SPACL: Shared-Private Architecture based on Contrastive Learning for Multi-domain Text Classification

Guoding Xiong[1], **Yongmei Zhou**[1,2*], **Deheng Wang**[1], **Zhouhao Ouyang**[3]

[1]School of Cyber Security, Guangdong University of Foreign Studies, Guangzhou 510006, China
[2]Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou 510006, China
[3] School of Computing, University of Leeds, Wood-house Lane, Leeds, West Yorkshire, LS2 9JT, United Kingdom
1184467802@qq.com,yongmeizhou@gdufs.edu.cn
1148684516@qq.com,tal-darim@foxmail.com

Abstract

With the development of deep learning in recent years, text classification research has achieved remarkable results. However, text classification task often requires a large amount of annotated data, and data in different fields often force the model to learn different knowledge. It is often difficult for models to distinguish data labeled in different domains. Sometimes data from different domains can even damage the classification ability of the model and reduce the overall performance of the model. To address these issues, we propose a shared-private architecture based on contrastive learning for multi-domain text classification which can improve both the accuracy and robustness of classifiers. Extensive experiments are conducted on two public datasets. The results of experiments show that the our approach achieves the state-of-the-art performance in multi-domain text classification.

1 Introduction

Text classification is one of the most basic tasks among the many tasks of Natural Langugae Processing(NLP). In recent years, the research work of text classification has produced a large number of applications and achieved remarkable results. With the continuous release of a large number of pretrained language models in recent years, such as BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019) and other pretrained models, text classification problems have been able to achieve good results on the basis of neural network and pretrained models. However, most text classification problems are highly domain-dependent in that the meaning of the same word may transform in different domains. For example, the word apple expresses the fruit in kitchen review (e.g., I have shifted to an apple for lunch), while in electronics review, it means a brand of electronic products(e.g., I can't understand how apple sell so much ipod video). A common strategy, training multiple classifiers for different domains, is used to solve above problems. However, text data in reality often have characteristics of multiple domains and the cost of labeling a large number of multi-domain data is too high. Therefore, it is very important and practical meaningful to improve the accuracy of text classification in multiple related domains. Multi-domain text classification(MDTC) (Li and Zong, 2008) is proposed to solve above problems, it aims to utilize textual information in different domains to improve the performance of model architecture, but there is no need to train a separate classifier for each domain. In recent years, deep learning has been widely used in MDTC problems, and has achieved excellent results (Wu and Huang, 2015; Wu and Guo, 2020). The method used in most studies is shared-private architecture. Private modules are used to capture domain-specific knowledge for each domain, and shared modules are used to capture domain-invariant knowledge (Liu et al., 2016).However, these researches only pay attention to how to obtain the shared knowledge of multiple domains and domain-specific knowledge better, but ignore the representation of the samples in the representation space. In order to solve the problems above, in this paper, we propose Shared-Private Architecture based on Contrastive Learning(SPACL), which uses contrastive learning to improve the representations of different types of samples in the representation space, thereby improving the performance level of downstream tasks. Different from previous studies,

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

958

our architecture can not only use conditional adversarial training to extract domain-invariant features, but also generate better sample representations for MDTC.

The contributions of this paper are summarized as follows: 1)In order to strengthen the alignment representations of data in different domains, we propose a shared-private architecture based on contrastive learning for multi-domain text classification which can improve both the accuracy and robustness of the text classfier. 2)We adopt a conditional adversarial network to interact domain-shared features and classification labels, which can be better adapted to multi-domain text classification. 3)Experiments are carried out on two public multi-domain datasets, and the experimental results compared with multiple baselines show that our proposed model architeture has achieved state-of-the-art results.

2 Related Work

2.1 Multi-domain text classification

Multi-domain text classification was proposed first to improve performance through fusing training data from multiple domains (Li and Zong, 2008). The biggest challenge of this task is that the same text may has different implications in different domains, and the cost of labeling each domain is too costly.

