Prompt-based Distribution Alignment for Domain Generalization in Text Classification

Chen Jia[†‡] and **Yue Zhang**[‡§]
[†]Fudan University, China
[‡]School of Engineering, Westlake University, China
[§]Institute of Advanced Technology, Westlake Institute for Advanced Study, China
{jiachen,zhangyue}@westlake.edu.cn

Abstract

Prompt-based learning (a.k.a. prompting) achieves high performance by bridging the gap between the objectives of language modeling and downstream tasks. Domain generalization ability can be improved by prompting since classification across different domains can be unified into the prediction of the same set of label words. The remaining challenge for domain generalization by prompting comes from discrepancies between the data distribution of different domains. To improve domain generalization with prompting, we learn distributional invariance across source domains via two alignment regularization loss functions. The first is vocabulary distribution alignment, which uses a Kullback-Leibler divergence regularization on source-domain vocabulary distributions. The second is feature distribution alignment, which uses a novel adversarial training strategy to learn domain invariant representation across source domains. Experiments on sentiment analysis and natural language inference show the effectiveness of our method and achieve state-of-the-art results on six datasets.

1 Introduction

Pretrained language models (PLMs) have achieved promising results on a range of natural language processing (NLP) tasks (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020). The framework of tuning a PLM by task data has achieved competitive performance (Devlin et al., 2019). However, it suffers significant performance degradation when the tuned PLM is directly applied to out-of-domain examples (Gururangan et al., 2020). To tackle the problem of domain shift where the training set and the test set come from different data distributions, unsupervised domain adaptation (Pan and Yang, 2009; Mansour et al., 2009) uses unlabeled target data cooperated with the labeled source data for training. However, in many real-world systems, access to unlabeled data in the target domain is also impossible. This paper focuses on **domain generalization (DG)**, the practical and challenge setting, where a model trained on multiple source domains can be directly generalized to a target domain without any labeled or unlabeled data from the target domain (Blanchard et al., 2011; Muandet et al., 2013).

Prompt-based learning (a.k.a. Prompting) (Brown et al., 2020; Gao et al., 2021; Han et al., 2021a; Liu et al., 2021a) makes better use of pretrained knowledge by bridging the gap between objectives of language modeling and downstream task. In particular, prompt-based text classification uses an identical projection from the probability distribution on label worlds to the probability distribution on classification classes and has achieved state-of-the-art results (Gao et al., 2021; Han et al., 2021b; Hu et al., 2022).

Intuitively, DG benefits from prompting in that different domains can share a unified set of label words, and thus no additional parameters such as output layers can carry domain-specific information to hinder the generalization performance for an unseen domain. However, there remains a crucial challenge for directly using prompting for DG. As shown in Figure 1 (*middle right*), different domains have intrinsically different feature distributions, and instances from different domains have different predicted vocabulary distributions (*top left*). For example, the preferable positive set of label words for book reviews can consist of "helpful" and "well-written" more frequently, but for film reviews, it can consist of "amazing" and "real", etc more frequently.

To tackle the above challenge, we propose two regularization loss functions to better align different domains, so that the gap between source domains and a new target domain can be reduced. The first alignment loss is **vocabulary distribution alignment (VDA)**, which is used to reduce the divergence of predicted vocabulary distribu-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10147 - 10157

