Learning Meta Soft Prompt for Few-Shot Language Models

Jen-Tzung Chien* Ming-Yen Chen* Jing-Hao Xue†
* Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
† Department of Statistical Science, University College London, London, United Kingdom

***Abstract*—Prompt-based learning is powerful to utilize the large-scaled pre-trained language model (PLM) for language understanding where the input sentences are augmented by either adding the hard prompt using word tokens or the soft prompt in a form of trainable tokens. However, the learned soft prompt in training domain may not really help a frozen PLM to handle domain shift in test domain. This paper presents an approach to incorporate meta learning into domain adaptation to train new soft prompt which sufficiently generalizes the frozen PLM to a number of domains. The meta soft prompt is then developed for few-shot unsupervised domain adaptation where a frozen PLM can be quickly adapted to a target domain. This soft prompt is optimized according to meta learning where the domain adaptation loss and the prompt-based classification loss are jointly minimized. The experiments on multi-domain natural language understanding show the benefits of the proposed meta soft prompt in pre-trained language model by using BERT under the few-shot setting.**

***Index Terms*—Meta learning, few-shot learning, prompt tuning, domain adaptation, language model**

I. Introduction

Traditionally, deep neural networks were trained by supervised learning and have achieved desirable performance in a variety of speech and language tasks [1]. Recently, self-supervised or unsupervised schemes have been successfully developed to improve the end performance even without additional training data. A famous example is the bidirectional encoder representations from transformers (BERT) [2] which is a pre-trained language model [3], [4] based on transformer [5] learned from a large corpus based on the unsupervised tricks consisting of the masked language modeling and the next sentence prediction. PLMs have been further extended to boost the performance of challenging tasks by enhancing the model capacity with very limited controlled parameters through utilizing the adapter learning [6] or model reprogramming [7]. With the emerging of the powerful generative pre-trained transformer like GPT-3 [8] or ChatGPT, it becomes crucial to conduct the cutting-edge research on how to utilize the pre-trained model for various downstream tasks via prompt-based learning or tuning. In general, the previous PLMs did not handle the issue that specific domain of input sentences is unseen and far from those in training sentences. The performance of PLMs will drop drastically. This phenomenon known as the domain shift is common in practice [9], [10]. The larger the domain shift, the higher the influence on the end performance. In addition, collecting a large amount of labeled data over a variety of domains is difficult and time-consuming. The unknown domain is hard to predict.

To activate fast learning in multi-domain language modeling, this study focuses on zero-shot or few-shot domain adaptation for prompt-based learning where there is few data or even no training data provided in target domain. This paper further copes with the issue of sub-optimal performance in domain adaptation which is caused by inconsistent objectives in both pre-training and fine-tuning stages [11], [12]. In order to address the aforementioned issues, this work proposes the so-called meta soft prompt to enhance the adaptability of PLM to an incoming target domain. In addition to domain adaptation [13], this method is designed by considering two learning paradigms. The first one is the soft prompt learning which converts the actual word tokens in hard prompt into some real-valued vectors as the continuous tokens in augmented input sentence. Prompt optimization is based on gradient updates by using the labeled data [14], [15]. Compared to the hard prompt, soft prompt looks more attractive since both specific domain expertise and trial-and-error process can be avoided [16], [17], [18]. The second one is the model-agnostic meta learning (MAML) [19], [20] which is performed to address the adaptability of soft prompt across multiple domains. A gradient-based meta learning algorithm [21], [22] is implemented to estimate the parameters of meta learner through a set of task-specific learners for individual learning over various meta-training tasks. In this paper, each individual meta-training task is treated as an unsupervised domain adaptation problem which is tackled by learning the soft prompt from source-domain labeled data as well as target-domain unlabeled data. After completing the meta training across diverse meta-training tasks, the trained meta soft prompt is feasible to smoothly adapt to any incoming target domain with a few unlabeled data. In the experiments, the competitive result is obtained by using the proposed prompt-based language model for multi-domain sentiment classification.

