher and N_S for the student.[2] Denote $p_T^{(j)}(\vec{y}|x_i)$ as the probability vector of x_i being assigned to the N classes $\mathcal{Y}$ based on Θ_T and the j-th probe ($j = 1, \cdots, N_T$). Similarly, the result probability from the student model is represented as $p_S^{(k)}(\vec{y}|x_i)$ with $k = 1, \cdots, N_S$. We define the exponential matching score between the two models as follows:

$$f_{T,S}(x_i) = \exp\{\frac{\sum_{j,k} p_T^{(j)}(\vec{y}|x_i) \cdot p_S^{(k)}(\vec{y}|x_i)}{N_T \cdot N_S}\}. \tag{5}$$

Inspired by [?], for each x_i, we randomly select a collection of samples with different ground-truth labels as negative samples, denoted as $\mathcal{N}(x_i)$. We propose the labeled contrastive prompt distillation loss to transfer the intermediate knowledge across models:

$$\mathcal{L}_{\text{CPD}}(X) = -\frac{1}{|X|} \sum_{(x_i,y_i)\in X} \frac{f_{T,S}(x_i)}{\sum_{x_i^{'}\in\mathcal{N}(x_i)} f_{T,S}(x_i^{'})}. \tag{6}$$

Likewise, for the unlabeled dataset $\tilde{X}$, we have the unlabeled contrastive prompt distillation loss:

$$\mathcal{L}_{\text{CPD}}(\tilde{X}) = -\frac{1}{|\tilde{X}|} \sum_{x_i\in\tilde{X}} \frac{f_{T',S}(x_i)}{\sum_{x_i^{'}\in\mathcal{N}(x_i)} f_{T',S}(x_i^{'})}. \tag{7}$$

The only difference is that the negative samples $\mathcal{N}(x_i)$ here cannot be directly extracted based on ground-truth labels (which are unavailable). As a simple heuristic rule, we directly infer the labels of $\tilde{X}$ using the same prompt and the label words over $\Theta_{T'}$ (which can be viewed as zero-shot learning).

In summary, the overall loss function of our Prompt-Distiller framework is:

$$\begin{aligned} \mathcal{L} =& \mathcal{L}_{\text{MLM}}(X) + \lambda_1 \cdot (\mathcal{L}_{\text{KD}}(X) + \mathcal{L}_{\text{KD}}(\tilde{X})) \\ &+ \lambda_2 \cdot (\mathcal{L}_{\text{CPD}}(X) + \mathcal{L}_{\text{CPD}}(\tilde{X})) \end{aligned} \tag{8}$$

where λ_1 and λ_2 are balancing hyper-parameters.

3. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate Prompt-Distiller on various aspects.

3.1. Datasets and Experimental Settings

In the experiments, we employ eight public datasets to evaluate the Prompt-Distiller framework, which are divided into four groups: natural language inference (MNLI [?], SNLI [?] and RTE [?]), question answering (MPQA [?]), question classification (TREC [?]) and sentiment analysis (MR [?] and SST-2 [?]).

We use RoBERTa-large [?] as the teacher model (with around 355M parameters). The default student model is BERT-small [?]. We also test the results of Prompt-Distiller when other sizes of student models are applied. In default, we set the hyper-parameters to be $K = 16$, $\alpha = 10$, $n = 10$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$. We keep all the prompts to be the same as in PET [?]. During model training, we fix the batch size and the learning rate to be 4 and 1e-5, respectively. Other hyper-parameters (λ_1, λ_2 and α) are tuned on the development sets.

Following [?], we test our model over five different few-shot training sets. The unlabeled datsets are randomly sampled from the original training sets with replacement (without overlaps with our few-shot training/development sets). The classification labels of these instances are removed. For evaluation, we report the average model performance in terms of accuracy (with the same random seeds for all methods). The proposed Prompt-Distiller method is implemented in PyTorch and run on Tesla V100 GPUs.

[1] In the exploratory experiments, we also add the loss function for hidden states proposed in [?] to Prompt-Distiller. The KD performance even drops significantly (over 5% in terms of accuracy). Therefore, we employ the probe-based approach for few-shot KD.

[2] For example, if the student model is BERT-base (with 12 transformer encoder layers), we have $N_S = 11$.

Method	SST-2	MPQA	MNLI
Full Implement.	**78.4**	**75.5**	**43.3**
w/o. contrastive learning	75.3	73.3	42.4
w/o. unlabeled data	63.1	64.3	39.2
w/o. both techniques	62.5	64.0	35.4

Table 2. The ablation study of Prompt-Distiller (%).