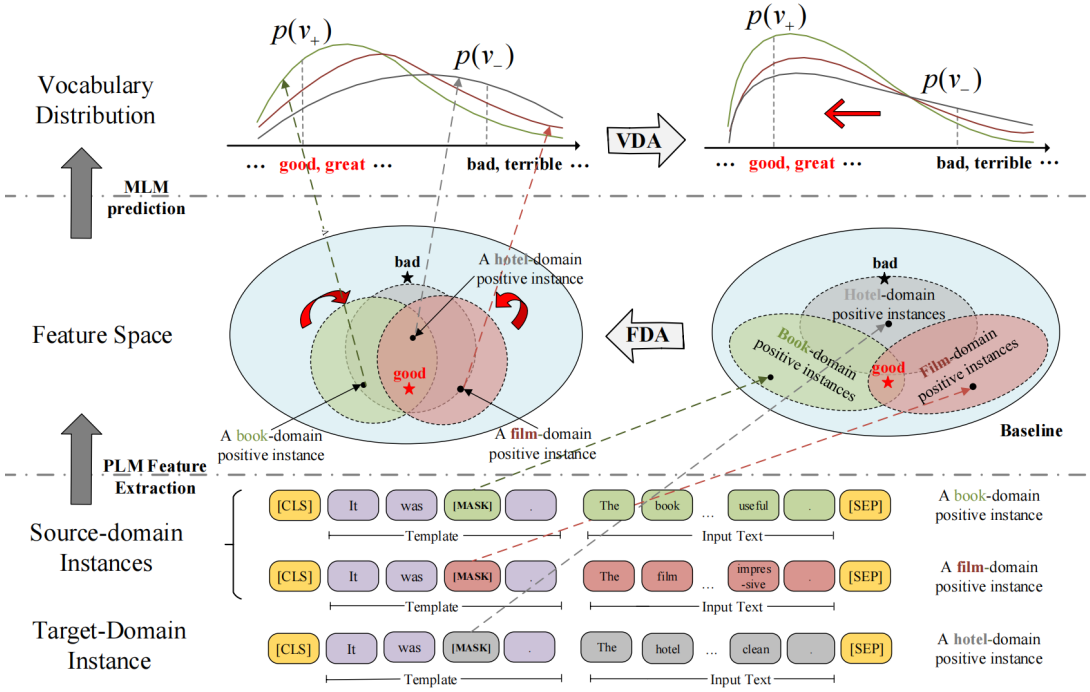


Figure 1: Distribution alignment for sentiment analysis. Positive instances from different domains tend to cluster into different regions in the feature space of prompting baseline (*middle right*). FDA makes the feature distributions of different domains similar (*middle*); VDA further reduces the distance of vocabulary distribution of instances from different domains (*top*). As a result, generalization performance of predicting the same set of label words, e.g., $\{v_+, v_-\}$ can be improved.

tions of different domains. To achieve this goal, we use Kullback-Leibler divergence regularization between predicted vocabulary distributions of different source domains in the same category. As a result, the optimized distribution alignment for the predicted vocabulary can improve the generalization performance for unseen domains, as shown in Figure 1 (*top*).

The second alignment loss is **feature distribution alignment (FDA)**, which is used to learn a generalized feature space corresponding to the predicted token for different domains. To achieve this goal, we use an adversarial training strategy (Goodfellow et al., 2014) to reduce the domain discrepancy (Ben-David et al., 2007). Different from traditional domain adversarial training for DA (Ganin et al., 2016), we learn the domain-invariant representation for each category across source domains. As a result, the optimized feature distribution alignment across source domains can improve DG performance, as shown in Figure 1 (*middle*).

We conduct experiments on three sentiment analysis datasets and three natural language inference datasets. Results show that our method can effectively learn distribution alignment and achieve the best results under both leave-one-domain-out evaluation and cross-dataset evaluation settings. To our knowledge, we are the first to use regularization to improve DG for prompting and the first to simultaneously learn domain invariance over representation and predicted probability in DG. The code will be released at https://github.com/jiachenwestlake/PDA.

2 Related Work

Prompt-based learning. Adapting the PLMs for downstream tasks via fine-tuning has become a dominant framework for NLP in recent years (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020). Prompt-based learning applies a fixed function to condition the model, so that the language model gets additional instructions to perform the downstream task. Prompt tuning the PLMs with manually designed prompts has achieved promising results on few-shot classification tasks such as sentiment analysis and natural language inference (Gao et al., 2021; Liu et al., 2021c). However, designing prompting function is challenging and requires heuristics. To this end, recent work propose to apply prompts as learnable parameters, such as

soft prompts (Lester et al., 2021; Vu et al., 2021; Gu et al., 2021), P-tuning V2 (Liu et al., 2021b) and prefix tuning (Li and Liang, 2021). Prompts capture task-specific knowledge with much smaller additional parameters than its competitors, such as Adapter (Wang et al., 2021; Pfeiffer et al., 2021) and LoRA (Hu et al., 2021). However, relatively little work has considered that the domain discrepancy can affect the performance of prompt-based learning. This paper focuses on improving the domain generalization ability for prompt-based learning via a domain invariance learning strategy.

Domain generalization. Different from domain adaptation (DA) (Pan and Yang, 2009; Mansour et al., 2009), DG improves out-of-domain (OOD) robustness with no need to explicitly know the data distribution in target domain (Blanchard et al., 2011; Muandet et al., 2013). While a plethora of DG methods have been proposed for objective recognition during the last decade (Li et al., 2017; Gulrajani and Lopez-Paz, 2020), invariance learning has shown high success and has become a prevalent approach for DG. Such algorithms include domain adversarial training (Li et al., 2018; Albuquerque et al., 2020; Xiao et al., 2021); Deep CORAL (Sun and Saenko, 2016), which optimizes the second-order statistics over feature space and IRM (Arjovsky et al., 2019), which tackles domain shift with intrinsic relationship between feature representation and labeling prediction. In contrast to these work, we tackle DG via simultaneously learning the domain invariance for both feature representation and predicted probability using novel regularization methods with prompting. Besides, recent work on DA with PLMs use domain-invariant feature regularization based on adversarial fine-tuning (Vernikos et al., 2020; Wang et al., 2020; Wu and Shi, 2022), which are difficult to directly apply for DG. In terms of prompting for classification on OOD instances, a recent work PADA (Ben-David et al., 2022) first generates a prompt for each instance and then applies the example-specific prompt to a T5 model for classification. Orthogonal to this example-based method, we aim to tackle a more general problem setting, generalization for any data distribution.

3 Background

3.1 Prompting for Text Classification

A typical framework of prompt-based text classification is shown in Figure 2, where the sentiment

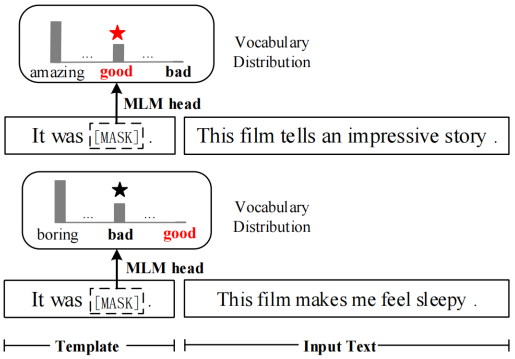


Figure 2: An example of prompt-based sentiment analysis for film reviews.