II. Soft Prompt & Domain Adaptation

A. Prompt-based language model

Pre-trained language models (PLMs) have shown convincing results in recent years. To leverage the benefit of PLM, the task-specific layer or head was added on top of PLM to adjust the PLM parameters to a specific downstream task. More recently, prompt-based learning is recognized as a powerful method to improve the fine-tuning of PLM after GPT-3 [8]

has been publicly released. This method allows training a model with a cloze-style input sentence which adds a textual string prompt to original sentence that has some unfilled slots. The resulting language model needs to predict the word for unfilled slots instead of predicting the class label. Due to the success of GPT-3, such a learning scheme was further introduced to employ smaller PLMs [16], [17], [18] by fine-tuning the pre-trained model. Using this approach, hard prompt is formed by a string of "discrete" word tokens from the vocabulary. But, determining a suitable prompt for a specific domain or task requires the domain expertise and needs many trial-and-errors. Although hard prompt template is fixed, the PLM and the masked language model (MLM) head should be fine-tuned for a downstream task. Alternatively, the soft prompt was constructed in the continuous space with good adaptability. Since the soft prompt $(v_1 v_2 \ldots)$ can be easily updated by gradient descent, the optimal prompt is obtained in a handcraft-free way. Using soft prompt, the parameters of PLM and MLM head are frozen. Only the soft prompt tokens are estimated. Relative to fine-tuning and hard prompt, the number of controlled parameters in soft prompt is greatly reduced. Figure 1 compares the hard (left) and soft (right) prompt-based language models.

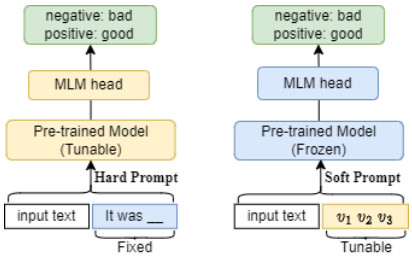

Fig. 1. Hard and soft prompts using pre-trained models.

B. Domain adaptation

In general, fine-tuning, hard prompt or soft prompt using PLM only works well for a specific test domain, and could not easily handle the diverse input sentences which may come from a variety of domains in practical scenarios. As a result, the researches on unsupervised domain adaptation (UDA) [23], [24] have become crucial for handling multi-domain tasks. UDA aims to mitigate the issue of domain shift caused by different distributions from source and target data where the class labels of adaptation data in target domain are missing. Previously, UDA was employed in BioBERT [25] where the pre-training phase was continued by adaptation using domain-specific data. In [26], UDA was performed by simultaneously running the domain-specific pre-training as well as task-specific pre-training by using task-specific unlabeled data [27]. The learned representation was closer to task distribution. In [28], UDA through language model was recently proposed and implemented by combining MLM task and downstream task where the model robustness and the sample efficiency were improved during the adaptation to target domain. Nevertheless, there is existing a difference between the pre-training objective for PLM task and the fine-tuning objective for downstream task, which may constrain the utilization of knowledge based on PLM [11]. This study introduces UDA to adapt the soft prompt across different domains via meta learning in accordance with a consistent hybrid objective for domain adaptation and prompt learning.

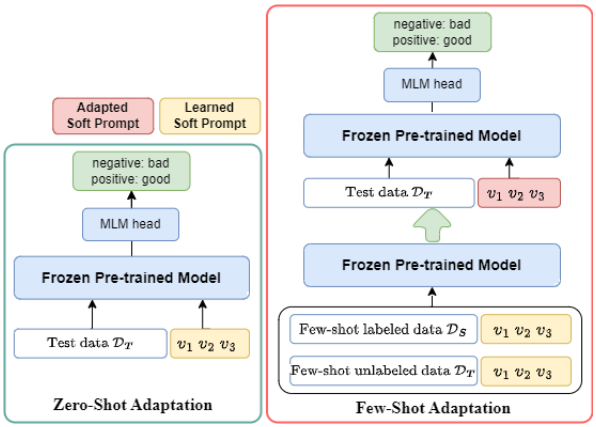

Fig. 2. Soft prompt language models with zero-shot and few-shot domain adaptation where the frozen PLM is used.

