

Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains

Haojie Pan^{1,2*}, Chengyu Wang^{2*}, Minghui Qiu^{2†}, Yichang Zhang², Yaliang Li², Jun Huang²

¹ Zhejiang Lab

² Alibaba Group

{haojie.phj, chengyu.wcy, minghui.qmh}@alibaba-inc.com

{yichang.zyc, yaliang.li, huangjun.hj}@alibaba-inc.com

Abstract

Pre-trained language models have been applied to various NLP tasks with considerable performance gains. However, the large model sizes, together with the long inference time, limit the deployment of such models in real-time applications. One line of model compression approaches considers knowledge distillation to distill large teacher models into small student models. Most of these studies focus on single-domain only, which ignores the transferable knowledge from other domains. We notice that training a teacher with transferable knowledge digested across domains can achieve better generalization capability to help knowledge distillation. Hence we propose a Meta-Knowledge Distillation (Meta-KD) framework to build a meta-teacher model that captures transferable knowledge across domains and passes such knowledge to students. Specifically, we explicitly force the meta-teacher to capture transferable knowledge at both instance-level and feature-level from multiple domains, and then propose a meta-distillation algorithm to learn single-domain student models with guidance from the meta-teacher. Experiments on public multi-domain NLP tasks show the effectiveness and superiority of the proposed Meta-KD framework. Further, we also demonstrate the capability of Meta-KD in the settings where the training data is scarce.

1 Introduction

Pre-trained Language Models (PLM) such as BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) have achieved significant success with the two-stage “pre-training and fine-tuning” process. Despite the performance gain achieved in various NLP tasks, the large number of model parameters

*H. Pan and C. Wang contributed equally to this work.

†M. Qiu is the corresponding author.

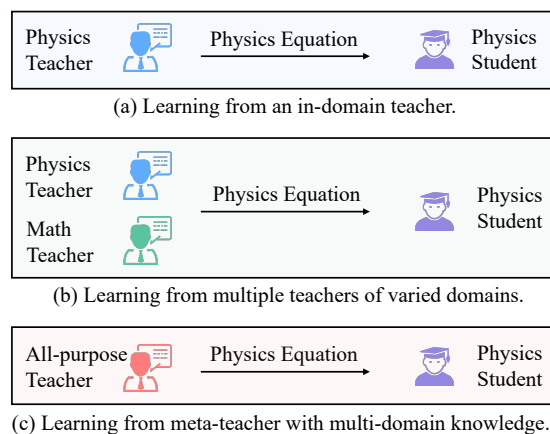


Figure 1: A motivation example of academic learning. A physics student may learn physics equations better with a powerful all-purpose teacher.

and the long inference time have become the bottleneck for PLMs to be deployed in real-time applications, especially on mobile devices (Jiao et al., 2019; Sun et al., 2020; Iandola et al., 2020). Thus, there are increasing needs for PLMs to reduce the model size and the computational overhead while keeping the prediction accuracy.

Knowledge Distillation (KD) (Hinton et al., 2015) is one of the promising ways to distill the knowledge from a large “teacher” model to a small “student” model. Recent studies show that KD can be applied to compress PLMs with acceptable performance loss (Sanh et al., 2019; Sun et al., 2019b; Jiao et al., 2019; Turc et al., 2019; Chen et al., 2020a). However, those methods mainly focus on single-domain KD. Hence, student models can only learn from their in-domain teachers, paying little attention to acquiring knowledge from other domains. It has been shown that it is beneficial to consider cross-domain information for KD, by either training a teacher using cross-domain data or multiple teachers from multiple domains (You et al., 2017; Liu et al., 2019; Yang et al., 2020;

Peng et al., 2020). Consider an academic scenario in Figure 1. A typical way for a physics student to learn physics equations is to directly learn from his/her physics teacher. If we have a math teacher to teach him/her basic knowledge of equations, the student can obtain a better understanding of physics equations. This “knowledge transfer” technique in KD has been proved efficient only when two domains are close to each other (Hu et al., 2019). In reality, however, it is highly risky as teachers of other domains may pass non-transferable knowledge to the student model, which is irrelevant to the current domain and hence harms the overall performance (Tan et al., 2017; Wang et al., 2020). Besides, current studies find multi-task fine-tuning of BERT does not necessarily yield better performance across all the tasks (Sun et al., 2019a).

To address these issues, we leverage the idea of meta-learning to capture transferable knowledge across domains, as recent studies have shown that meta-learning can improve the model generalization ability across domains (Finn et al., 2017; Javed and White, 2019; Yin, 2020; Ye et al., 2020). We further notice that meta-knowledge is also helpful for cross-domain KD. Re-consider the example in Figure 1. If we have an “all-purpose teacher” (i.e., the meta-teacher) who has the knowledge of both physics principles and mathematical equations (i.e., the general knowledge of the two courses), the student may learn physics equations better with the teacher, compared to the other two cases. Hence, it is necessary to train an “all-purpose teacher” model for domain-specific student models to learn.