Setting	Src data	Tgt data	Data Distributions
Standard	-	Labeled	$p^T(x, y)$
Domain adaptation	Labeled	Unlabeled	$p^S(x, y), p^T(x)$
Domain generalization	Labeled	-	$p^S(x, y)$

Table 1: Comparison between the settings of standard training, domain adaptation and domain generalization.

orientation of a film review can be judged by the probability of sentiment words.

Given a sequence of tokens $\mathbf{s} = [x_1, \ldots, x_t]$, the prompted input $\widetilde{\mathbf{s}}$ is represented as:

$$\widetilde{\mathbf{s}} := [\text{CLS}]\ \text{tmp}([\text{MASK}])\ \mathbf{s}\ [\text{SEP}] \quad (1)$$

where $\text{tmp}([\text{MASK}])$ denotes prompting template.

Feeding $\widetilde{\mathbf{s}}$ into a PLM $\mathcal{M}$, we obtain hidden representation at $[\text{MASK}]$, $\mathcal{M}(\widetilde{\mathbf{s}}) \in \mathbb{R}^H$, where H represents the hidden dimension. Then, a linear classifier $f : \mathbb{R}^H \to \mathcal{C}_{|\mathcal{V}|}$ parameterized by $\boldsymbol{\theta}^f \in \mathbb{R}^{H \times |\mathcal{V}|}$ outputs the probability over vocabulary $\mathcal{V}$, where $\mathcal{C}_{|\mathcal{V}|}$ denotes a $(|\mathcal{V}| - 1)$-simplex.

Prompting for binary text classification defines a set of label words $\mathcal{V}_p = \{v_+, v_-\} \subset \mathcal{V}$, e.g., $v_+ := \text{good}$, $v_- := \text{bad}$. Predicted probability is:

$$p(v_+|\mathbf{s}, f, \mathcal{M}) = \frac{\exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}^f_{v_+})}{\sum_{v \in \{v_+, v_-\}} \exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}^f_{v})} \quad (2)$$

where $\boldsymbol{\theta}^f_{v_+}$ and $\boldsymbol{\theta}^f_{v_-}$ are parameters corresponding to v_+ and v_-, respectively. Similarly, the probability for negative category is $p(v_-) = 1 - p(v_+)$.

Given training data $\hat{S} = \{(\mathbf{s}_i, y_i)\}_{1 \le i \le m}$ with cardinality m, the objective is a cross-entropy loss,

$$\mathcal{L}_{class} = -\sum_{i=1}^{m} [y_i \log p(v_+) + (1 - y_i) \log p(v_-)] \quad (3)$$

3.2 Domain Generalization

Transfer learning aims to tackle the problem of domain shift, i.e., test data are drawn from different

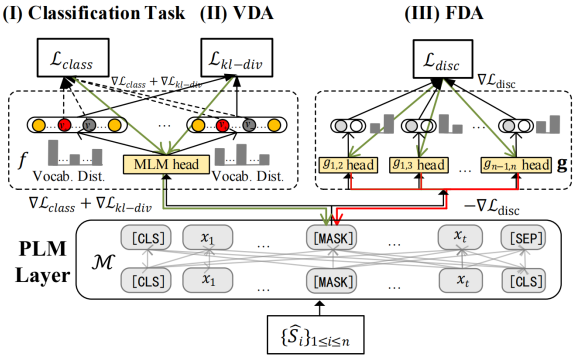


Figure 3: Overall structure for the proposed method. Green arrows indicate positive gradient (minimize the corresponding loss), while red arrows indicate negative gradient (maximize the corresponding loss).

distribution from the training data. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the output space. We define a domain as a *distribution* $p(x)$ on $\mathcal{X}$ or a joint distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$.

Domain adaptation uses unlabeled target data $x^T \sim p^T(x)$ and labeled source data $(x^S, y^S) \sim p^S(x, y)$ for training (Mansour et al., 2009; Ben-David et al., 2010). In contrast, domain generalization aims to solve a more practical and challenging problem setting, where both labeled and unlabeled target data is unavailable during training, and thus the available data for training is only the source labeled data $(x^S, y^S) \sim p^S(x, y)$ (Blanchard et al., 2011; Muandet et al., 2013). A comparison between DA and DG w.r.t. the training data condition is given in Table 1.

4 Prompt-based Distribution Alignment

The overall framework of our method is shown in Figure 3. The proposed method consists of three parts: (I) Prompt-tuning for text classification (§3.1). (II) vocabulary distribution alignment (VDA) via Kullback-Leibler divergence regularization (§4.1) and (III) feature distribution alignment (FDA) via labeling awareness domain adversarial training (§4.2).

4.1 Vocabulary Distribution Alignment

In order to improve DG with prompting, we first aim to learn domain invariance w.r.t. the predicted vocabulary distribution. A regularization method is proposed to reduce the divergence of vocabulary distributions between different source domains. In contrast to labeling regularization (Wang et al., 2018), our method is more general and more robust in two points: (i) vocabulary distribution from PLMs involves rich domain information; (ii) our method can invariably adjust to different tasks with different categories or different label words.