C. Joint soft prompt and few-shot domain adaptation

In this paper, soft prompt learning is collaborated with unsupervised domain adaptation where zero-shot and few-shot settings are considered as illustrated in Figure 2 where the frozen PLM is utilized and the labeled data in target domain $\mathcal{D}_T$ are missing. The left-hand-side subfigure shows the zero-shot adaptation where there is no additional adaptation data. Only the test sentence in target domain is provided. Soft prompt is updated by using test data. On the other hand, the right-hand-side subfigure displays the scenario of joint soft prompt estimation and few-shot domain adaptation. In this scenario, there are two stages where the frozen PLM is both applied. The first stage is to learn soft prompt where the labeled data in source domain $\mathcal{D}_S$ and the few-shot unlabeled data in target domain $\mathcal{D}_T$ are provided for prompt-based learning. Then, the second stage is to further adapt the soft prompt by using the test sentence in target domain. To cope with multi-domain tasks, such an adaptive soft prompt is hereafter estimated through meta learning.

III. Few-Shot Learning & Meta Soft Prompt

In practical situation, the input query is usually originated from various domains. It becomes crucial to develop a general solution to soft prompt across different domains and present a fast adaptation scheme to a new domain where multi-domain tasks are handled. The underlying idea of this paper is to enhance the adaptability of soft prompt language model to a new target domain where meta learning or learning to learn is performed to acquire multi-domain knowledge from a number of tasks which are designed to solve the problem on hands.

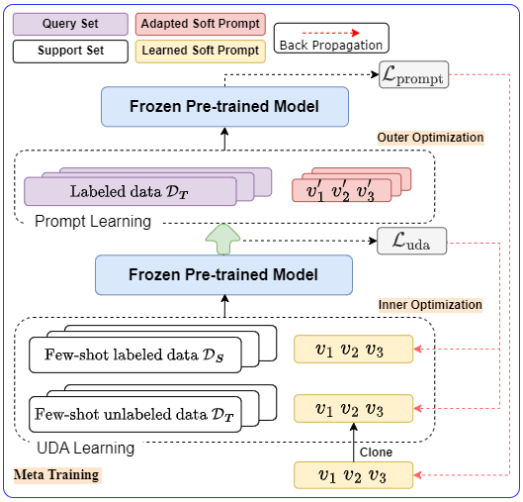

Fig. 3. Multi-domain learning for meta soft prompt language model via few-shot unsupervised domain adaptation (UDA).

A. System overview and meta learning

This study presents the model agnostic soft prompt for few-shot domain adaptation. An overview of the proposed method is shown in Figure 3. There are three parts in the learning process including input data, soft prompt and PLM using BERT. Input data contain *support set* and *query set*. In meta training, *support set* is used to update the prompt parameters of learner (v_1, v_2, v_3). *Query set* is then used to calculate the gradients for updating the parameters of learner (v_1, v_2, v_3) followed by updating the parameters of meta learner (v_1', v_2', v_3'). Soft prompt is treated as the trainable parameters, and is appended to an input sequence and a mask token denoted by [MASK]. PLM is responsible to encode the input sentence and extract the features of mask token. These features are used to calculate MLM head outputs which are employed to measure the probability distribution over vocabulary words to implement the meta soft prompt language model. Unlike traditional meta learning approaches [19], which train the whole model from a large number of few-shot classification problems to allow the model to quickly adapt to the unseen few-shot classification tasks, the proposed meta learning framework aims to tackle domain shift issue by introducing soft prompt learning conditioned on a frozen PLM. All of the meta-training tasks are designed to simulate the scenarios which will be encountered during test stage. Therefore, the frozen PLM is feasible to generalize multiple source domains to a target domain by only tuning the parameters of soft prompt.

