Asymmetric Mutual Learning for Multi-source Unsupervised Sentiment Adaptation with Dynamic Feature Network

Rui Li*, **Cheng Liu**, **Dazhi Jiang**
Department of Computer Science, Shantou University
{ruili, cliu, dzjiang}@stu.edu.cn

Abstract

Recently, fine-tuning the pre-trained language model (PrLM) on labeled sentiment datasets demonstrates impressive performance. However, collecting labeled sentiment dataset is time-consuming, and fine-tuning the whole PrLM brings about much computation cost. To this end, we focus on multi-source unsupervised sentiment adaptation problem with the pre-trained features, which is more practical and challenging. We first design a dynamic feature network to fully exploit the extracted pre-trained features for efficient domain adaptation. Meanwhile, with the difference of the traditional source-target domain alignment methods, we propose a novel asymmetric mutual learning strategy, which can robustly estimate the pseudo-labels of the target domain with the knowledge from all the other source models. Experiments on multiple sentiment benchmarks show that our method outperforms the recent state-of-the-art approaches, and we also conduct extensive ablation studies to verify the effectiveness of each the proposed module.

1 Introduction

Sentiment classification (Cambria et al., 2020) aims to predict the sentiment label for each textual data automatically (Susanto et al., 2022), which is one of the most popular natural language processing (NLP) tasks with many important applications, such as social media monitoring (Ortigosa et al., 2014), market research (Jabbar et al., 2019), conversation sentiment detection (Tu et al., 2022), *etc*. Very recently, the pre-trained language models (PrLMs), *e.g.*, BERT (Devlin et al., 2019), have demonstrated significant improvements on wide-range of NLP tasks, including the sentiment classification. This framework includes two steps: the transformer-based (Vaswani et al., 2017) model is first pre-trained on large unlabeled corpus, and then fine-turned on the labeled datasets for the downstream tasks. However, as illustrated in Figure 1(Top), different domain texts often contain different subject words, and have different sentiment descriptions, which lead to decreased performance induced by domain-shift (Pan and Yang, 2010)(Figure 1(Bottom)).

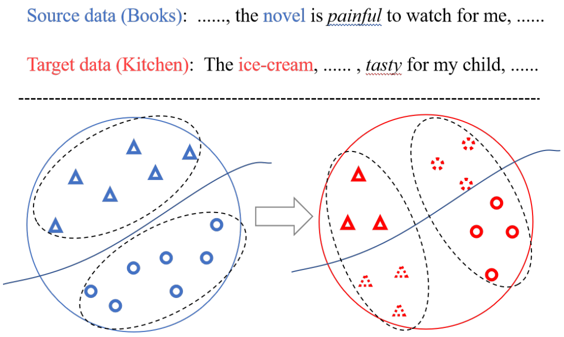
Source data (Books): ......, the novel is painful to watch for me, ...
Target data (Kitchen): The ice-cream, ... , tasty for my child, ...

Figure 1: Illustration of domain-shift for sentiment classification. Top: different domain reviews have different subject words (marked with blue) with different sentiment descriptions (marked with underlines). Bottom: the different data distributions will lead to performance degradation.

Unsupervised domain adaptation (UDA) is another hot research topic in machine learning to address the domain-shift. It aims to transfer the knowledge from the source domain to the target domain with the labeled source data and unlabeled target data. This is often achieved by minimizing a specific distance between the source and target domains to learn the shared domain-invariant features. For example, Guo et al. (2020) exploit several distance metrics (*e.g.*, Maximum Mean Discrepancy (MMD) Gretton et al., 2012, Correlation Alignment (CORAL) Sun et al., 2016) for domain adaptation in the context of text classification tasks. While, Li et al. (2017) use a domain classifier to obtain domain-invariant sentiment features via adversarial training (Ganin et al., 2016) between the source and target features.

Despite the progress of recent cross-domain sentiment analysis works which mainly focus on

Proceedings of the 29th International Conference on Computational Linguistics, pages 6934–6943
October 12–17, 2022.

single-source domain adaptation setting, it is still cumbersome to apply these domain alignment methods on the multi-source sentiment adaptation tasks, which is more practical in real-world scenarios. Simply combining all the source domains into a single dataset may deliver worse performance compared with the best result from one of source dataset (Guo et al., 2018), due to the various source distributions. Meanwhile, with the number of source domains increasing, the corresponding computation cost and complexity will be dramatically increased (Dai et al., 2020; Xue et al., 2020). In particular, the recent dominant frameworks PrLMs (Devlin et al., 2019; Yang et al., 2019b) are adopted as the backbone for feature extraction, which usually contains a large amount of training parameters (Ye et al., 2020).