Some early studies mainly used domain transfer learning techniques for MDTC. The structural correspondence learning (SCL) algorithm was proposed to select source domains most likely to adapt well to given target domains (Blitzer et al., 2007). Pan et al. (2010) proposed a spectral feature alignment (SFA) method to align domain-specific words from different domains into unified clusters, with the help of domain-independent words as a bridge. Wu and Huang (2015) proposed a novel approach based on multi-task learning to train sentiment classifiers for different domains in a collaborative way. Liu et al. (2015) proposed a multi-task deep neural network(MTDNN) for learning representations across multiple tasks, not only leveraging large amounts of cross-task data, but also benefiting from a regularization effect that leads to more general representations to help tasks in new domains. Liu et al. (2017) proposed an adversarial multi-task learning framework, alleviating the shared and private latent feature spaces from interfering with each other.

The most recent prior works on MDTC include Meta Fine-Tuning(MFT) for multi-domain text classification (Wang et al., 2020). Dual Adversarial Co-Learning(DACL) for Multi-Domain Text Classification (Wu and Guo, 2020), Conditional Adversarial Networks(CAN) for Multi-Domain Text Classification (Wu et al., 2021a) and Mixup Regularized Adversarial Networks(MRAN) for Multi-Domain Text Classification (Wu et al., 2021b). MFT uses meta-learning and domain transfer technology to learn highly transferable knowledge from typical samples in various domains. Both DACL and CAN leverage adversarial training to obtain the shared domain features. MRAN adopts the domain and category mixup regularizations to enrich the intrinsic features in the shared latent space and enforce consistent predictions in-between training instances. However, these methods ignore the distance of samples in the feature space when learning multi-domain feature representations, which is an important guideline to help classification. Furthermore, they did not consider the interaction between the extracted features and class labels, which is often important to improve their correlation. Different from the above studies, the work our proposed further advances the line of study by deploying contrastive learning. It can also model the interactions between shared domain features and classes to enhance their representations through a conditional adversarial network. We assume that data in various domains is insufficient, and make full use of data from multiple domains to improve overall system performance.

2.2 Contrastive Learning

Recently, related researches show that contrastive learning is an effective self-supervised learning method. Chen et al. (2020) proposed simple framework for contrastive learning of visual representations(SimCLR) to improve the quality of the learned representations by contrastive learning. Meng et al. (2021) present a self-supervised learning framework, COCO-LM, that pretrains Language Models by COrrecting and COntrasting corrupted text sequences. Giorgi et al. (2020) present Deep Contrastive Learning for Unsupervised Textual Representations(DeCLUTR) to enclose the performance gap between unsupervised and supervised pretraining for universal sentence encoders. One of the key aspects of con-
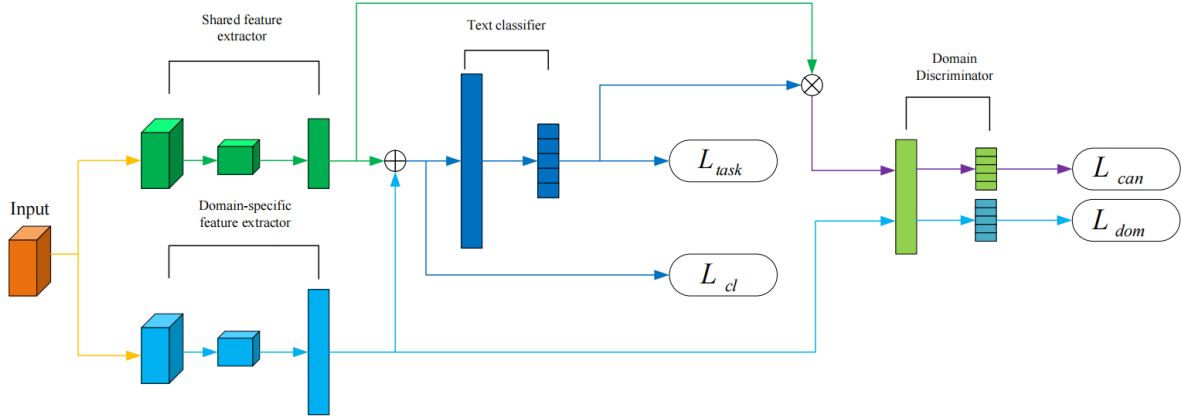
Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