We first recall the computing of predicted probability in §3.1. Given an prompted input $\widetilde{\mathbf{s}}$, the predicted probability over the vocabulary $\mathcal{V}$ is:

$$p(v|\mathbf{s}, f, \mathcal{M}) = \frac{\exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}_v^f)}{\sum_{v' \in \mathcal{V}} \exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}_{v'}^f)} \quad (4)$$

where $\boldsymbol{\theta}_v^f$ are the parameters w.r.t. a token $v \in \mathcal{V}$.

Noting that different categories in a classification task intrinsically have different vocabulary distributions. We thus make VDA across source domains for each category. Given training data of k categories for n source domains $\{\{\hat{S}_i^j\}_{1\le j\le k}\}_{1\le i\le n}$, the optimization objective of VDA is the Kullback-Leibler divergence between average vocabulary distributions for each source domain pair in each category,

$$\mathcal{L}_{kl-div} = \sum_{j=1}^{k} \sum_{l \neq t} \mathrm{D_{KL}} \left(\overline{p}(\mathcal{V}|\hat{\mathcal{S}}_l^j) || \overline{p}(\mathcal{V}|\hat{\mathcal{S}}_t^j) \right) \quad (5)$$

where the average vocabulary distribution for domain l in category j is represented as:

$$\overline{p}(\mathcal{V}|\hat{\mathcal{S}}_l^j) = \frac{1}{|\hat{\mathcal{S}}_l^j|} \sum_{\mathbf{s} \in \hat{\mathcal{S}}_l^j} p(\mathcal{V}|\mathbf{s}, f, \mathcal{M}) \quad (6)$$

As a result, VDA can make the probability of preferable label words across different domains more similar, as shown in Figure 1 (*top*). Thus, when using the unified set of label words for unseen target domains, DG ability can be improved.

4.2 Feature Distribution Alignment

In addition, we also learn domain-invariant representation across source domains.

Given n source domains, we equip each domain pair $(i, j), 1 \le i < j \le n$ with a domain classifier $g_{ij} : \mathbb{R}^H \to \{0, 1\}$ parameterized by $\boldsymbol{\theta}^{g_{ij}} \in \mathbb{R}^{H \times 2}$ to differentiate two domains. We assume that the i-th domain and the j-th domain $(i < j)$ are categorized by $y = 0$ and $y = 1$, respectively. We denote all the $n(n-1)/2$ domain classifiers as $\mathbf{g} = \{g_{ij}\}_{1\le i<j\le n}$.

Domain discrimination. Following (Ganin et al., 2016), given a training instance $\mathbf{s}$ from either the i-th or j-th domain, the probability of the i-th domain is:

$$p(y = 0|\mathbf{s}, g_{ij}, \mathcal{M}) = \frac{\exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}_i^{g_{ij}})}{\sum_{k \in \{i,j\}} \exp(\mathcal{M}(\widetilde{\mathbf{s}})^\top \boldsymbol{\theta}_k^{g_{ij}})} \quad (7)$$

Given the training samples from n source domains $\{\hat{S}_i\}_{1\le i\le n}$, we use the cross-entropy loss to represent the objective of domain discrimination,

$$\widetilde{\mathcal{L}}_{disc}(\hat{S},\mathbf{g},\mathcal{M}) = -\sum_{i<j}^{n}\Big[\sum_{\mathbf{s}\in\hat{S}_i}\log p(y=0|\mathbf{s}) + \sum_{\mathbf{s}'\in\hat{S}_j}\log\left(1-p(y=0|\mathbf{s}')\right)\Big] \tag{8}$$

Different from traditional domain adversarial training for DA (Ganin et al., 2016), we further use labeling information of source domains and propose label-aware domain adversarial training.

Label-aware domain discrimination. Given training data of k categories for n source domains $\{\{\hat{S}_i^l\}_{1\le l\le k}\}_{1\le i\le n}$, domain discrimination is conducted on each category. Accordingly, we need $kn(n-1)$ domain classifiers, $\mathbf{G}=\cup_{l=1}^{k}\mathbf{g}^l$, where $\mathbf{g}^l=\{g_{ij}^l\}_{1\le i<j\le n}$. The objective of label-aware domain discrimination is represented as:

$$\mathcal{L}_{disc} = \sum_{l=1}^{k}\widetilde{\mathcal{L}}_{disc}(\hat{S}^l,\mathbf{g}^l,\mathcal{M}) \tag{9}$$

where each $\widetilde{\mathcal{L}}_{disc}(\hat{S}^l,\mathbf{g}^l,\mathcal{M})$ is defined in Eq. (8).

Domain adversarial training. Domain adversarial training can be seen as a two-player minimax game where the domain classifiers $\mathbf{G}$ tend to minimize the label-aware domain discrimination loss while the PLM $\mathcal{M}$ tends to maximize the loss, to make representations in category cluster across source domains. Formally, the label-aware domain adversarial training objective is represented as:

$$\max_{\mathcal{M}}\min_{\mathbf{G}}\mathcal{L}_{disc} \tag{10}$$

4.3 Learning Algorithm

Joint training objective. Given source-domain training samples, VDA and FDA objectives are optimized jointly for learning the PLM $\mathcal{M}$ and task classifier f (MLM head), formally represented as:

$$\min_{\mathcal{M},f}\left\{\mathcal{L}_{class}+\mathcal{L}_{kl-div}-\min_{\mathbf{G}}\mathcal{L}_{disc}\right\} \tag{11}$$

where the text classification objective $\mathcal{L}_{class}$, VDA objective $\mathcal{L}_{kl-div}$ and FDA objective $\mathcal{L}_{disc}$ are defined as above.