Therefore, there is a strong motivation to develop an efficient multi-source unsupervised sentiment adaptation framework which can generalize well to the target domain with no labeled target data. Recent self-training methods (He et al., 2018; Zou et al., 2019; Liu et al., 2021) achieve advanced performance on many unsupervised domain adaptation tasks by iteratively updating the pseudo-labels of the target data with current adapted model, and the model can be retrained with these self-annotated data. However, the pseudo-labels are not always reliable due to the distribution shift between the source domain and the target domain, and the incorrect pseudo-labels can significantly hurt the final adaptation performance. Several techniques are proposed to reduce the negative effect of noisy pseudo labels, such as high-confidence threshold (Zou et al., 2019), self-ensemble bootstrapping (He et al., 2018), mutual information maximization (Ye et al., 2020), *etc.*, which demonstrate improved performance on single-source cross-domain adaptation tasks. In this paper, we propose a novel Asymmetric Mutual Learning (AML) strategy to estimate the pseudo-labels robustly, and we show this strategy is well-suited to the unsupervised multi-source domain adaptation setting. Specifically, we design a classification model for each source domain. For each source model, the pseudo-labels of target data are derived from the ensembles of all the other source models. In contrast with traditional deep mutual learning (Zhang et al., 2018) which distills the knowledge of a single dataset with multiple models, our AML can utilize the knowledge from multiple datasets under different distributions. Therefore, each source model can be enhanced with the other source models. Unlike traditional self-training methods which generate pseudo-label by itself, our AML is more robust to the noisy pseudo-label.

In addition, we tend to use the features extracted from BERT for efficient sentiment adaptation, and this feature-based adaptation method is more memory-friendly compared with fine-tuning BERT. To fully exploit BERT features, we propose a dynamic network (Yang et al., 2019a) for better aggregating the features from different layers, which is referred to Dynamic Feature Networks (DFN). Compared with attention-based fusion (Vaswani et al., 2017) which only scales the features, our DFN can dynamically adjust the parameters of the network according to each instance input for better performance. Together with the two proposed modules (AML and DFN), we achieve new state-of-the-art performance on the widely-used sentiment benchmarks (Blitzer et al., 2007) under unsupervised multi-source setting. We summarize our contributions as follows:

- We propose a novel asymmetric mutual learning (AML) method, which is designed for multi-source unsupervised sentiment adaptation task and beneficial for real-world sentiment analysis applications.
- To achieve efficient adaptation on sentiment classification, we propose a dynamic feature network (DFN), which allows to dynamically assemble multiple parameters for the extracted features, and not update the encoder of PrLMs during adaptation training.
- We demonstrate that the proposed model achieves SOTA performance on multiple sentiment adaptation benchmarks, and the ablation studies verify the effectiveness of each proposed module.

The remainder of the paper is organized as follows. Section 2 introduce the related work, followed by the proposed framework in Section 3. Experimental results are reported in Section 4. Conclusion is drawn in the last Section 5.

2 Related work

In this section, we mainly focus on recent related methods based on Deep Neural Networks (DNNs) due to their superior performance.

Sentiment Classification: aims to predict the sentiment polarity of a given texts. Dang et al. (2020) compare many DNN-based methods, such as Convolutional Neural Networks (CNNs) (Kim, 2014), Recurrent Neural Networks (RNNs) (Zhou et al., 2016), *etc*. However, these methods often use word embedding or TF-IDF as the representations of the texts, which can not capture the context information within a sentence. Recently, with the advent of pre-trained language models which achieve impressive performance in many NLP tasks (Devlin et al., 2019), more and more works adopt these PrLMs as the backbone for sentiment analysis (Sun et al., 2019a; Dang et al., 2020). Despite their great success, the performance of these models is still suffering from domain-shift of the datasets (Li et al., 2021).