959

Figure 1: The overall model atchitecture of SPACL. A shared feature extractor is used to capture the shared knowledge. Each domain-specific extractor is trained to extract the domain-specific knowledge. A domain classifier is trained to predict the domain label of the input sample. A text classifier is trained to predict the class of samples and calculate the loss of contrastive learning. L_{task} is the loss function of text classification. L_{cl} is the loss fuction of contrastive learning. L_{dom} is the loss function of domain classfication. L_{can} is the conditoning adverserial loss function which extracts the shared knowledge across domains.

trastive learning is the sampling of positive pairs. Gao et al. (2021) add dropout noise to keep a good alignment for positive pairs. Fang et al. (2020) uses data augmentation to generate positive pairs from the original sentences.

We develop our model architecture with contrastive learning. In our experiments, we select a sample and combine it with itself to get a positive pair. And then combine it with other different kinds of samples to get negative pairs. A contrastive loss is used to control the distance between samples of different classes in the sample space so that enhance the ability of the text classifier.

3 Methodology

3.1 Model Architecture

In this paper, we consider MDTC tasks in the following settings.

Specifically, there exists M domains $\{D_i\}_{i=1}^{M}$. The labeled training collection of the m-th domain is denoted by $X_1^m = \left\{ \left(x_j^m, y_j^m \right) \mid j \in [1, N_1^m] \right\}$, where x_j^m and y_j^m are the input texts and the label of the j-th sample of the m-th domain. N_l^m is the total number of the labeled samples of the m-th domain. The unlabeled training collection of the m-th domain is denoted by $X_u^m = \{(x_k^m) \mid k \in [1, N_u^m]\}$, where x_k^m and N_u^m are the input texts of the j-th sample and the sample size of the m-th domain. N_L represents the amount of labeled data for all domains and N_U represents the amount of unlabeled data for all domains. The goal of MDTC is to improve the overall system performance by utilizing the training sets of M domains. The classification performance of the system is measured by the average classification accuracy across M domains.

3.2 Domain-specific Representation Learning

In order to ensure the validity of our extracted domain-specific features, we add a simple and effective domain discriminator D_d, which takes the extracted domain-specific features as input and outputs the predicted domain category, so as to optimize the domain discrimination ability. The h_p is the output of the domain-specific extractor for the given instance X. The domain classifier $D_d (h_p; \theta_d) \to \hat{d}$ maps the domain-specific feature representation to a domain label prediction. θ_d denotes the parameters of the domain classifier D_d. The discriminator D_d is trained to minimize the prediction loss on labeled and unlabeled instances of multiple domains:

$$L_{\text{dom}} = -\frac{1}{N_U + N_L} \sum_{m=1}^{M} \sum_{j=1}^{N_l^m + N_u^m} d_j^m \log \hat{d_j^m} + \left(1 - d_j^m\right) \log \left(1 - \hat{d_j^m}\right) \tag{1}$$

where the $\hat{d}$ is prediction probaiblities of domain labels of domain discriminator D_d and the d is the true domain label of input text.

3.3 Conditional Adversarial Network

Motivated by some previous works of domain separation learning (Bousmalis et al., 2016; Shen et al., 2018), we adopt a conditional adversarial network for SPACL to extract domain shared features. After the domain-specific learning, we freeze the parameters θ_d of the domain discriminator D_d to ensure that the discriminator has good domain recognition capabilities. At the same time, in order to ensure that the features we extract can express shareability across domains, we also adopt a negative entropy loss so that the domain classifier cannot accurately identify the domain of the shared-representation the input text.