In this paper, we propose the Meta-Knowledge Distillation (Meta-KD) framework, which facilitates cross-domain KD. Generally speaking, Meta-KD consists of two parts, *meta-teacher learning* and *meta-distillation*. Different from the K-way N-shot problems addressed in traditional meta-learning (Vanschoren, 2018), we propose to train a “meta-learner” as the meta-teacher, which learns the transferable knowledge across domains so that it can fit new domains easily. The meta-teacher is jointly trained with multi-domain datasets to acquire the *instance-level* and *feature-level* meta-knowledge. For each domain, the student model learns to solve the task over a domain-specific dataset with guidance from the meta-teacher. To improve the student’s distillation ability, the meta-distillation module minimizes the distillation loss from both intermediate layers, output layers, and

transferable knowledge, combined with domain-expertise weighting techniques.

To verify the effectiveness of Meta-KD, we conduct extensive experiments on two NLP tasks across multiple domains, namely natural language inference (Williams et al., 2018) and sentiment analysis (Blitzer et al., 2007). Experimental results show the effectiveness and superiority of the proposed Meta-KD framework. Moreover, we find our method performs well especially when i) the in-domain dataset is very small or ii) there is no in-domain dataset during the training of the meta-teacher. In summary, the contributions of this study can be concluded as follows:

- To the best of our knowledge, this work is the first to explore the idea of meta-teacher learning for PLM compression across domains.
- We propose the Meta-KD framework to address the task. In Meta-KD, the meta-teacher digests transferable knowledge across domains, and selectively passes the knowledge to student models with different domain expertise degrees.
- We conduct extensive experiments to demonstrate the superiority of Meta-KD and also explore the capability of this framework in the settings where the training data is scarce.

The rest of this paper is summarized as follows. Section 2 describes the related work. The detailed techniques of the Meta-KD framework are presented in Section 3. The experiments are reported in Section 4. Finally, we conclude our work and discuss the future work in Section 5.¹

2 Related Work

Our study is close to the following three lines of studies, introduced below.

2.1 Knowledge Distillation (KD)

KD was first proposed by (Hinton et al., 2015), aiming to transfer knowledge from an ensemble or a large model into a smaller, distilled model. Most of the KD methods focus on utilizing either the dark knowledge, i.e., predicted outputs (Hinton et al., 2015; Chen et al., 2020b; Furlanello et al., 2018; You et al., 2017) or hints, i.e., the intermediate

¹The experimental code can be found in <https://github.com/alibaba/EasyTransfer/tree/master/scripts/metaKD>.

representations (Romero et al., 2015; Yim et al., 2017; You et al., 2017) or the relations between layers (Yim et al., 2017; Tarvainen and Valpola, 2017) of teacher models. You et al. (2017) also find that multiple teacher networks together can provide comprehensive guidance that is beneficial for training the student network. Ruder et al. (2017) show that multiple expert teachers can improve the performances of sentiment analysis on unseen domains. Tan et al. (2019) apply the multiple-teachers framework in KD to build a state-of-the-art multilingual machine translation system. Feng et al. (2021) considers to build a model to automatically augment data for KD. Our work is one of the first attempts to learn a meta-teacher model that digest transferable knowledge from multiple domains to benefit KD on the target domain.

2.2 PLM Compression

Due to the massive number of parameters in PLMs, it is highly necessary to compress PLMs for application deployment. Previous approaches on compressing PLMs such as BERT (Devlin et al., 2019) include KD (Hinton et al., 2015), parameter sharing (Ullrich et al., 2017), pruning (Han et al., 2015) and quantization (Gong et al., 2014). In this work, we mainly focus on KD for PLMs. In the literature, Tang et al. (2019) distill BERT into BiLSTM networks to achieve comparable results with ELMo (Peters et al., 2018). Chen et al. (2021) studies cross-domain KD to facilitate cross-domain knowledge transferring. Zhao et al. (2019) use dual distillation to reduce the vocabulary size and the embedding size. DistillBERT (Sanh et al., 2019) applies KD loss in the pre-training stage, while BERT-PKD (Sun et al., 2019b) distill BERT into shallow Transformers in the fine-tuning stage. TinyBERT (Jiao et al., 2019) further distills BERT with a two-stage KD process for hidden attention matrices and embedding matrices. Ad-aBERT (Chen et al., 2020a) uses neural architecture search to adaptively find small architectures. Our work improves the prediction accuracy of compressed PLMs by leveraging cross-domain knowledge, which is complementary to previous works.

2.3 Transfer Learning and Meta-learning

TL has been proved to improve the performance on the target domain by leveraging knowledge from related source domains (Pan and Yang, 2010; Mou et al., 2016; Liu et al., 2017; Yang et al., 2017). In most NLP tasks, the “shared-private” architecture

is applied to learn domain-specific representations and domain-invariant features (Mou et al., 2016; Liu et al., 2017; Chen et al., 2018, 2019). Compared to TL, the goal of meta-learning is to train meta-learners that can adapt to a variety of different tasks with little training data (Vanschoren, 2018). A majority of meta-learning methods for include metric-based (Snell et al., 2017; Pan et al., 2019), model-based (Santoro et al., 2016; Bartunov et al., 2020) and model-agnostic approaches (Finn et al., 2017, 2018; Vuorio et al., 2019). Meta-learning can also be applied to KD in some computer vision tasks (Lopes et al., 2017; Jang et al., 2019; Liu et al., 2020; Bai et al., 2020; Li et al., 2020). For example, Lopes et al. (2017) record per-layer meta-data for the teacher model to reconstruct a training set, and then adopts a standard training procedure to obtain the student model. In our work, we use instance-based and feature-based meta-knowledge across domains for the KD process.