Unsupervised Domain Adaptation: is an attractive topic for dealing with the domain-shift problem. The mainstream is to reduce the distribution discrepancy between the source and the target domains (Ganin et al., 2016; Guo et al., 2018). For sentiment classification tasks, some previous works aim to identify domain-invariant pivot words (Ziser and Reichart, 2018; Li et al., 2018). However, pivot words identification is tedious and may be inaccurate. Ganin et al. (2016) and Li et al. (2017) tend to minimize the whole sentence representation by a binary domain classifier. As for the more challenging multi-source adaptation setting, mixture-of-experts (Guo et al., 2018) aligns the each domain-pair based on MMD for simplicity, and ensemble all the source prediction based on the distance metric. While, Zhao et al. (2018) uses a multi-class domain classifier to align multi-domain distributions and Dai et al. (2020) incorporates pseudo-labels to further improve the performance. Fu and Liu (2022) share a similar idea but using BERT as the backbone. In contrast to most domain-alignment methods which become complex with the number of source domains increasing, we turn to self-training methods which demonstrate effective performance for UDA (Zou et al., 2019; Liu et al., 2021; Dai et al., 2020), and the asymmetric mutual learning (AML) is proposed for robust pseudo-label generation in multi-source adaptation setting.

Dynamic Networks: aim to adjust the networks' architectures or parameters conditioned on each input (Yang et al., 2019a). SkipNet (Wang et al., 2018) can decide whether a block is kept and not, which can significantly reduce the inference time. CondConv (Yang et al., 2019a) can select the best combination of the convolution parameters dynamically, which increase model capacity with marginal computation cost. DyCNN (Chen et al., 2020) shares a similar idea, while uses softmax function to derive the attention coefficiency for each convolution kernel. In this paper, we are interested in adapting the features extracted from BERT for simplicity and efficiency. Thus, we propose a dynamic feature network (DFN) to fully exploit the features and adjust the network parameters accordingly for better performance.

3 Method

In this section, we first introduce the overall framework for multi-source unsupervised sentiment adaptation. Next, we provide further details of each proposed module. The detailed training procedures are presented in the last section.

3.1 Overall Framework

For multi-source unsupervised domain adaptation setting, there are k labeled source domains $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^{k}$ (where $\mathcal{S}_i = \{x_t^{\mathcal{S}_i}, y_t^{\mathcal{S}_i}\}_{t=1}^{|\mathcal{S}_i|}$) and an unlabeled target domain $\mathcal{T} = \{x_t^{\mathcal{T}}\}_{t=1}^{|\mathcal{T}|}$, $|\cdot|$ indicates number of samples in the domain. All these domains have different data distributions: $\mathcal{P}_{\mathcal{S}_i} \neq \mathcal{P}_{\mathcal{T}}$ and $\mathcal{P}_{\mathcal{S}_i} \neq \mathcal{P}_{\mathcal{S}_j}$. Our goal is to train a sentiment classification model with $\mathcal{S}$ and $\mathcal{T}$, which generalizes well to the target dataset.

Many previous works adopt statistic metrics or adversarial training to align the distributions between each domain-pairs, this strategy becomes unstable and complicated with the number of source domains increasing. As shown in Figure 2, we use BERT as the feature extractor for the text input, and build a classifier head for each source domain. Without explicit alignment, we design an asymmetric mutual learning method to estimate the pseudo labels of the target data directly, so that all the source classifier can be adapted to the target domain and mutually enhanced, simultaneously.

Since BERT is a large-scale pre-trained language model, we tend to only use its features for efficient adaptation and inference at test time. To this end, we propose the dynamic feature network to fully exploit the extracted features from different layers by adjusting the network parameters (See section 3.3 for details).

During test stage, we simply average the outputs of all the classifiers as the final prediction.