Training process. The training process is shown in Algorithm 1. In each step, given a minibatch of training data from n source domains in the same category, where each domain has β instances. The text classification task (lines 3-6) updates parameters of PLM and MLM head based on the $n\beta$ data points. The VDA task (lines 7-10) updates parameters of PLM and MLM head. The FDA task (lines 11-15) updates parameters of domain classifiers using positive gradients (line 13) while updating parameters of PLM using negative gradients (line 14). The VDA and FDA tasks equally process $2n(n-1)\beta$ data points in each step.

Algorithm 1 Training Process.

Input data: Training samples of n source domains.
Input parameters: parameters of PLM $\boldsymbol{\theta}^{\mathcal{M}}$, task classifier $\boldsymbol{\theta}^{f}$ and domain classifiers $\cup_{l=1}^{k}\{\boldsymbol{\theta}^{g_{ij}^k}\}_{1\le i<j\le n}$. **Hyperparameter:** learning rate η and combinational coefficient γ_1, γ_2.
Output: Tuned PLM $\mathcal{M}$ and the MLM head f

Training process begin
1: **while** Stopping conditions are not met **do**
2: Minibatch of $y\in\mathcal{Y}$: $\{\{\mathbf{s}_i^k,y\}_{1\le k\le\beta}\}_{1\le i\le n}$
Parameter updates using each minibatch
3: **for** $i\in\{1,\dots,n\}$ **do**
4: $\mathcal{L}_{class}\leftarrow\frac{1}{\beta}\sum_{k=1}^{\beta}\mathrm{CE}(f(\mathcal{M}(\mathbf{s}_i^k)),y_i^k)$
Minimizing text classification obj.
5: $[\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}]\leftarrow[\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}]-\eta(1-\gamma_1-\gamma_2)\nabla_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}}\mathcal{L}_{class}$
6: **end for**
7: **for** $i,j\in\{1,\dots,n\},i\neq j$ **do**
8: $\mathcal{L}_{kl-div}\leftarrow\mathrm{D_{KL}}(f(\mathcal{M}(\mathbf{s}_i))||f(\mathcal{M}(\mathbf{s}_j)))$
Minimizing the KL-divergence
9: $[\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}]\leftarrow[\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}]-\eta\gamma_1\nabla_{\boldsymbol{\theta}^f,\boldsymbol{\theta}^{\mathcal{M}}}\mathcal{L}_{kl-div}$
10: **end for**
11: **for** $i,j\in\{1,\dots,n\},i<j$ **do**
12: $\mathcal{L}_{disc}\leftarrow\frac{1}{2\beta}\sum_{k}^{\beta}[\mathrm{CE}(g_{ij}(\mathcal{M}(\mathbf{s}_i^k)),1)+\sum_{t}^{\beta}\mathrm{CE}(g_{ij}(\mathcal{M}(\mathbf{s}_j^k)),0)]$
Minimizing domain discrimination obj.
13: $\boldsymbol{\theta}^{g_{ij}^y}\leftarrow\boldsymbol{\theta}^{g_{ij}^y}-\eta\gamma_2\nabla_{\boldsymbol{\theta}^{g_{ij}^y}}\mathcal{L}_{disc}$
Maximizing domain discrimination obj.
14: $\boldsymbol{\theta}^{\mathcal{M}}\leftarrow\boldsymbol{\theta}^{\mathcal{M}}+\eta\gamma_2\nabla_{\boldsymbol{\theta}^{\mathcal{M}}}\mathcal{L}_{disc}$
15: **end for**
16: **end while**

5 Experiments

We evaluate our method on both sentiment classification and natural language inference.

5.1 Experimental Setup

Dataset. The statistics of the six datasets are listed in Table 2. The binary sentiment analysis datasets include Amazon reviews (Blitzer et al., 2007), IMDB (Thongtan and Phienthrakul, 2019) and SST-2 (Socher et al., 2013). For natural language inference, we use three datasets that consist of three categories: {entailment, neutral, contradiction}, including a smaller version of MNLI[1] Ben-David et al. (2022), SNLI (Bowman et al., 2015) and SICK (Marelli et al., 2014).

Evaluation. We use the standard leave-one-domain-out evaluation for DG (Gulrajani and Lopez-Paz, 2020) on amazon reviews and MNLI.

[1] https://github.com/eyalbd2/PADA.

Sentiment Analysis				
Dataset	**Domain**	**#Training**	**#Dev**	**#Test**
Amazon	*book* (B)	1,600	400	400
	DVD (D)	1,600	400	400
	electronics (E)	1,600	400	400
	kitchen (K)	1,600	400	400
IMDB	*movie*	6,400†	1,000	25,000
SST-2	*movie*	6,920	872	1,821
Natural Language Inference				
Dataset	**Domain**	**#Training**	**#Dev**	**#Test**
MNLI	*fiction* (F)	2,547	1,972	1,972
	government (G)	2,541	1,944	1,944
	slate (S)	2,605	1,954	1,954
	telephone (T)	2,754	1,965	1,965
	travel (T')	2,541	1,975	1,975
SNLI	general	13,000†	2,000	9,831
SICK	*image&video*	9,501	500	500

Table 2: Statistics of datasets. † indicates subset of the original dataset to build a fair upper-bound w.r.t. data size.