3 The Meta-KD Framework

In this section, we formally introduce the Meta-KD framework. We begin with a brief overview of Meta-KD. After that, the techniques are elaborated.

3.1 An Overview of Meta-KD

Take text classification as an example. Assume there are K training sets, corresponding to K domains. In the k -th dataset $\mathcal{D}_k = \{X_k^{(i)}, y_k^{(i)}\}_{i=1}^{N_k}$, $X_k^{(i)}$ is the i -th sample² and $y_k^{(i)}$ is the class label of $X_k^{(i)}$. N_k is the total number of samples in \mathcal{D}_k . Let \mathcal{M}_k be the large PLM fine-tuned on \mathcal{D}_k . Given the K datasets, the goal of Meta-KD is to obtain the K student models $\mathcal{S}_1, \dots, \mathcal{S}_K$ that are small in size but has similar performance compared to the K large PLMs, i.e., $\mathcal{M}_1, \dots, \mathcal{M}_K$.

In general, the Meta-KD framework can be divided into the following two stages:

- **Meta-teacher Learning:** Learn a meta-teacher \mathcal{M} over all domains $\bigcup_{k=1}^K \mathcal{D}_k$. The model digests transferable knowledge from each domain and has better generalization while supervising domain-specific students.
- **Meta-distillation:** Learn K in-domain students $\mathcal{S}_1, \dots, \mathcal{S}_K$ that perform well in their

² $X_k^{(i)}$ can be a sentence, a sentence pair or any other textual units, depending on the task inputs.

respective domains, given only in-domain data \mathcal{D}_k and the meta-teacher \mathcal{M} as input.

During the learning process of the meta-teacher, we consider both *instance-level* and *feature-level* transferable knowledge. Inspired by prototype-based meta-learning (Snell et al., 2017; Pan et al., 2019), the meta-teacher model should memorize more information about prototypes. Hence, we compute sample-wise *prototype scores* as the *instance-level* transferable knowledge. The loss of the meta-teacher is defined as the sum of classification loss across all K domains with *prototype-based, instance-specific weighting*. Besides, it also learns *feature-level transferable knowledge* by adding a *domain-adversarial loss* as an auxiliary loss. By these steps, the meta-teacher is more generalized and digests transferable knowledge before supervising student models.

For meta-distillation, each sample is weighted by a *domain-expertise score* to address the meta-teacher’s capability for this sample. The *transferable knowledge* is also learned for the students from the meta-teacher. The overall meta-distillation loss is a combination of the Mean Squared Error (MSE) loss from intermediate layers of both models (Sun et al., 2019b; Jiao et al., 2019), the soft cross-entropy loss from output layers (Hinton et al., 2015), and the *transferable knowledge distillation loss*, with *instance-specific domain-expertise weighting* applied.

3.2 Meta-teacher Learning

We take BERT (Devlin et al., 2019) as our base learner for text classification due to its wide popularity. For each sample $X_k^{(i)}$, the input is: [CLS], $tok_{k,1}^{(i)}, tok_{k,2}^{(i)}, \dots$, [SEP], where $tok_{k,n}^{(i)}$ is the n -th token in $X_k^{(i)}$. The last hidden outputs of this sequence is denoted as $h_{[CLS]}, h(tok_{k,1}^{(i)}), h(tok_{k,2}^{(i)}), \dots, h(tok_{k,N}^{(i)})$, where $h(tok_{k,j}^{(i)})$ represents the last layer embedding of the j -th token in $X_k^{(i)}$, and N is the maximum sequence length. For simplicity, we define $h(X_k^{(i)})$ as the average pooling of the token embeddings, i.e., $h(X_k^{(i)}) = \sum_{n=1}^N h(tok_{k,n}^{(i)})$.

Learning Instance-level Transferable Knowledge. To select *transferable instances across domains*, we compute a *prototype score* $t_k^{(i)}$ for each sample $X_k^{(i)}$. Here, we treat the *prototype representation* for the m -th class of the k -th domain:

$p_k^{(m)} = \frac{1}{|\mathcal{D}_k^{(m)}|} \sum_{X_k^{(i)} \in \mathcal{D}_k^{(m)}} h(X_k^{(i)})$, where $\mathcal{D}_k^{(m)}$ is the k -th training set with the m -th class label. The *prototype score* $t_k^{(i)}$ is:

$$t_k^{(i)} = \alpha \cos(p_k^{(m)}, h(X_k^{(i)})) + \zeta \sum_{k'=1}^{K(k' \neq k)} \cos(p_{k'}^{(m)}, h(X_k^{(i)})),$$

where \cos is the cosine similarity function, α is a pre-defined hyper-parameter and $\zeta = \frac{1-\alpha}{K-1}$. We can see that the definition of the prototype score here is different from previous meta-learning, as we require that an instance $X_k^{(i)}$ should be close to its class prototype representation in the embedding space (i.e., $p_k^{(m)}$), as well as the prototype representations in out-of-domain datasets (i.e., $p_{k'}^{(m)}$ with $k' = 1, \dots, K, k' \neq k$). This is because the meta-teacher should learn more from instances that are *prototypical across domains* instead of *in-domain only*. For the text classification task, the cross-entropy loss of the meta-teacher is defined using the cross-entropy loss with the prototype score as a weight assigned to each instance.