B. Soft prompt language model

Soft prompt language model adopts a soft prompt template consisting of a set of trainable vectors that are added to the input sentence as the description of target task. The label words are defined as the highest probability words that PLM will predict. In a binary sentiment classification task, the input sentence is formed by L words as $\mathbf{x} = \{x_1, \cdots, x_L\}$ and the corresponding label is $y \in \mathcal{Y} = \{\text{positive}, \text{negative}\}$. Given the template function $T(\cdot)$, the input sentence $\mathbf{x}$ can be converted to a MLM input to BERT-based PLM, $\mathbf{x}_{\text{prompt}} = T(\mathbf{x} = \{e(\mathbf{x}), v_1, \cdots, v_k, e([\text{MASK}])\})$, where the $e(\cdot)$ is the embedding function of PLM $\mathcal{F}$. For the label words, a verbalizer $\mathcal{M} : \mathcal{Y} \mapsto \mathcal{V}^*$ is given to map the label space to a set of label words $\mathcal{V}^* \subset \mathcal{V}$, where $\mathcal{V}$ is the vocabulary of $\mathcal{F}$. Then, we can treat $\mathcal{F}$ as a function of mapping $\mathbf{x}_{\text{prompt}}$ to a vocabulary distribution of mask token as $p(\mathcal{V}_y(\mathbf{x}_{\text{prompt}})) = \mathcal{F}(\mathbf{x}_{\text{prompt}}) \triangleq \mathbf{v}_{\text{mask}}$. The conditional likelihood of the predicted label word $y^* \in \mathcal{V}_y^*$ with respect to mask token given $\mathbf{x}_{\text{prompt}}$, $p(\mathcal{V}_y^* \leftarrow [\text{MASK}]|\mathbf{x}_{\text{prompt}}, \theta)$, is yielded by

$$p(y^*|\mathbf{x}_{\text{prompt}}, \theta) = \frac{\exp(\mathbf{v}_{\text{mask}}(\mathcal{V}_y^*))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{v}_{\text{mask}}(\mathcal{V}_{y'}^*))} \tag{1}$$

where $\mathbf{v}_{\text{mask}}(\mathcal{V}_y^*)$ denotes the probability of label word $\mathcal{V}_y^*$ in the vocabulary distribution $\mathbf{v}_{\text{mask}}$. The overall training objective for soft prompt learning is expressed by a cross-entropy loss as

$$\mathcal{L}_{\text{prompt}}(\mathbf{x}, y, \theta) = -\sum_{y' \in \mathcal{Y}} y' \log p(y^*|\mathbf{x}_{\text{prompt}}, \theta) \tag{2}$$

where θ denotes the trainable parameters for soft prompt.

C. Unsupervised domain adaptation

Unsupervised domain adaptation as shown in Figure 3 implements the inner optimization that adapts the soft prompt language model to a target domain. This study presents an approach to UDA by combining soft prompt learning and masked language modeling using labeled data from source domain $\mathcal{D}_S$ and unlabeled data from target domain $\mathcal{D}_T$. The soft prompt (v_1, v_2, v_3) in UDA is updated by using the objectives for masked language model $\mathcal{L}_{\text{mlm}}$ (via cross-entropy loss for predicting mask tokens) and soft prompt $\mathcal{L}_{\text{prompt}}$ where the unlabeled data $\mathcal{D}_T$ and labeled data $\mathcal{D}_S$ are adopted, respectively. The adapted soft prompt is able to simultaneously capture the target domain information from MLM objective and the task language knowledge from prompt objective. UDA objective is built as $\mathcal{L}_{\text{uda}} = \mathcal{L}_{\text{prompt}} + \lambda \mathcal{L}_{\text{mlm}}$ with a hyperparameter λ. α is implemented via the popular optimizer using AdamW.

D. Nested optimization procedure

The overall learning and adaptation for meta soft prompt are performed via a nested loop. In the inner loop, the learners start with the parameters of meta learner and update the parameters based on $\mathcal{L}_{\text{uda}}$ by using *support set* in each meta-training task. In the outer loop, the meta learner is optimized according to the performance of those learners evaluated on *query set* in each meta-training task. The whole process is repeated until reaching convergence. Parameter updating for learners is expressed as

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta), \quad 1 \le i \le m. \tag{3}$$

$\mathcal{D}_i^{\text{sup}}$ represents the *support set* of meta-training task $\mathcal{T}_i$. Meta learner $\widehat{\theta}$ is estimated by minimizing the prompt loss over