The h_s is the output of the shared extractor F_s for the given instance X. The h_s is the output of the shared extractor F_s for the given instance X. The h_p is the output of the shared extractor F_p for the given instance X. The final joint representations h is the concatenated vector of private features h_p and shared features h_s. The text classifier C outputs the probability distribution of the prediction labels which are denoted as h_c. The domain classifier $D_d\left(h_c \otimes h; \theta_d\right) \to \hat{d}$ maps the joint feature representation h and the class prediction h_c to a domain label $\hat{d}$. The loss can be defined as:

$$L_{\text{can}} = \frac{1}{N_U + N_L} \sum_{m=1}^{M} \sum_{j=1}^{N_l^m + N_u^m} d_j^m \log \hat{d_j^m} + \left(1 - d_j^m\right) \log \left(1 - \hat{d_j^m}\right) \tag{2}$$

where $h_c \otimes h$ denotes the cross-covariance of the two vectors which is calculated by multilinear conditioning (Long et al., 2018).

3.4 Contrastive Learning

Intuitively, we hope that the distance between the final joint representation vectors of samples of different categories is as far as possible, so as to make the final text classifier C easier to distinguish. Therefore, we adopt a contrastive learning approach to generate better joint representation vectors. Specifically, assuming that given a batch of samples, we will sample a pair of positive examples and other sets of negative examples in the batch. The class label of every sample denotes y. Given a final joint representation h_i of a sample, from a batch we can get an positive pair (h_i, h_{pos}) and other negative sample pairs $\{(h_i, h_{neg}) \mid h_i \in y, h_{neg} \notin y\}$.

The loss of contrastive learning is defined as:

$$L_{\text{cl}} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \log \frac{\exp\left(\text{sim}\left(h_i, h_{\text{pos}}\right)\right)}{\sum_{\{(h_i, h_{neg}) \mid h_i \in y, h_{neg} \notin y\}} \exp\left(\text{sim}\left(h_i, h_n\right)\right)} \tag{3}$$

where $\text{sim}(u, v) = u^T v / \|u\|_2 \|v\|_2$ denotes the cosine distance between the two vectors u and v. N_b and $\|.\|_2$ denote the number of batch size and the L2 norm.

3.5 Objective Function

The multi-domain text classification task is a binary classification task. Therefore, we define the task loss is :

$$L_{\text{task}} = -\frac{1}{N_L} \sum_{m=1}^{M} \sum_{j=1}^{N_l^m} y_j^m \log \hat{y_j^m} + \left(1 - y_j^m\right) \log \left(1 - \hat{y_j^m}\right) \tag{4}$$

The text classifier C takes the final joint representation as input, and outputs the prediction labels which denote $\hat{y}$.

The final loss function is the combination of above losses:

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

961

Computational Linguistics

$$L = L_{\text{task}} + L_{\text{dom}} + \alpha L_{\text{can}} + \beta L_{\text{cl}} \tag{5}$$

where α and β are hyperparameters for balancing different losses.

4 Experiment

4.1 Dataset

We evaluate SPACL on two standard datasets in our experiments: the Amazon review dataset (Blitzer et al., 2007) and the FDU-MTL dataset (Liu et al., 2015). The Amazon review dataset contains reviews in four domains: books, DVDs, electronics, and kitchen. The data for each domain has 1000 positive samples and 1000 negative samples. This dataset is already preprocessed into a bag of features (unigrams and bigrams) which loses word order information. The FDU-MTL datasets contains a total of 16 domains: books, electronics, DVDs, kitchen, apparel, camera, health, music, toys, video, baby, magazine, software, sport, IMDB, and MR. Each domain of FDU-MTL dataset contains a development set of 200 samples, a test set of 400 samples, a training set of about 1400 samples, and about 2000 unlabeled samples.

4.2 Baselines

To evaluate SAPCL, we compare it with the following baselines.