Learning Feature-level Transferable Knowledge. Apart from the cross-entropy loss, we propose the *domain-adversarial loss* to increase the meta-teacher’s ability for learning feature-level transferable knowledge.

For each sample $X_k^{(i)}$, we first learn an $|h(X_k^{(i)})|$ -dimensional domain embedding of the true domain label $d_k^{(i)}$ by mapping one-hot domain representations to the embeddings, denoted as $\mathcal{E}_D(X_k^{(i)})$. A sub-network is then constructed by:

$$h_d(X_k^{(i)}) = \tanh((h(X_k^{(i)}) + \mathcal{E}_D(X_k^{(i)}))W + b),$$

where W and b are sub-network parameters. The domain-adversarial loss for $X_k^{(i)}$ is defined as:

$$\mathcal{L}_{DA}(X_k^{(i)}) = - \sum_{k=1}^K \mathbf{1}_{k=z_k^{(i)}} \cdot \log \sigma(h_d(X_k^{(i)})),$$

where σ is the K -way domain classifier, and $\mathbf{1}$ is the indicator function that returns 1 if $k = z_k^{(i)}$, and 0 otherwise. Here, $z_k^{(i)} \neq d_k^{(i)}$ is a false domain label of $X_k^{(i)}$ ³. Hence, we deliberately maximize the probability that the meta-teacher makes the *wrong*

³For ease of implementation, we shuffle the domain labels of all instances in a mini-batch.

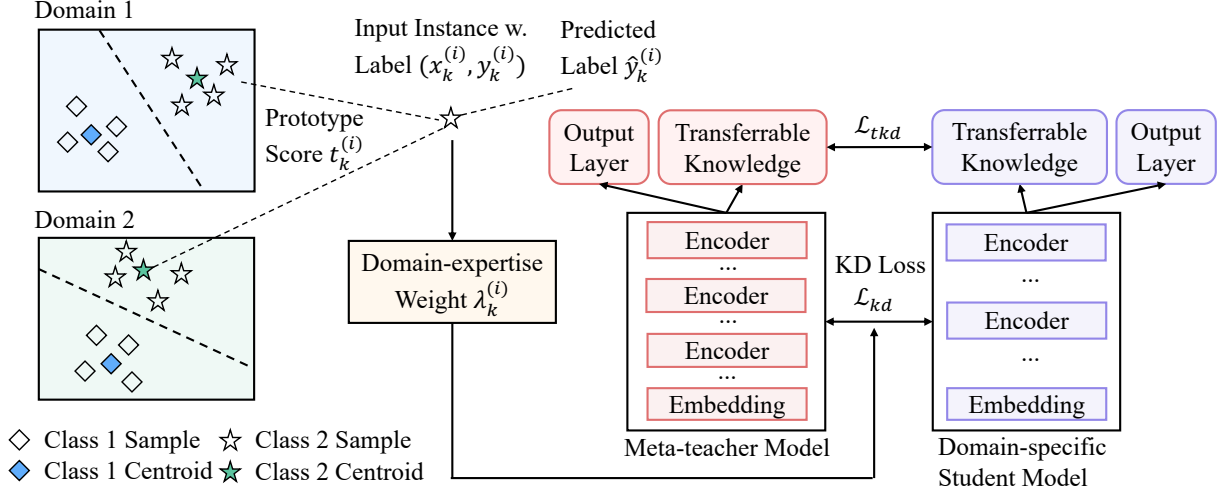


Figure 2: An overview of meta-distillation and the neural architecture that we adopt for knowledge distillation.

predictions of domain labels. We call $h_d(X_k^{(i)})$ as the *transferable knowledge* for $X_k^{(i)}$, which is more insensitive to domain differences.

Let $\mathcal{L}_{CE}(X_k^{(i)})$ be the normal cross-entropy loss of the text classification task. The total loss of the meta-teacher \mathcal{L}_{MT} is the combination of weighted $\mathcal{L}_{CE}(X_k^{(i)})$ and $\mathcal{L}_{DA}(X_k^{(i)})$, shown as follows:

$$\mathcal{L}_{MT} = \sum_{X_k^{(i)} \in \bigcup_{k=1}^K \mathcal{D}_k} \frac{t_k^{(i)} \mathcal{L}_{CE}(X_k^{(i)}) + \gamma_1 \mathcal{L}_{DA}(X_k^{(i)})}{\sum_{k=1}^K |\mathcal{D}_k|},$$

where γ_1 is the factor to represent how the domain-adversarial loss contributes to the overall loss.