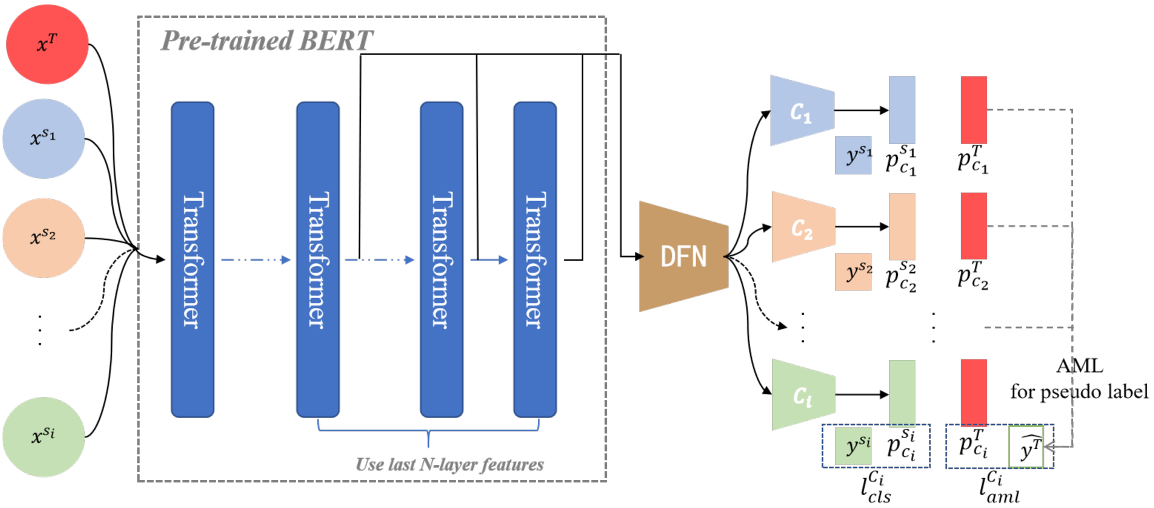

Figure 2: Overview of the proposed framework. We use the pre-trained BERT for feature extraction. DFN indicates dynamic feature network for exploiting better features with dynamic parameters. AML indicates asymmetric mutual learning for robust pseudo labels.

3.2 BERT Feature Extraction

We first extract text features with BERT (Devlin et al., 2019), which consists of several transformer layers. For each layer, we will use the representation of the first `CLS` token as features. Since the transformer layer is based on the self-attention module (Vaswani et al., 2017), the `CLS` token representation should contain all the information from a input sentence.

Given a text input x, the extracted feature from the *last* l^{th} layer can be expressed as follows:

$$\mathrm{f}_l = \mathrm{Transformer}_l^{\mathtt{CLS}}(x), \tag{1}$$

As shown in Figure 2, multiple `CLS` features from last N transformer layers in BERT can be obtained and fed to the proposed dynamic feature network (DFN) for further exploitation.

3.3 Dynamic Feature Network

The text features from BERT are powerful due to the contextualize information (Peters et al., 2019). Sun et al. (2019b) and Merchant et al. (2020) also demonstrate that features from different layers present different behaviors. Since we aim for efficient adaptation without updating the large amount of parameters within BERT extractor, fully exploiting the representations from different layers is necessary for better adaptation performance. Therefore, we propose a novel dynamic feature network (DFN) to adjust the network's parameters based on the input for dynamic adaptation of features.

As shown in Figure 3, the parameters of a fully-connected layer in DFN are a function of the input features. DFN can be adaptive to each input and choose the optimal parameters automatically. There are two advantages: our model can not only fully exploit the different layer features by discovering the optimal aggregation manner, but also increase the model's representation power with marginal computation cost.

We define the dynamic parameters of DFN as $\mathbf{W_f}$, which is conditioned on the input features $\mathbf{f}$. Following Merchant et al. (2020), we select the last N layers of BERT for input features to DFN, which means $\mathbf{f} = \{\mathrm{f}_1, \mathrm{f}_2, ..., \mathrm{f}_N\}$. Similar to Chen et al. (2020), the dynamic parameters of the DFN can be derived from combination of N linear parameters $\{\mathbf{W}_l\}_{l=1}^N$, which is defined as follows:

$$\mathbf{W_f} = \sum_{l=1}^{N} \pi_l(\mathbf{f})\mathbf{W}_l$$
$$\text{s.t.} \quad \pi_l(\mathbf{f}) \in [0,1], \quad \sum_{l=1}^{N} \pi_l(\mathbf{f}) = 1 \tag{2}$$

where the linear parameters $\{\mathbf{W}_l\}_{l=1}^N$ are trainable parameters, and $\pi_l(\mathbf{f})$ denotes the weight for the parameters $\mathbf{W}_l$, which is conditioned on the input feature $\mathbf{f}$. As shown in Figure 3, we use a two-fully connected layers followed by a softmax activation to compute the dynamic weight $\pi_l(\mathbf{f})$ for each $\mathbf{W}_l$. Therefore, the parameters within the DFN can be dynamically adjusted based on each input $\mathbf{f}$ to fully