Task	Template	Label words
Sentiment Analysis	It was [MASK], <Text>	{(good), (bad)}
NLI	<Premise>, <Hypothesis>, [SOFT], ..., [SOFT], [MASK]	{(no, false), (yes, true), (uncertain, neutral)}

Table 3: Prompt design.

Besides, we also consider a more challenge setting, the cross-dataset evaluation, where amazon reviews and MNLI are used as the source dataset to train sentiment analysis and natural language inference models, respectively. Then, the sentiment analysis model is tested on IMDB and SST-2, while the natural language inference model is tested on SNLI and SICK. We use macro-F1 as the performance metric in each experiment.

Training details. Following Gao et al. (2021), we use manually designed prompting templates on both sentiment analysis and natural language inference datasets. The details of prompt design are listed in Table 3. We use RoBERTa$_{\text{BASE}}$ (Liu et al., 2019) as the default PLM, building the model on the OpenPrompt framework (Ding et al., 2022). The whole model is trained up to 20 epochs with a minibatch size of 4 for each category in each domain. We use AdamW with an initial learning rate of $1e^{-5}$, weight decay rate of 0.01 and warm up steps of 500 for optimization.

5.2 Leave-one-domain-out Results

We report the results of leave-one-domain-out evaluation in Table 4. The data settings for our method and all baselines except for the upper-bound are the same (source domains w/o the target domain).

Prompting outperforms fine-tuning w/o prompting by 0.8% and 1.8% on amazon reviews and MNLI respectively, which shows that prompting benefits from the unified feature space of language modeling and classification task. Besides, both of the two invariance-based regularization methods VDA and FDA can improve the prompting baseline, achieving 92.8%(amazon), 80.1%(MNLI) and 92.9%(amazon), 79.4%(MNLI), respectively. This shows that both VDA and FDA can effectively learn domain invariance to improve DG. Moreover, FDA cooperating with VDA can further improve the performance and give the best results on two datasets. This shows that FDA can effectively promote the vocabulary distributional invariance across source domains and thus improve DG performance.

Compared with other invariance-based methods, Deep CORAL (Sun and Saenko, 2016) and IRM (Arjovsky et al., 2019), we focus on vocabulary distributional invariance, which can effectively leverage rich pretrained knowledge in PLMs. Compared with the reported results of PADA (Ben-David et al., 2022) on MNLI, our method performs better, which shows the effectiveness of learning distributional invariance for DG.

5.3 Cross-dataset Results

To evaluate DG performance for more diverse distinctions in text genre and topic, we report cross-dataset evaluation results in Table 5. The training data settings for our method and all the baselines except for the upper-bound are the same (only the source dataset). Further, we include another PLM, BERT$_{\text{BASE}}$ to show the robustness of our method to different PLMs.

The results show a similar trend as the leave-one-domain-out results, and achieve more significant improvements over the baseline prompting method compared with the leave-one-domain-out evaluation, over 5% on the SICK dataset, over 3.5% on the SST-2 dataset and about 3% on the other two datasets. Besides, our method outperforms other baselines and achieves the best results on all the four datasets, which shows the effectiveness of learning domain invariance via VDA and FDA for DG.

Method			Amazon					MNLI					
Prompting	VDA	FDA	DEK→B	BEK→D	BDK→E	BDE→K	Avg.	GSTT'→F	FSTT'→G	GFTT'→S	GSFT'→T	GSTF→T'	Avg.
×	×	×	90.9	89.9	92.0	93.5	91.6	76.9	77.2	76.4	74.3	78.3	76.6
×	×	✓	91.2	90.1	91.8	93.7	91.7	77.3	78.9	74.3	75.7	78.5	76.9
✓	×	×	92.1	90.8	92.2	94.4	92.4	78.0	77.8	80.0	76.2	79.9	78.4
✓	✓	×	92.6	91.9	92.5	94.2	92.8	78.6	82.3	**80.2**	78.3	81.2	80.1
✓	×	✓	**93.1**	92.0	92.1	94.5	92.9	79.5	83.8	77.3	76.2	80.3	79.4
✓	✓	✓	92.9	**92.2**	**93.3**	**94.8**	**93.3**†	**80.8**	**85.8**	79.7	**79.4**	**83.0**	**81.7**†
Deep CORAL (Sun and Saenko, 2016)			91.9	91.3	90.9	93.5	91.9	77.6	76.3	78.2	75.3	78.2	77.1
IRM (Arjovsky et al., 2019)			92.3	91.2	91.9	94.5	92.5	78.1	75.2	79.4	76.2	79.2	77.6
PADA‡ (Ben-David et al., 2022)			86.8	86.9	89.0	92.6	88.8	76.4	83.4	76.9	78.9	82.5	79.6
Upper-bound (all domains)			*94.2*	*92.6*	*93.9*	*94.9*	*93.9*	*81.7*	*86.4*	*81.4*	*81.6*	*84.8*	*83.2*

Table 4: Leave-one-domain-out evaluation on amazon reviews and MNLI. ‡ the results of PADA on amazon are reproduced by ours and the results of PADA on MNLI come from the original paper. † indicates statistical significance with $p < 0.01$ by t-test when compared to all baselines.