3.3 Meta-distillation

We take BERT as our meta-teacher and use smaller BERT models as student models. The distillation framework is shown in Figure 2. In our work, we distill the knowledge in the meta-teacher model considering the following five elements: input embeddings, hidden states, attention matrices, output logits, and transferable knowledge. The KD process of input embeddings, hidden states and attention matrices follows the common practice (Sun et al., 2019b; Jiao et al., 2019). Recall that \mathcal{M} and \mathcal{S}_k are the meta-teacher and the k -th student model. Let $\mathcal{L}_{embd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$, $\mathcal{L}_{hidn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ and $\mathcal{L}_{attn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ be the sample-wise MSE loss values of input embeddings, hidden states and attention matrices of the two models, respectively. Here, $\mathcal{L}_{embd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$, $\mathcal{L}_{hidn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ and $\mathcal{L}_{attn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ refer to the sum of MSE values among multiple hidden layers. We refer

interested readers to Jiao et al. (2019) for more details. $\mathcal{L}_{pred}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ is the cross-entropy loss of “softened” output logits, parameterized by the temperature (Hinton et al., 2015). A naive approach to formulating the total KD loss \mathcal{L}_{kd} is the sum of all previous loss functions, i.e.,

$$\mathcal{L}_{kd} = \sum_{X_k^{(i)} \in \mathcal{D}_k} (\mathcal{L}_{embd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \mathcal{L}_{hidn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \mathcal{L}_{attn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \mathcal{L}_{pred}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})).$$

However, the above approach does not give special considerations to the transferable knowledge of the meta-teacher. Let $h_d^{\mathcal{M}}(X_k^{(i)})$ and $h_d^{\mathcal{S}}(X_k^{(i)})$ be the transferable knowledge of the meta-teacher and the student model w.r.t. the input $X_k^{(i)}$. We further define the *transferable knowledge distillation loss* $\mathcal{L}_{TKD}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})$ as follows:

$$\mathcal{L}_{tkd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) = \frac{1}{|\mathcal{D}_k|} \sum_{X_k^{(i)} \in \mathcal{D}_k} MSE(h_d^{\mathcal{M}}(X_k^{(i)}) W_d^{\mathcal{M}}, h_d^{\mathcal{S}}(X_k^{(i)}))$$

where $W_d^{\mathcal{M}}$ is a learnable projection matrix to match the dimension between $h_d^{\mathcal{M}}(X_k^{(i)})$ and $h_d^{\mathcal{S}}(X_k^{(i)})$, and MSE is the MSE loss function w.r.t. single element. In this way, we encourage student models to learn more transferable knowledge from the meta-teacher.

We further notice that although the knowledge of the meta-teacher should be highly transferable, there still exists the domain gap between the meta-teacher and domain-specific student models. In this

work, for each sample $X_k^{(i)}$, we define the domain expertise weight $\lambda_k^{(i)}$ as follows:

$$\lambda_k^{(i)} = \frac{1 + t_k^{(i)}}{\exp(\hat{y}_k^{(i)} - y_k^{(i)})^2 + 1},$$

where $\hat{y}_k^{(i)}$ is the predicted result of $X_k^{(i)}$'s class label. Here, the weight $\lambda_k^{(i)}$ is large when the meta-teacher model i) has a large prototype score $t_k^{(i)}$ and ii) makes correct predictions on the target input, i.e., $\hat{y}_k^{(i)} = y_k^{(i)}$. We can see that the weight reflects how well the meta-teacher can supervise the student on a specific input. Finally, we derive the complete formulation of the KD loss \mathcal{L}'_{kd} as follows:

$$\begin{aligned} \mathcal{L}'_{kd} = & \sum_{X_k^{(i)} \in \mathcal{D}_k} \lambda_k^{(i)} (\mathcal{L}_{embd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \\ & \mathcal{L}_{hidn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \mathcal{L}_{attn}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}) + \\ & \mathcal{L}_{pred}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)})) + \gamma_2 \mathcal{L}_{tkd}(\mathcal{M}, \mathcal{S}_k, X_k^{(i)}), \end{aligned}$$

where γ_2 is the *transferable KD factor* to represent how the transferable knowledge distillation loss contributes to the overall loss.

4 Experiments

In this section, we conduct extensive experiments to evaluate the Meta-KD framework on two popular text mining tasks across domains.

4.1 Tasks and Datasets

We evaluate Meta-KD over natural language inference and sentiment analysis, using the following two datasets MNLI and Amazon Reviews. The data statistics are in Table 1.

- **MNLI** (Williams et al., 2018) is a large-scale, multi-domain natural language inference dataset for predicting the entailment relation between two sentences, containing five domains (genres). After filtering samples with no labels available, we use the original development set as our test set and randomly sample 10% of the training data as a development set in our setting.
- **Amazon Reviews** (Blitzer et al., 2007) is a multi-domain sentiment analysis dataset, widely used in multi-domain text classification tasks. The reviews are annotated as positive or negative. For each domain, there are 2,000 labeled reviews. We randomly split the data into train, development, and test sets.

Dataset	Domain	#Train	#Dev	#Test
MNLI	Fiction	69,613	7,735	1,973
	Gov.	69,615	7,735	1,945
	Slate	69,575	7,731	1,955
	Telephone	75,013	8,335	1,966
	Travel	69,615	7,735	1,976
Amazon Reviews	Book	1,631	170	199
	DVD	1,621	194	185
	Elec.	1,615	172	213
	Kitchen	1,613	184	203

Table 1: Statistics of the two datasets.