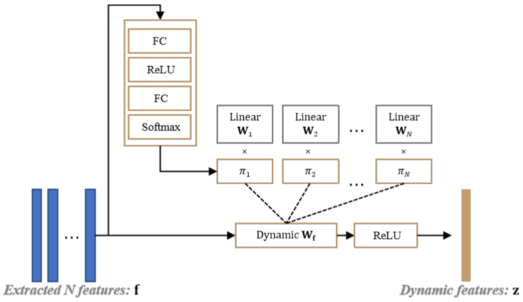

Figure 3: The architecture of the dynamic feature network. 'FC' denotes the fully-connected layer.

exploit the optimal aggregation manner for better feature adaptation performance.

The final dynamic feature representation produced by the DFN is defined as:

$$\mathbf{z} = g(\mathbf{W_f} \cdot \mathbf{f} + \mathbf{b}) \tag{3}$$

where $g(\cdot)$ is ReLU activation function used in this work.

3.4 Asymmetric Mutual Learning

We argue that explicitly reducing the distribution gap between each source-target pair is cumbersome for multi-source adaptation setting, especially for the domain adversarial training scheme (Ganin et al., 2016), which is often unstable and increase the training difficulties (Guo et al., 2018). A promising alternative is to estimate the pseudo-labels for the target domain to guide the adaptation to the target domain, iteratively. Nevertheless, the pseudo-label generated by the model itself, *i.e.*, self-training, will inevitably contain noise (Liu et al., 2021), which can hurt performance seriously.

As shown in Figure 2, we build a classification head C_i for each source domain $\mathcal{S}_i$. Therefore, we can generate the pseudo-label for one classifier by ensembling the output from all the other source domains $\{C_j\}_{j \neq i}$, which is called Asymmetric Mutual Learning (AML). We define the output of C_i as $p_{\theta_{C_i}}(\mathbf{z})$, where $\mathbf{z}$ is the output from DFN, and θ_{C_i} includes the trainable parameters within the DFN. The pseudo-label of a target representation $\mathbf{z}^{\mathcal{T}}$ for C_i can be derived as $\hat{y^{\mathcal{T}}} = \frac{\sum_{j \neq i} p_{\theta_{C_j}}(\mathbf{z}^{\mathcal{T}})}{k-1}$, where k is the number of source domain. The objective function for C_i can be formulated as follows:

$$\min_{\theta_{C_i}} \ell_{cls}^{C_i} + \ell_{aml}^{C_i}, \tag{4}$$

where ℓ_{cls} and ℓ_{aml} indicate supervised classification loss and asymmetric mutual loss, respectively. Both are defined as follows:

$$\ell_{cls}^{C_i} = \mathbb{E}_{\mathbf{z}^{\mathcal{S}_i}, y^{\mathcal{S}_i}}[-y^{\mathcal{S}_i} \log p_{\theta_{C_i}}(\mathbf{z}^{\mathcal{S}_i})] \tag{5}$$

$$\ell_{aml}^{C_i} = \mathbb{E}_{\mathbf{z}^{\mathcal{T}}} || p_{\theta_{C_i}}(\mathbf{z}^{\mathcal{T}}) - \hat{y^{\mathcal{T}}} ||_2 \tag{6}$$

where the superscript denotes the domain name.

Different from the traditional mutual learning (Zhang et al., 2018) with a single dataset, our proposed AML framework is targeted for multi-source unsupervised domain adaptation, and has multiple branches which corresponds to multiple source domains. It not only maintain the domain-specific information by training separate classifiers for corresponding source domains, but also exploit the complementary knowledge from all the other source domains to estimate the pseudo-label of target data for adaptation. It is noting that the Eq. 4 contains two parts, the first supervised source training enables diversified source classifiers, which in turn provide more robust ensembled pseudo-labels of target data for the AML training.

Therefore, during adaptation process, all the source models can be collaboratively enhanced with each other, and exploit robust target knowledge from diverse source models for better multi-source adaptation results.

3.5 Training Procedures

We proceed with the training by alternately optimizing C_i for each source classifier based on the loss objective shown in Eq. 4. The detailed optimization procedure is summarized in Algorithm 1.

During test, we average the output of all the classifiers $\{C_i\}_{i=1}^{k}$ for the final prediction.