Algorithm 1: Meta soft prompt learning & adaptation

Require: $\mathcal{D}$: multiple source and target domains
$\mathcal{D}_{\text{train}} = \{\mathcal{T}_i\}_{i=1}^m$: a set of training tasks from $p(\mathcal{T})$
α: step size of updating soft prompt
Initialize: soft prompt parameters θ
repeat
  for *each meta-training task* $\mathcal{T}_i \in \mathcal{D}_{train}$ **do**
    sample $\mathcal{D}_S \sim \mathcal{D}$ as source domain
    sample $\mathcal{D}_T \sim \{\mathcal{D} - \mathcal{D}_S\}$ as target domain
    sample *support set*
      $\mathcal{D}_i^{\text{sup}} = \{\mathbf{x}_S^j, y_S^j\}_{j=1}^{k_s} \cup \{\mathbf{x}_T^j\}_{j=1}^{k_t}$
    sample *query set* $\mathcal{D}_i^{\text{qry}} = \{\mathbf{x}_T^j, y_T^j\}_{j=1}^{k_q}$
    adapt θ to each domain θ_i' using *support set*
      $\theta_i' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta)$
    calculate gradient of $\mathcal{L}_{\text{prompt}}$ using *query set*
      $g_i \leftarrow \nabla_{\theta_i'} \mathcal{L}_{\text{prompt}}(\mathcal{D}_i^{\text{qry}}, \theta_i')$
  end
  update soft prompt $\widehat{\theta} \leftarrow \theta - \alpha \sum_{i=1}^m g_i$
until *training converged*

individual learners θ_i'. Meta learning is performed across various meta-training tasks $\mathcal{T}_i$ sampled from $p(\mathcal{T})$

$$\widehat{\theta} = \arg\min_{\theta} \sum_{\mathcal{T}_i} \mathcal{L}_{\text{prompt}}(\mathcal{D}_i^{\text{qry}}, \theta - \alpha \nabla_\theta \mathcal{L}_{\text{uda}}(\mathcal{D}_i^{\text{sup}}, \theta)). \quad (4)$$

$\mathcal{D}_i^{\text{qry}}$ represents the *query set* of $\mathcal{T}_i$. The overall learning and adaptation of meta soft prompt is shown in Algorithm 1. This procedure is a kind multi-task learning over various domains.

IV. Experiments

The proposed meta soft prompt language model was evaluated for multi-domain text classification by using the FDU-MTL dataset [29] which is known as a challenging dataset consisting of 16 domains, as shown in Table III, which were broadly categorized into Amazon product reviews and movie reviews. In the evaluation, one domain was chosen as target domain, and the remaining domains were seen as source domains. A binary sentiment classification with the classes of positive and negative for reviews was performed. Classification accuracy was reported in different sets of evaluation.

A. Evaluation for latent visualization

First, the goodness of meta soft prompt language model for unseen domain data is evaluated. Figure 4 compares 2-D visualization using t-SNE [30] for latent representation of the mask tokens by using the prompted data in 'books' reviews. Basically, the mask tokens using the proposed language model are separate in the classes of positive and negative reviews, and are domain-matching in source and target domains. In the experimental setting, the soft prompt length 5 was used and λ=0.8 was fixed.

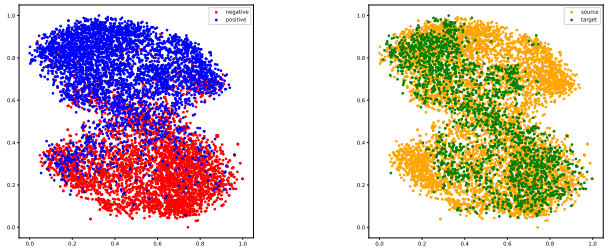

Fig. 4. Visualization on reviews of 'books' target domain using the proposed meta soft prompt. (left) Positive and negative reviews are in blue and red, and (right) source and target domain reviews are in orange and green, respectively.

TABLE I
Accuracy (%) on 4 domains of FDU-MTL dataset. Number in brackets shows the soft prompt length (S).

Domain	**FT**	**SP** (2)	**MSP** (2)	**SP** (5)	**MSP** (5)
books	80.8	**85.2**	83.6	86.8	**88.0**
DVD	79.3	83.2	**83.4**	84.4	**85.6**
electronics	79.4	82.1	**82.6**	84.8	**85.1**
kitchen	79.5	84.5	**86.2**	86.1	**87.6**

B. Evaluation for document classification

The proposed meta soft prompt (MSP) is further compared with traditional fine-tuning (FT) method and the previous soft prompt (SP) [14]. The experiments on review or document classification were conducted on four domains: books, DVD, electronics and kitchen. For a fair comparison, the number of learned parameters were set identically in all methods. For the conventional fine-tuning, the number of trainable parameters in the task-specific layer 768×2 was chosen. As a consequence, the length of soft prompt (S) was set to 2 as each prompt was encoded as a 768-dimensional word embedding. The case S=5 is also evaluated. The accuracy is shown in Table I which indicates that the proposed MSP outperforms FT and SP in various domains especially when the soft prompt length S is increased to 5.