3.2. Main Results

Main experimental results are shown in Table 1. Two learning paradigms of PLMs are used for comparison, namely standard fine-tuning (FT) and prompt-tuning (PT). For each paradigm, we also list the performance of the respective teacher and student models as upper and lower bounds. The baselines include vanilla KD [**?**], BERT-PKD [**?**] and TinyBERT [**?**] for FT, and Prompt-KD for PT (which distills the logits of the teacher MLM head without unlabeled data and contrastive learning).

Based on the results, we have the following findings. i) Due to the lack of labeled training data for KD, KD approaches for FT yield poor results, which even (surprisingly) perform worse than FT without KD (the lower bound). This means the trained teacher models are severely overfitted, conveying no useful knowledge for distillation. ii) The naive Prompt-KD approach achieves comparable performance to PT without KD, showing that the naive approach is also not sufficient. iii) The Prompt-Distiller framework outperforms all the baselines by a large margin across all datasets.

3.3. Detailed Analysis

Ablation Study. The ablation results of our method are presented in Table 2. We can see that learning from dual teachers contributes larger to the model improvement. This is natural as it directly guides the student on how to do text classification based on the (relatively large) unlabeled dataset. The proposed contrastive learning technique further improves the model accuracy by probing intermediate representations of teacher models.

Dataset Scale Analysis. We further vary the number of training instances per class K from 16 to 512. The trend of performance of Prompt-Distiller and the method without KD (i.e., PET) is reported in Figure 2. As seen, the margin between PET and Prompt-Distiller is consistently large across different Ks, showing that it enhances the model accuracy regardless of the training data size and not restricted to few-shot learning. Yet, it has a greater impact with small training sets.

Model Scale Analysis. We test other sizes of BERT models and report the performance in Table 3. The statistics of these models are in [**?**]. We can also see the accuracy improvement is consistent when our distillation method is applied. Hence, we suggest distilling from larger prompt-tuned models can benefit small models that can be efficiently deployed online.

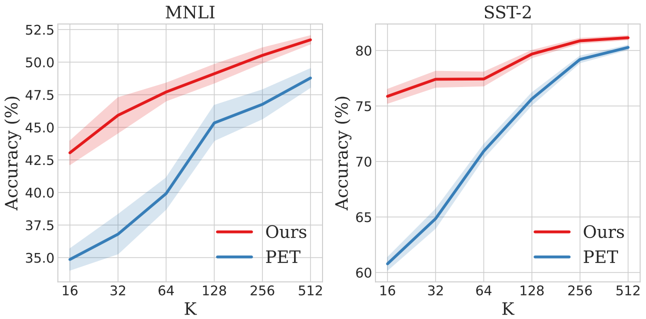


Fig. 2. Dataset scale analysis (%).

Student Model	SST-2	MPQA	MNLI
BERT-base	66.1/87.8	67.2/84.0	40.4/51.5
BERT-small	62.5/77.5	63.9/74.3	34.8/43.3
BERT-tiny	56.8/62.4	60.6/67.1	33.6/34.8

Table 3. Model scale analysis (%). A/B refers to the model performance w/o and w/ KD, respectively.

Model	Top Tokens
Prompt-tuned Teacher	**great**, brilliant, good, $\cdots$
Pre-trained Teacher	fun, **great**, good, $\cdots$
Prompt-Distiller	**great**, shaken, $\cdots$

Table 4. List of tokens generated by the MLM head with high probabilities where the input is "it all adds up to good fun. it was [MASK]." for sentiment analysis.

Case Study. In Table 4, we show a case for sentiment analysis where the results of both prompt-tuned and pre-trained teachers are listed, together with the student model. It is evident both teachers provide sufficient knowledge for the student to learn, even for the zero-shot pre-trained teacher.

4. CONCLUSION AND FUTURE WORK

In this work, we have presented Prompt-Distiller, the first few-shot KD algorithm for prompt-based learners based on dual contrastive learning. It forces the student model to learn from both its pre-trained and prompt-tuned teacher models. Experimental results over a variety of NLP tasks and datasets show that the Prompt-Distiller framework consistently outperforms strong baselines by a large margin. In the future, we would like to extend Prompt-Distiller to other NLP tasks (such as text generation) and other large-scale PLMs (such as the GPT series).

Acknowledgment. This work is partially supported by the Alibaba Group through the Alibaba Innovative Research Program under Grant No. H20210412.