4 Experiments

In this section, we extensively evaluate our model on two widely-used sentiment adaptation benchmarks, Amazon view datasets [1] and Skytrax view datasets [2]. First, we introduce the datasets, experiment setup, and implementation details. Then, the performance of recent state-of-the-art adaptation methods are reported for comparisons. Besides, we also conduct detailed ablation studies to verify the contribution of each proposed module.

4.1 Experimental Settings

Amazon view dataset: contains reviews from four products, namely, books (B), DVD (D), electron-

[1] https://www.cs.jhu.edu/ mdredze/datasets/sentiment/

[2] https://github.com/quankiquanki/skytrax-reviews-dataset

Method	D, E, K → B	B, E, K → D	B, D, K → E	B, D, E → K	Avg.
Previous methods					
DANN (Ganin et al., 2016)	0.779	0.789	0.849	0.864	0.820
MDAN (Zhao et al., 2018)	0.786	0.807	0.853	0.863	0.827
MoE (Guo et al., 2018)	0.794	0.834	0.866	0.880	0.843
2ST-UDA (Dai et al., 2020)	0.799	0.839	0.851	0.877	0.841
CTDA (Fu and Liu, 2022)	0.800	0.839	0.866	0.880	0.846
Our methods					
Single-best	0.837	0.831	0.857	0.872	0.849
Source-combined	0.832	0.843	0.863	0.885	0.856
Our model	**0.852**	**0.856**	**0.880**	**0.892**	**0.870**

Table 1: Comparison of multi-source unsupervised domain adaptation results on Amazon review datasets. The best results are denoted with **bold**.

Algorithm 1 Pseudo-code of AML

```
Input: Extracted last $N$ BERT features for all the domains, mini-batch size $B$, learning rates $\zeta_{C_i}$ for each classifier $C_i$, $i \in [0, k]$;
Output: $\theta_{C_i}$, $i \in [0, k]$;
for $epoch$ = 1 to $N$ do
    for $i$ = 1 to $k$ do
        for each mini-batch in the $\mathcal{S}_i$ domain do
            Randomly sample target features;
            Compute the pseudo-label $\hat{y}$ with the other $\{C_j\}_{j \neq i}$
            Update $C_i$ via: $\theta_{C_i} \leftarrow \texttt{Adam}(\nabla_{\theta_{C_i}}(\ell_{cls}^{C_i} + \ell_{aml}^{C_i}), \theta_{C_i}, \zeta_{C_i})$;
        end for
    end for
end for
```

ics (E), kitchen (K). Each produce represents one domains, and has 1,000 positive reviews (label 1) and 1,000 negative reviews (label 2), while has different number of unlabeled reviews. Following similar multi-source unsupervised domain adaptation adopted in (Fu and Liu, 2022), we select one of domain as target domain, the rest domains are used as multi-source domains.

Skytrax view dataset: includes two air-travel-related reviews from *skytrax* website, *i.e.,* Airline (AL) and Airport (AP), which contain 41,396 and 17,721 reviews, respectively. The data distribution discrepancy between the Amazon product views and air-travel reviews should be large, we use all four product datasets as source domains and one of Skytrax view dataset as the target domain, to demonstrate the effectiveness of our proposed method in this challenging settings. To align with Amazon view datasets, we randomly sample 1,000 positive and 1,000 negative reviews from AL and AP domains for training.

Implementation details: In all experiments, we use the pre-trained BERT$_{base}$-uncased (Devlin et al., 2019) to extract features from the last 4 transformer layers, which demonstrates both effectiveness and efficiency, reported in (Peters et al., 2019; Merchant et al., 2020). All the classifiers have the same architecture, which includes a DFN module, followed by two fully-connected layers. We use Adam (Kingma and Ba, 2015) optimizer and set the learning rate to 10^{-4}, weight decay to 10^{-4}, batch size to 16.

4.2 Experimental Results

Results on Amazon review benchmarks: Table 1 compares the sentiment classification accuracies of our method and recent multi-source unsupervised adaptation methods on Amazon review benchmarks. It is noted that our method achieves the best performance on all the multi-source adaptation tasks, which demonstrates the superiority of our proposed method.