The multi-task learning with bidirectional language (MT-BL) method utilizes extraction of task-invariant features by leveraging potential information among related tasks, which improves the performance of a single task (Yang and Shang, 2019). The multinomial adversarial network (MAN) learns features that are invariant across multiple domains by resorting to its ability to reduce the divergence among the feature distributions of each domain (Chen and Cardie, 2018). This method trains the domain discriminator by two loss functions: the least square loss (MAN-L2) and the negative log-likelihood loss (MAN-NLL). Dual adversarial co-learning (DACL) deploys dual adversarial regularizations to align features across different domains, aiming to improve the classifiers' generalization capacity with the learned features (Wu and Guo, 2020). Conditional adversarial networks (CANs) introduce a conditional domain discriminator to model the domain variance in both shared feature representations and class-aware information simultaneously and adopts entropy conditioning to guarantee the transferability of the shared features (Wu et al., 2021a). The collaborative multi-domain sentiment classification (CMSC) train the models by three loss functions: the least square loss (CMSC-LS), the hinge loss (CMSC-SVM), and the log loss (CMSC-Log) (Wu and Huang, 2015). The adversarial multi-task learning for text classification (ASP-MTL) alleviates the shared and private latent feature spaces from interfering with each other (Liu et al., 2017). All the comparison methods use the standard partitions of the datasets. Thus, we cite the results from (Wu and Huang, 2015; Liu et al., 2017; Chen and Cardie, 2018; Yang and Shang, 2019; Wu and Guo, 2020; Wu et al., 2021a) for fair comparisons.

Domain	CMSC-LS	CMSC-SVM	CMSC-Log	MAN-NLL	MAN-L2	DACL	CAN	SPACL(proposed)
Books	82.10	82.26	81.81	82.98	82.46	83.45	83.76	**84.65**
DVD	82.40	83.48	83.73	84.03	83.98	**85.50**	84.68	85.20
Elec.	86.12	86.76	86.67	87.06	87.22	87.40	**88.34**	88.20
Kit.	87.56	88.20	88.23	88.57	88.53	90.00	90.03	**90.10**
Avg	84.55	85.18	85.11	85.66	85.55	86.59	86.70	**87.03**

Table 1: MDTC results on the Amazon review dataset

4.3 Experimental Setting

In our experiment, we set the hyperparameters α=0.001, β=0.1. The experiment uses the Adam optimizer with the learning rate of 0.0001. The vector size of the shared feature extractor is 64 while the vector size of the domain-specific feature extractor is 128. The dropout rate is 0.5. ReLU is the activation function.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

962

Domain	MT-BL	ASP-MTL	MAN-L2	MAN-NLL	SPACL(proposed))
books	89.0	84.00	87.6	86.8	**90.2**
electronics	**90.2**	86.80	87.4	88.8	90.0
dvd	88.0	85.50	88.1	**88.6**	88.5
Kitchen	**90.5**	86.20	89.8	89.9	90.0
apparel	87.2	87.00	87.6	87.6	**88.0**
camera	89.5	89.20	91.4	90.7	**91.2**
health	**92.5**	88.20	89.8	89.4	90.2
music	**86.0**	82.50	85.9	85.5	**86.0**
toys	**92.0**	88.0	90.0	90.4	91.1
video	88.0	84.5	89.5	**89.6**	88.7
baby	88.7	88.20	90.0	**90.2**	89.9
magazine	92.5	92.20	92.5	**92.9**	92.5
software	**91.7**	87.20	90.4	90.9	89.5
sports	**89.5**	85.7	89.0	89.0	88.2
IMDb	88.0	85.5	86.6	87.0	**88.7**
MR	75.7	**76.7**	76.1	**76.7**	76.5
AVG	88.6	86.1	88.2	88.4	**88.7**

Table 2: MDTC results on the FDU-MTL dataset

Method	Book	DVD	Electronics	Kitchen	AVG
SPACL w/o C	83.10	83.05	85.10	86.20	84.36
SPACL w/o CL	84.10	82.50	84.00	85.05	83.90
SPACL w/o D	83.05	80.01	82.05	83.17	82.07
SPACL(full)	**84.65**	**85.20**	**88.20**	**90.10**	**87.03**

Table 3: Ablation study on the Amazon review dataset

The batch size is 128. MLP feature extractors are the feature extractor of the experiment on the Amazon review dataset with an input size of 5000. MLP feature extractor is composed of two hidden layers, with size 1,000 and 500, respectively. CNN feature extractor with a single convolutional layer is the feature extractor of the experiment on the FDU-MTL review dataset. Each CNN feature extractor uses different kernel sizes (3, 4, 5) with input size of 1000. Text classifier and discriminator are MLPs with one hidden layer of the same size as their input (128 + 64 for text classifier and 128 for discriminator).