4.2 Baselines

For the teacher side, to evaluate the cross-domain distillation power of the meta-teacher model, we consider the following models as baseline teachers:

- **BERT-single**: Train the BERT teacher model on the target distillation domain only. If we have K domains, then we will have K BERT-single teachers.
- **BERT-mix**: Train the BERT teacher on a combination of K -domain datasets. Hence, we have one BERT-mix model as the teacher model for all domains.
- **BERT-mtl**: Similar to the ‘‘one-teacher’’ paradigm as in BERT-mix, but the teacher model is generated by the multi-task fine-tuning approach (Sun et al., 2019a).
- **Multi-teachers**: It uses K domain-specific BERT-single models to supervise K student models, ignoring the domain difference.

For the student side, we follow TinyBERT (Jiao et al., 2019) to use smaller BERT models as our student models. In single-teacher baselines (i.e., BERT-single, BERT-mix and BERT-mtl), we use TinyBERT-KD as our baseline KD approach. In multi-teachers, because TinyBERT-KD does not naturally support distilling from multiple teacher models, we implement a variant of the TinyBERT-KD process based on MTN-KD (You et al., 2017), which uses averaged softened outputs as the incorporation of multiple teacher networks in the output layer. In practice, we first learn the representations of the student models by TinyBERT, then apply MTN-KD for output-layer KD.

Method	Fiction	Government	Slate	Telephone	Travel	Average
BERT _B -single	82.2	84.2	76.7	82.4	84.2	81.9
BERT _B -mix	84.8	87.2	80.5	83.8	85.5	84.4
BERT _B -mtl	83.7	87.1	80.6	83.9	85.8	84.2
Meta-teacher	85.1	86.5	81.0	83.9	85.5	84.4
BERT _B -single $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	78.8	83.2	73.6	78.8	81.9	79.3
BERT _B -mix $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	79.6	83.3	74.8	79.0	81.5	79.6
BERT _B -mtl $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	79.7	83.1	74.2	79.3	82.0	79.7
Multi-teachers $\xrightarrow{\text{MTN-KD}}$ BERT _S	77.4	81.1	72.2	77.2	78.0	77.2
Meta-teacher $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	80.3	83.0	75.1	80.2	81.6	80.0
Meta-teacher $\xrightarrow{\text{Meta-distillation}}$ BERT _S	80.5	83.7	75.0	80.5	82.1	80.4

Table 2: Results over MNLI (with five domains) in terms of accuracy (%). Here $X \xrightarrow{A} Y$ means it uses X as the teacher and Y as the student, with A as the KD method, hereinafter the same.

Method	Books	DVD	Electronics	Kitchen	Average
BERT _B -single	87.9	83.8	89.2	90.6	87.9
BERT _B -mix	89.9	85.9	90.1	92.1	89.5
BERT _B -mtl	90.5	86.5	91.1	91.1	89.8
Meta-teacher	92.5	87.0	91.1	89.2	89.9
BERT _B -single $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	83.4	83.2	89.2	91.1	86.7
BERT _B -mix $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	88.4	81.6	89.7	89.7	87.3
BERT _B -mtl $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	90.5	81.6	88.7	90.1	87.7
Multi-teachers $\xrightarrow{\text{MTN-KD}}$ BERT _S	83.9	78.4	88.7	87.7	84.7
Meta-teacher $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	89.9	84.3	87.3	91.6	88.3
Meta-teacher $\xrightarrow{\text{Meta-distillation}}$ BERT _S	91.5	86.5	90.1	89.7	89.4

Table 3: Results over Amazon reviews (with four domains) in terms of accuracy (%).

4.3 Implementation Details

In the implementation, we use the original BERT_B model (L=12, H=768, A=12, Total Parameters=110M) as the initialization of all of the teachers, and use the BERT_S model (L=4, H=312, A=12, Total Parameters=14.5M) as the initialization of all the students⁴.

The hyper-parameter settings of the meta-teacher model are as follows. We train 3-4 epochs with the learning rate to be 2e-5. The batch size and γ_1 are chosen from {16, 32, 48} and {0.1, 0.2, 0.5}, respectively. All the hyper-parameters are tuned on the development sets.

⁴<https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/TinyBERT>

For meta-distillation, we choose the hidden layers in {3, 6, 9, 12} of the teacher models in the baselines and the meta-teacher model in our approach to learn the representations of the student models. Due to domain difference, we train student models in 3-10 epochs, with a learning rate of 5e-5. The batch size and γ_2 are tuned from {32, 256} and {0.1, 0.2, 0.3, 0.4, 0.5} for intermediate-layer distillation, respectively. Following Jiao et al. (2019), for prediction-layer distillation, we run the method for 3 epochs, with the batch size and learning rate to be 32 and 3e-5. The experiments are implemented on PyTorch and run on 8 Tsela V100 GPUs.