Most previous unsupervised multi-source sentiment adaptation methods adopt word embedding as features, which is lack of contextualized information for each word. We show that the BERT features adopted by our method can achieve relatively better results. '*Source-combined*' indicates that we train the DFN-based source classifier using the combination of all the labeled data in the multi-source domains, which is demonstrated to be a strong baseline for multi-source unsupervised domain adaptation tasks (Guo et al., 2018). As shown in Table 1, '*Source-combined*' achieves compara-

Method	B, D, E, K → AL	B, D, E, K → AP	Avg.
Single-best	0.841	0.687	0.764
Source-combined	0.832	0.680	0.756
Our model	**0.850**	**0.695**	**0.772**

Table 2: Adaptation performance from multiple product review domains (Amazon) to one of air-travel review domains (Skytrax). The best results are denoted with **bold**.

Method	D, E, K → B	B, E, K → D	B, D, K → E	B, D, E → K
Without DFN				
Last features	0.831	0.845	0.873	0.873
Avg features	0.836	0.839	0.868	0.878
Our model	**0.852**	**0.856**	**0.880**	**0.892**

Table 3: Ablation study the effect of the proposed DFN module. The best results are denoted with **bold**.

ble or better results on all the adaptation setting with an average accuracy of 85.6%, which outperforms previous methods by around 1 to 3 percentage points. In addition, our AML-based training strategy achieves accuracy of 87.0% on average, which surpasses the strong baseline (*Source-combined*') method and the most recent method (Fu and Liu, 2022) (84.6%) by 1.4 and 2.4 percentage points, respectively.

Adaptation Results from Amazon to Skytrax: Table 2 reports the performance of adaptation to AL and AP domains by using all the Amazon review datasets as multi-source domains. '*Source-best*' indicates the best performance achieved under the single-source domain adaptation setting, the corresponding source domain often has more similar distribution to the target domain. Due to the large domain gap between the product reviews in Amazon and air-travel-related reviews in Skytrax, negative transfer is often occurred. As shown in Table 2, the '*Source-combined*' performance is worse than that of '*Single-best*' baseline in both tasks, which indicates that the extra source data are not fully leveraged, and the distribution shift among multi-source domains brings about the negative effects. While, our model based on the AML strategy can consistently improve the final adaptation performance, which achieves 85.0% and 69.5% accuracy on the AL domain and AP domain, respectively. The corresponding average accuracy is better than the baselines by 1 to 2 percentage points.

4.3 Ablation Study

Effectiveness of AML: We make comparison among '*Source-combined*', '*Single-best*' and our AML-based method to illustrate the effectiveness of the AML module. We show that simply combining all the source data together may hurt the final result due to various source domain distributions. As shown in Table 1, for the $\mathrm{D, E, K} \rightarrow \mathrm{B}$ task, '*Source-best*' surpasses '*Source-combined*' by 0.5 percentage point. As shown in Table 2, both two tasks demonstrate the same results. While, our AML-based adaptation can achieve consistently improved performance without negative transfer in all multi-source sentiment adaptation tasks. We consider that AML makes best of each domain's specific knowledge and enables collaboration among multiple source classifiers to address the negative transfer, so that delivers better performance than both '*Source-best*' and '*Source-combined*' baselines. We also tried the traditional self-training methods on the adaptation from Amazon domains to the Skytrax domain, and found that the accuracy quickly drops caused by the noisy label generated by the model itself. We speculate that during our proposed AML adaptation, each source model learns from the other models, which will not accumulate the same errors as done in self-training. Therefore, AML is more robust to the noisy-labels, and more effective and suitable in the multi-source unsupervised domain adaptation tasks.

Effectiveness of DFN: We conduct several experiments to verify the effect of DFN. As shown in Table 3, we first use the last transformer layer features of BERT as input, which indicates '*last features*'. We also use the same features (last 4 transformer layers) as input, but just average them without DFN module, which indicates '*Avg features*'. It can be observed that simply averaging the features from different layers of BERT delivers comparable performance with only using the last features (less than 1 percentage point in most cases). However, our DFN module can dynamic adjust the network's parameters for better exploiting the input features from different layers. Therefore, the corresponding performance is consistently better than the above two baselines as reported in Table 3, which demonstrate the effectiveness of DFN.