4.4 Results

We conduct the experiments on the Amazon review dataset and FDU-MTL dataset following the setting of (Chen and Cardie, 2018). A 5-fold cross-validation is conducted on the Amazon review dataset. All data is divided into five folds: three folds are used as the training set, one fold is used as the validation set, and the remaining one fold is used the test set. The experimental results on the Amazon review dataset are shown in Table 1 and the results on the FDU-MTL dataset are shown in Table 2. The best performance is shown in bold.

From Table 1, we can see that our proposed SPACL architecture is able to achieve the best average accuracy across multiple domains on the Amazon review dataset. This suggests our proposed model architecture is more effective than other baselines. From the experimental results on FDU-MTL in the Table 2, the average accuracy of our proposed SPACL is superior to the other methods. The experimental results once again demonstrate the effectiveness of our proposed method.

The reasons for the above results are as follows: 1)Our model utilizes a conditional adversarial network to correlate the extracted shared features and predicted class labels, thereby improving the overall

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

963

generalization performance of the model architecture. 2)Our model architecture expands the distance between samples of different classes in the sample space and the distance of samples of the same class through the method of comparative learning. Therefore, our model performs better at multi-domain text classification tasks.

4.5 Ablation Study

To validate the contribution of conditional adversarial networks and contrastive learning in our model architecture, we conduct extensive ablation experiments on the Amazon review dataset. In particular, we studied two kinds of ablation variants: (1)SPACL w/o C, the variant model architecture of our SPACL without conditional adversarial learning on shared feature extractor; (2)SPACL w/o CL, the variant model architecture of our SPACL without contrastive learning on the final joint representation; (3)SPACL w/o D, the variant model architecture of our SPACL without domain-specific representation learning; The ablation experiment results are shown in the Table 3, where we can see all variants of produce poor results, the full model architecture provides the best performance. Therefore, this validated our model architecture of the components in the presence of necessity. From the results of the ablation experiments, we can see that using contrastive learning to improve the sample representation benefits the performance of our model.

5 Conclusion

In this paper, we proposed a shared-private architecture based on contrastive learning to use across different domains of all the available resources for multi-domain text classification. The model architecture expands the distance between shared-representations of samples of different categories in the sample space by introducing contrastive learning, thereby further improving the discriminative ability of the model architecture. In addition, the model architecture uses a conditional adversarial network to establish the correlation between domain shared features and classification prediction labels which improves the overall performance of the model architecture. The experimental results on two benchmarks show that the SPACL model architecture can effectively improve the performance of the system on the multi-domain text classification task. In the future, we will explore a better solution to transfer knowledge from different domains for multi-domain text classification.

Acknowledgements

References

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.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. *arXiv preprint arXiv:1802.05694*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

964

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, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Mengting Hu, Yike Wu, Shiwan Zhao, Honglei Guo, Renhong Cheng, and Zhong Su. 2019. Domain-invariant feature distillation for cross-domain sentiment classification. *arXiv preprint arXiv:1908.09122*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *Proceedings of ACL-08: HLT, Short Papers*, pages 257–260.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

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*.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Meta fine-tuning neural language models for multi-domain text mining. *arXiv preprint arXiv:2003.13003*.

Yuan Wu and Yuhong Guo. 2020. Dual adversarial co-learning for multi-domain text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6438–6445.

Fangzhao Wu and Yongfeng Huang. 2015. Collaborative multi-domain sentiment classification. In *2015 IEEE international conference on data mining*, pages 459–468. IEEE.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021a. Conditional adversarial networks for multi-domain text classification. *arXiv preprint arXiv:2102.10176*.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021b. Mixup regularized adversarial networks for multi-domain text classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7733–7737. IEEE.

Qi Yang and Lin Shang. 2019. Multi-task learning with bidirectional language models for text classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 958-965, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

965