4.4 Experimental Results

Table 2 and Table 3 show the general testing performance over MNLI and Amazon Reviews of baselines and Meta-KD, in terms of accuracy. From the results, we have the following three major insights:

- Compared to all the baseline teacher models, using the meta-teacher for KD consistently achieves the highest accuracy in both datasets. Our method can help to significantly reduce model size while preserving similar performance, especially in Amazon review, we reduce the model size to 7.5x smaller with only a minor performance drop (from 89.9 to 89.4).
- The meta-teacher has similar performance as BERT-mix and BERT-mtl, but shows to be a better teacher for distillation, as Meta-teacher $\xrightarrow{\text{TinyBERT-KD}}$ BERT_S and Meta-teacher $\xrightarrow{\text{Meta-distillation}}$ BERT_S have better performance than other methods. This shows the meta-teacher is capable of learning more transferable knowledge to help the student. The fact that Meta-teacher \rightarrow Meta-distillation has better performance than other distillation methods confirms the effectiveness of the proposed Meta-KD method.
- Meta-KD gains more improvement on the small datasets than large ones, e.g. it improves from 86.7 to 89.4 in Amazon Reviews while 79.3 to 80.4 in MNLI. This motivates us to explore our model performance on domains with few or no training samples

4.5 Ablation Study

We further investigate Meta-KD’s capability with regards to different portion training data for both of two phases and explore how the transferable knowledge distillation loss contributes to final results.

4.5.1 No In-domain Data during Meta-teacher Learning

In this set of experiments, we consider a special case where we assume all the “fiction” domain data in MNLI is unavailable. Here, we train a meta-teacher without the “fiction” domain dataset and use the distillation method proposed in Jiao et al. (2019) to produce the student model for the “fiction” domain with in-domain data during distillation. The results are shown in Table 4. We find that KD from the meta-teacher can have large

Method	Accuracy
BERT _{B-s} (fiction)	82.2%
Meta-teacher (w/o fiction)	81.6%
<hr/>	
BERT _{B-s} (fiction) $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	78.8%
<hr/>	
BERT _{B-s} (govern) $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	75.3%
BERT _{B-s} (telephone) $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	75.6%
BERT _{B-s} (slate) $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	77.1%
BERT _{B-s} (travel) $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	74.1%
<hr/>	
Meta-teacher $\xrightarrow{\text{TinyBERT-KD}}$ BERT _S	78.2%

Table 4: Results under the setting where no in-domain data used for meta-teacher learning on MNLI. Here, “BERT_{B-s}” refers to the “BERT_{B-single}” method. The distillation is performed on the “fiction” domain data. We report accuracy on the domain dataset.

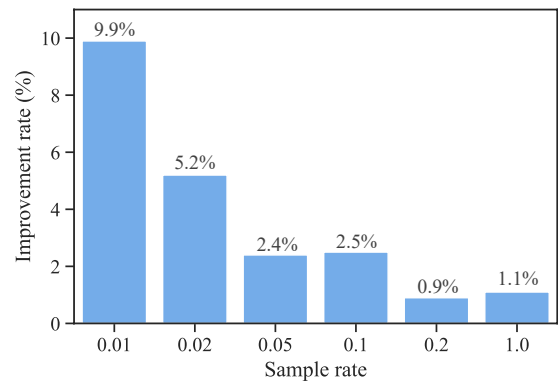


Figure 3: Improvement rate w.r.t different portion (sample rate) of training data in usage.

improvement, compared to KD from other out-domain teachers. Additionally, learning from the out-domain meta-teacher has a similar performance to KD from the in-domain “fiction” teacher model itself. It shows the Meta-KD framework can be applied in applications for emerging domains.

4.5.2 Few In-domain Data Available during Meta-distillation

We randomly sample a part of the MNLI dataset as the training data in this setting. The sample rates that we choose include 0.01, 0.02, 0.05, 0.1 and 0.2. The sampled domain datasets are employed for training student models when learning from the in-domain teacher or the meta-teacher. The experimental results are shown in Figure 3, with results reported by the improvement rate in averaged accuracy. The experimental results show that when less data is available, the improvement rate is much larger. For example, when we only have 1% of the original MNLI training data, the accuracy can be

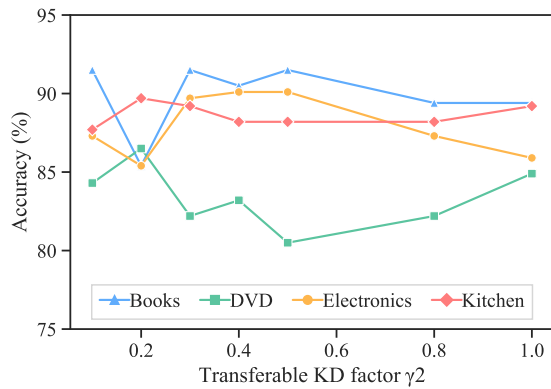


Figure 4: Model performance w.r.t. the transferable KD factor γ_2

increased by approximately 10% when the student tries to learn from the meta-teacher. It shows Meta-KD can be more beneficial when we have fewer in-domain data.

4.5.3 Influence of the Transferable Knowledge Distillation Loss

Here, we explore how the *transferable KD factor* γ_2 affects the distillation performance over the Amazon Reviews dataset. We tune the value of γ_2 from 0.1 to 1.0, with results are shown in Figure 4. We find that the optimal value of γ_2 generally lies in the range of 0.2 - 0.5. The trend of accuracy is different in the domain “DVD” is different from those of the remaining three domains. This means the benefits from transferable knowledge of the meta-teacher vary across domains.