Method			Amazon→				MNLI→			
Prompting	VDA	FDA	IMDB		SST-2		SNLI		SICK	
			$\text{RoBERTa}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$
×	×	×	88.5	85.2	86.2	84.8	75.4	64.1	53.6	52.5
×	×	✓	89.8	85.5	88.4	84.7	74.9	65.2	57.2	54.7
✓	×	×	89.4	86.0	87.6	85.1	76.6	64.8	56.7	55.2
✓	✓	×	90.8	86.8	90.9	86.3	78.7	67.0	58.2	56.2
✓	×	✓	91.2	87.2	89.2	85.9	77.2	66.4	60.5	57.8
✓	✓	✓	**92.1**†	**88.5**†	**91.3**†	**86.8**†	**79.3**†	**67.6**†	**62.0**†	**60.4**†
Deep CORAL (Sun and Saenko, 2016)			89.8	85.8	87.6	85.7	77.3	65.5	57.0	56.3
IRM (Arjovsky et al., 2019)			89.0	84.8	86.7	84.2	76.2	65.8	58.7	57.0
Upper-bound (target dataset)			*94.3*	*92.3*	*94.3*	*90.7*	*88.2*	*83.0*	*90.2*	*89.6*

Table 5: Cross-dataset evaluation on four datasets. † indicates statistical significance with $p < 0.01$ by t-test when compared to all baselines.

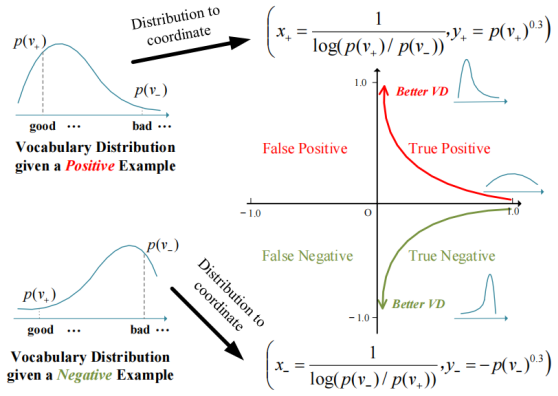
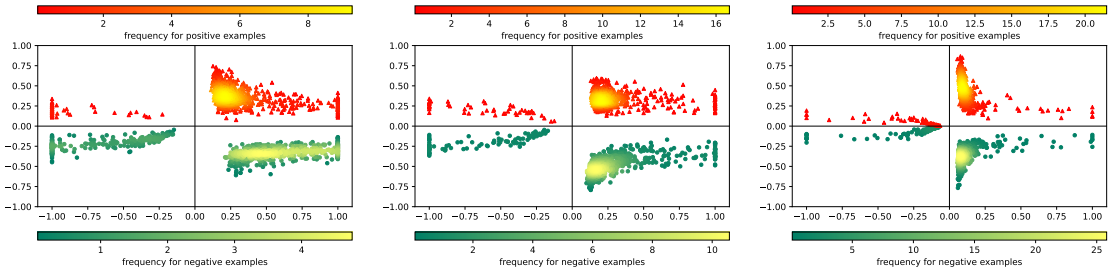


Figure 4: Techniques for visualizing the vocabulary distribution used in Figure 5.

5.4 Visualizing Distribution Alignment

Visualizing vocabulary distribution. To intuitively show how VDA and FDA can learn domain-generalized vocabulary distribution, we project the probability of label words, e.g., {good, bad} over total vocabulary into a 2-dimensional space via the coordinate computation shown in Figure 4. In particular, the x-axis (horizontal) indicates divergence between probabilities of negative and positive label words and the y-axis (vertical) indicates the probability of positive label word. Thus, x-coordinate near the origin and y-coordinate away from the origin mean that the probability of positive label word is higher while the vocabulary distribution is sharper, which can reflect the quality of vocabulary distribution. Besides, the four regions can indicate FP/FN/TP/TN as shown in Figure 4.

Based on the above computation, we visualize the sample of a hold-out domain, *book* in Figure 5. First, VDA can make scatters much nearer to the origin than the prompting baseline, which shows that VDA can effectively generalize the source-domain vocabulary distribution to unseen target domain such that the probability of negative label word tend to be lower. Furthermore, the composition of VDA and FDA can make the vocabulary distribution sharper. This shows that collaborative



(a) Prompting (*baseline*). (b) Prompting + VDA. (c) Prompting + VDA + FDA.

Figure 5: Visualizing vocabulary distribution for instances of hold-out domain, *book* on amazon reviews.

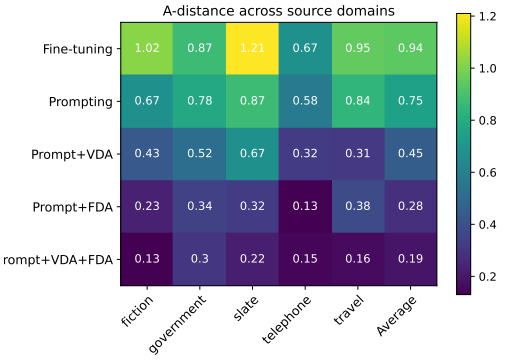


Figure 6: Average $\mathcal{A}$-distance across source domains for each hold-out domain on MNLI.