TABLE II
Accuracy (%) and number of trainable parameters (N) for the methods with/without UDA using S=10 with 4 or 8 shots. MoE-Tr [10] requires a large-scaled trainable model.

Domain	MoE-Tr	SP	MSP	SP-4	MSP-4	MSP-8
books	90.0	87.5	**88.6**	87.9	88.7	**89.0**
DVD	89.3	86.2	**86.9**	87.0	**88.5**	88.1
electronics	90.6	87.4	**88.4**	87.9	89.2	**90.3**
kitchen	90.8	88.5	**89.8**	89.2	90.5	**90.7**
UDA	yes	no	no	yes	yes	yes
N	264M	7.68K	7.68K	7.68K	7.68K	7.68K

In Table II, the comparison is extended to a recent work called the transformer-based multi-source domain adaptation (MoE-Tr) [10], which introduced a mixture of experts by using multiple trainable PLMs in adaptation where the number of trainable parameters is N=264M. MoE-Tr performed the domain adversarial learning [31] where many shots were used. Instead, MSP has N=7.68K and only adopts one single frozen

PLM. Unsupervised domain adaptation works for SP and MSP. The improvements by increasing shot number and soft prompt length are obtained.

TABLE III
ACCURACY (%) ON 16 DOMAINS OF FDU-MTL DATASET.

Domain	MT-DNN	ASP-MTL	MAN-L2	MAN-NLL	BERT	MSP
books	82.2	84.0	87.6	86.8	87.0	**89.0**
DVD	84.2	85.5	88.1	**88.6**	85.6	88.1
electronics	81.7	86.8	87.4	88.8	88.3	**90.3**
kitchen	80.7	86.2	89.8	89.9	91.0	**90.7**
apparel	85.0	87.0	87.6	87.6	90.0	**92.0**
camera	86.2	89.2	**91.4**	90.7	90.0	90.8
health	85.7	88.2	89.8	89.4	88.3	**91.3**
music	84.7	82.5	85.9	85.5	86.8	**87.8**
toys	87.7	88.0	90.0	90.4	**91.3**	90.8
video	85.0	84.5	89.5	**89.6**	88.0	88.4
baby	88.0	88.2	90.0	90.2	**91.5**	91.3
magazine	89.5	92.2	92.5	**92.9**	92.8	90.2
software	85.7	87.2	90.4	90.9	89.3	**90.9**
sports	83.2	85.7	89.0	89.0	90.8	**91.8**
IMDb	83.2	85.5	86.6	87.0	85.8	**88.3**
MR	75.5	76.7	76.1	76.7	74.0	**80.4**
AVG	84.3	86.1	88.2	88.4	88.1	**89.5**

Next, the proposed MSP language model is compared with the previous multi-task learning methods over various domains of FDU-MTL dataset. The experiments were setup by choosing one domain as the target for testing and the remaining domains as the sources for training. After the meta training on different source domains is done, the learned meta soft prompt is adapted to a target domain by using 8 shots of unlabeled reviews. The adapted meta soft prompt is collaborated with the frozen PLM to conduct evaluation on the target domain. The results are shown in Table III. These results are consistently compared with strong baselines including the multi-task deep neural network (MT-DNN) [32], the adversarial multi-task learning (ASP-MTL) [29], the multinomial adversarial network with the least square loss (MAN-L2), the negative log-likelihood loss (MAN-NLL) [33], and the BERT model [2] which is fine-tuned on each domain. ASP-MTL [29] used abundant unlabeled samples from the target domain. Instead, MSP only adopted 8 unlabeled samples for domain adaptation. The results reveal that MSP obtains the highest averaged accuracy in most of domains.