5 Conclusion and Future Work

In this work, we propose the Meta-KD framework which consists of meta-teacher learning and meta distillation to distill PLMs across domains. Experiments on two widely-adopted public multi-domain datasets show that Meta-KD can train a meta-teacher to digest knowledge across domains to help better teach in-domain students. Quantitative evaluations confirm the effectiveness of Meta-KD and also show the capability of Meta-KD in the settings where the training data is scarce i.e. there is no or few in-domain data. In the future, we will examine the generalization capability of Meta-KD in other application scenarios and apply other meta-learning techniques to KD for PLMs.

Acknowledgements

This work is supported by Open Research Projects of Zhejiang Lab (No. 2019KD0AD01/004). Any

opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

- Haoli Bai, Jiaxiang Wu, Irwin King, and Michael R. Lyu. 2020. Few shot network compression via cross distillation. In *AAAI*, pages 3203–3210.
- Sergey Bartunov, Jack W. Rae, Simon Osindero, and Timothy P. Lillicrap. 2020. Meta-learning deep energy-based memory models. In *ICLR*.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.
- Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. 2019. Multi-domain gated CNN for review helpfulness prediction. In *The World Wide Web Conference*, pages 2630–2636.
- Cen Chen, Chengyu Wang, Minghui Qiu, Dehong Gao, Linbo Jin, and Wang Li. 2021. Cross-domain knowledge distillation for retrieval-based question answering systems. In *The World Wide Web Conference*.
- Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *NAACL-HLT*, pages 602–607.
- Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. 2020a. Adabert: Task-adaptive BERT compression with differentiable neural architecture search. In *IJCAI*, pages 2463–2469.
- Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. 2020b. Online knowledge distillation with diverse peers. In *AAAI*, pages 3430–3437.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.
- Lingyun Feng, Minghui Qiu, Yaliang Li, Hai-Tao Zheng, and Ying Shen. 2021. Learning to augment for data-scarce domain bert knowledge distillation. In *AAAI*.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135.
- Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic model-agnostic meta-learning. In *NeurIPS*, pages 9537–9548.

- Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *ICML*, pages 1602–1611.
- Yunchao Gong, Liu Liu, Ming Yang, and Lubomir D. Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *CoRR*.
- Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural network. In *NeurIPS*, pages 1135–1143.
- Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
- Mengting Hu, Yike Wu, Shiwan Zhao, Honglei Guo, Renhong Cheng, and Zhong Su. 2019. Domain-invariant feature distillation for cross-domain sentiment classification. *CoRR*.
- Forrest N. Iandola, Albert E. Shaw, Ravi Krishna, and Kurt Keutzer. 2020. Squeezebert: What can computer vision teach NLP about efficient neural networks? *CoRR*.
- Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. 2019. Metadistiller: Network self-boosting via meta-learned top-down distillation. In *ICML*.
- Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning. In *NeurIPS*, pages 1818–1828.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling BERT for natural language understanding. *CoRR*.
- Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. 2020. Few sample knowledge distillation for efficient network compression. In *CVPR*, pages 14627–14635.
- Benlin Liu, Yongming Rao, Jiwen Lu, Jie Zhou, and Cho jui Hsieh. 2020. Metadistiller: Network self-boosting via meta-learned top-down distillation. In *ECCV*.
- Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. 2019. Mkd: a multi-task knowledge distillation approach for pretrained language models. *CoRR*.
- Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *ACL*, pages 1–10.
- Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. 2017. Data-free knowledge distillation for deep neural networks. *CoRR*.
- Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? In *EMNLP*, pages 479–489.
- Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, pages 1345–1359.
- Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *CVPR*, pages 2239–2247.
- Shuke Peng, Feng Ji, Zehao Lin, Shaobo Cui, Haiqing Chen, and Yin Zhang. 2020. MTSS: learn from multiple domain teachers and become a multi-domain dialogue expert. In *AAAI*, pages 8608–8615.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.
- Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *ICLR*.
- Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2017. Knowledge adaptation: Teaching to adapt. *CoRR*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*.
- Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *ICML*, pages 1842–1850.
- Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019a. How to fine-tune BERT for text classification? In *CCL*, pages 194–206.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019b. Patient knowledge distillation for BERT model compression. *CoRR*.
- Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In *ACL*, pages 2158–2170.
- Ben Tan, Yu Zhang, Sinno Jialin Pan, and Qiang Yang. 2017. Distant domain transfer learning. In *AAAI*, pages 2604–2610.
- Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In *ICLR*.

- Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pages 1195–1204.
- Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *CoRR*.
- Karen Ullrich, Edward Meeds, and Max Welling. 2017. Soft weight-sharing for neural network compression. In *ICLR*.
- Joaquin Vanschoren. 2018. Meta-learning: A survey. *CoRR*.
- Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. In *NeurIPS*, pages 1–12.
- Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. In *EMNLP*, page 3094–3104.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, pages 1112–1122.
- Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *WSDM*, pages 690–698.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.
- Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. In *ICLR*.
- Zhiqian Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, Suhang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *ACL*, pages 3014–3024.
- Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 7130–7138.
- Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *CoRR*.
- Shan You, Chang Xu, Chao Xu, and Dacheng Tao. 2017. Learning from multiple teacher networks. In *SIGKDD*, pages 1285–1294.
- Sanqiang Zhao, Raghav Gupta, Yang Song, and Denny Zhou. 2019. Extreme language model compression with optimal subwords and shared projections. *CoRR*.