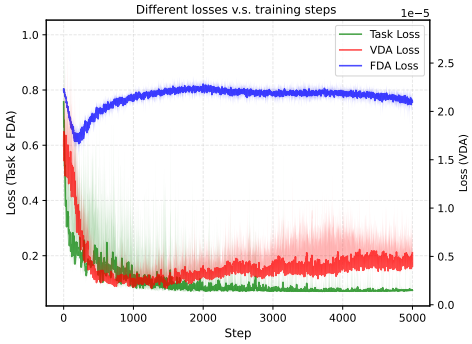


Figure 7: Task loss $\mathcal{L}_{class}$, VDA loss $\mathcal{L}_{kl-div}$ and FDA loss $\mathcal{L}_{disc}$ against training step on amazon reviews.

effects of VDA and FDA can generalize the vocabulary distribution to unseen target domain.

Domain invariance learning. We visualize the proxy $\mathcal{A}$-distance (Ben-David et al., 2007) for leave-one-domain-out evaluation. As shown in Figure 6, the feature space with prompting achieves better invariance than fine-tuning w/o prompting, which supports the motivation of using prompting for domain generalization. Further, both VDA and FDA can reduce domain discrepancy upon the prompting baseline, which shows that both VDA via KL-divergence regularization and FDA via domain adversarial training can improve domain invariance learning. Furthermore, combination of VDA and FDA can further reduce the average domain discrepancy. This shows the effectiveness of using VDA and FDA simultaneously for reducing domain discrepancy for DG.

5.5 Analysis on Training Tasks

Training procedure. We show the trend of loss for each subtask in our method against training step in Figure 7, the loss of classification task and VDA quickly reach low plateaus with small fluctuations.

Method	Amazon	MNLI	$\overline{\Delta}$
Prompting (*baseline*)	92.4	78.4	-
+ VDA	92.8	80.1	0.0
+ VDA (label-agnostic)	92.1	77.8	−1.5
+ FDA	92.9	79.4	0.0
+ FDA (label-agnostic)	92.6	78.6	−0.6
+ VDA + FDA	93.3	81.7	0.0
+ VDA + FDA (label-agnostic)	92.6	80.4	− 1.0

Table 6: Importance of labeling awareness for distribution alignment on domain generalization.

While the loss of FDA via domain adversarial training descents with a similar trend as other tasks at the beginning 0∼200 steps. This is because at the beginning, the domain discriminators are stronger than the feature embedding. However, with the gradient ascent, the domain classification loss then converges to a much higher stable value than loss of other tasks.

The importance of labeling awareness. As listed in Table 6, when conducting distribution alignment without specially computing in each category (a.k.a. labeling agnostic), both VDA and FDA suffer large descents of -1.5% and -0.6%, respectively. This shows that labeling awareness is important for

learning domain invariance for DG. Interestingly, the labeling agnostic VDA even becomes lower than the prompting baseline by -0.6%, which is because vocabulary distributions are distinct between different classification categories, thus aligning vocabulary distribution in different categories across domains can hurt the performance.

6 Conclusion

We investigated how to improve out-of-domain (OOD) robustness for prompt-based learning, by learning domain invariance via vocabulary distribution alignment and feature distribution alignment with prompting. Experiments show that we achieve the best results on six datasets of sentiment analysis and natural language inference under both the standard leave-one-domain-out evaluation setting and a novel cross-dataset evaluation setting compared with a range of strong baselines. Moreover, the proposed distribution alignment method can be seen as a general regularization technique for domain generalization beyond the text classification task.

Limitations

Our work focuses on the text classification task to investigate how to use invariance learning to improve out-of-domain generalization with prompting. However, the proposed distribution alignment method can be a general approach for domain generalization in NLP. It can be the future work to consider more tasks beyond text classification.

Another limitation is the computational complexity of the proposed FDA, which is $\mathrm{poly}(k \cdot n \cdot n)$. Although it can only slightly increase the training time in our experiments, but there could be a tradeoff between the training time and the classification accuracy when the number of source domains is much larger.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. We gratefully acknowledge funding from the National Natural Science Foundation of China (NSFC No. 61976180). Yue Zhang is the corresponding author.

References

Isabela Albuquerque, Joao Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. 2020. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. 2007. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19:137.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, 24:2178–2186.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. Openprompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*.

Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, et al. 2021a. Pre-trained models: Past, present and future. *AI Open*.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021b. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5542–5550.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision*, pages 624–639.

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

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021c. Gpt understands, too. *arXiv:2103.10385*.

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*.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237,

New Orleans, Louisiana. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer.

Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414. Association for Computational Linguistics.

Giorgos Vernikos, Katerina Margatina, Alexandra Chronopoulou, and Ion Androutsopoulos. 2020. Domain adversarial fine-tuning as an effective regularizer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3103–3112.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3094–3104.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. 2018. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *NAACL-HLT*.

Hui Wu and Xiaodong Shi. 2022. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2438–2447.

Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees Snoek. 2021. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning*, pages 11351–11361. PMLR.