V. CONCLUSIONS

This paper has presented the meta soft prompt language model with few-shot domain adaptation. The learned meta soft prompt was appended to input data and adapted to different domains by using the frozen pre-trained model with few-shot unlabeled samples in target domain. The results have shown that the meta soft prompt could successfully boost a frozen pre-trained model to capture domain-specific information and achieved desirable results by only training a few parameters. For future work, the proposed method could be integrated with other unsupervised domain adaptation in the inner optimization for meta learning. Furthermore, the proposed method is feasible to collaborate with not only masked language model but also sequence-to-sequence model or autoregressive model.

REFERENCES

[1] J.-T. Chien, "Deep Bayesian natural language processing," in *Proc. of Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2019, pp. 25–30.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2019, pp. 4171–4186.

[3] J.-T. Chien and C.-H. Chueh, "Latent dirichlet language model for speech recognition," in *Proc. of IEEE Spoken Language Technology Workshop*, 2008, pp. 201–204.

[4] J.-T. Chien, "Hierarchical Pitman–Yor–Dirichlet language model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1259–1272, 2015.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. of International Conference on Machine Learning*, 2019, pp. 2790–2799.

[7] C.-H. Yang, Y.-Y. Tsai, and P.-Y. Chen, "Voice2series: Reprogramming acoustic models for time series classification," in *Proc. of International Conference on Machine Learning*, 2021, pp. 11808–11819.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[9] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2020, pp. 4019–4028.

[10] D. Wright and I. Augenstein, "Transformer based multi-source domain adaptation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 7963–7974.

[11] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT understands, too," *arXiv preprint arXiv:2103.10385*, 2021.

[12] L.-J. Yang, I-P. Yeh, and J.-T. Chien, "Low-resource speech synthesis with speaker-aware embedding," in *Proc. of International Symposium on Chinese Spoken Language Processing*, 2022, pp. 235–239.

[13] W. Lin, M.-W. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[14] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 3045–3059.

[15] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5203–5212.

[16] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proc. of Conference of European Chapter of the Association for Computational Linguistics*, 2021, pp. 255–269.

[17] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2339–2352.

[18] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proc. of International Joint Conference on Natural Language Processing*, 2021, pp. 3816–3830.

[19] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. of International Conference on Machine Learning*, 2017, pp. 1126–1135.

[20] M.-Y. Chen, M. Rohmatillah, C.-H. Lee, and J.-T. Chien, "Meta learning for domain agnostic soft prompt," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[21] J.-T. Chien and W. X. Lieow, "Meta learning for hyperparameter optimization in dialogue system," *Proc. of Annual Conference of International Speech Communication Association*, pp. 839–843, 2019.

[22] J.-T. Chien and W. Lai, "Variational skill embeddings for meta reinforcement learning," in *Proc. of International Joint Conference on Neural Networks*, 2023, pp. 1–8.
[23] J.-T. Chien and Y.-Y. Lyu, "Partially adversarial learning and adaptation," in *Proc. of European Signal Processing Conference*, 2019, pp. 1–5.
[24] J.-T. Chien and C.-W. Huang, "Stochastic adversarial learning for domain adaptation," in *Proc. of International Joint Conference on Neural Networks*, 2020, pp. 1–7.
[25] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
[26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2020, pp. 8342–8360.
[27] X. Han and J. Eisenstein, "Unsupervised domain adaptation of contextualized embeddings for sequence labeling," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 4238–4248.
[28] C. Karouzos, G. Paraskevopoulos, and A. Potamianos, "UDALM: unsupervised domain adaptation through language modeling," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2579–2590.
[29] P. Liu, X. Qiu, and X. Huang, "Adversarial multi-task learning for text classification," in *Proc. of Annual Meeting of Association for Computational Linguistics*, 2017, pp. 1–10.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[31] J.-C. Tsai and J.-T. Chien, "Adversarial domain separation and adaptation," in *Proc. of International Workshop on Machine Learning for Signal Processing*, 2017, pp. 1–6.
[32] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics*, 2015, pp. 912–921.
[33] X. Chen and C. Cardie, "Multinomial adversarial networks for multi-domain text classification," in *Proc. of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2018, pp. 1226–1240.