5 Conclusion

In this paper, we propose a novel framework for multi-source unsupervised domain adaptation on sentiment classification. To achieve efficient adaptation with the recent large-scale and powerful pretrained BERT model, we propose a dynamic feature network to find the optimal network parameters for better features exploitation. Besides, instead of explicitly reducing the distribution discrepancy between domain pairs which becomes complex with the number of source domain increasing, we design a asymmetric mutual learning strategy to estimate the pseudo-label of the target data directly. We conduct extensive experiments and ablation studies that verify the effectiveness and superiority of our proposed model.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (Project No.62106136), in part by the Natural Science Foundation of Guangdong Province (Project No.2022A1515010434), in part by Shantou University under Project NTF20007 and Project NTF22012.

References

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

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *ACM International Conference on Information and Knowledge Management*, pages 105–114. ACM.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. 2020. Dynamic convolution: Attention over convolution kernels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11027–11036. Computer Vision Foundation / IEEE.

Yong Dai, Jian Liu, Xiancong Ren, and Zenglin Xu. 2020. Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. In *AAAI Conference on Artificial Intelligence*, pages 7618–7625. AAAI Press.

Nhan Cach Dang, María N. Moreno García, and Fernando de la Prieta. 2020. Sentiment analysis based on deep learning: A comparative study. *CoRR*, abs/2006.03541.

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*, pages 4171–4186. Association for Computational Linguistics.

Yanping Fu and Yun Liu. 2022. Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification. *Knowledge-Based Systems*, 245:108649.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI Conference on Artificial Intelligence*, pages 7830–7838.

Jiang Guo, Darsh J. Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Adaptive semi-supervised learning for cross-domain sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3467–3476. Association for Computational Linguistics.

Jahanzeb Jabbar, Iqra Urooj, JunSheng Wu, and Naqash Azeem. 2019. Real-time sentiment analysis on e-commerce application. In *16th IEEE International Conference on Networking, Sensing and Control*, pages 391–396. IEEE.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1746–1751. ACL.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations, ICLR*.

Tian Li, Xiang Chen, Shanghang Zhang, Zhen Dong, and Kurt Keutzer. 2021. Cross-domain sentiment classification with contrastive learning and mutual information maximization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 8203–8207. IEEE.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI Conference on Artificial Intelligence*, pages 5852–5859. AAAI Press.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *International Joint Conference on Artificial Intelligence*, pages 2237–2243. ijcai.org.

Hong Liu, Jianmin Wang, and Mingsheng Long. 2021. Cycle self-training for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 22968–22981.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online*, pages 33–44. Association for Computational Linguistics.

Alvaro Ortigosa, José M. Martín, and Rosa M. Carro. 2014. Sentiment analysis in facebook and its application to e-learning. *Comput. Hum. Behav.*, 31:527–541.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019*, pages 7–14. Association for Computational Linguistics.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *AAAI Conference on Artificial Intelligence*, pages 2058–2065.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–385. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019b. How to fine-tune BERT for text classification? In *Chinese Computational Linguistics - 18th China National Conference, CCL 2019*, volume 11856 of *Lecture Notes in Computer Science*, pages 194–206. Springer.

Yosephine Susanto, Erik Cambria, Ng Bee Chin, and Amir Hussain. 2022. Ten years of sentic computing. *Cogn. Comput.*, 14(1):5–23.

Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. 2018. Skipnet: Learning dynamic routing in convolutional networks. In *Computer Vision - ECCV*, volume 11217 of *Lecture Notes in Computer Science*, pages 420–436. Springer.

Qianming Xue, Wei Zhang, and Hongyuan Zha. 2020. Improving domain-adapted sentiment classification by deep adversarial mutual learning. In *AAAI Conference on Artificial Intelligence*, pages 9362–9369. AAAI Press.

Brandon Yang, Gabriel Bender, Quoc V. Le, and Jiquan Ngiam. 2019a. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, pages 1305–1316.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

Hai Ye, Qingyu Tan, Ruidan He, Juntao Li, Hwee Tou Ng, and Lidong Bing. 2020. Feature adaptation of pre-trained language models across languages and domains with robust self-training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 7386–7399. Association for Computational Linguistics.

Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. Deep mutual learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, João Paulo Costeira, and Geoffrey J. Gordon. 2018. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 8568–8579.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *COLING 26th International Conference on Computational Linguistics*, pages 3485–3495.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1241–1251. Association for Computational Linguistics.

Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *IEEE/CVF International Conference on Computer Vision*, pages 5981–5990.