Domain Generalization for Text Classification with Memory-Based Supervised Contrastive Learning

Qingyu Tan[*1,2] **Ruidan He**[†1] **Lidong Bing**[1] **Hwee Tou Ng**[2]
[1]DAMO Academy, Alibaba Group
[2]Department of Computer Science, National University of Singapore
{qingyu.tan,ruidan.he,l.bing}@alibaba-inc.com
{qtan6,nght}@comp.nus.edu.sg

Abstract

While there is much research on cross-domain text classification, most existing approaches focus on one-to-one or many-to-one domain adaptation. In this paper, we tackle the more challenging task of domain generalization, in which domain-invariant representations are learned from multiple source domains, without access to any data from the target domains, and classification decisions are then made on test documents in unseen target domains. We propose a novel framework based on supervised contrastive learning with a memory-saving queue. In this way, we explicitly encourage examples of the same class to be closer and examples of different classes to be further apart in the embedding space. We have conducted extensive experiments on two Amazon review sentiment datasets, and one rumour detection dataset. Experimental results show that our domain generalization method consistently outperforms state-of-the-art domain adaptation methods[1].

1 Introduction

Text classification is a highly important and widely studied natural language understanding task. Recent success in self-supervised pre-training has significantly improved the state-of-the-art performance in sentiment analysis. However, sentiment classification is widely known as a domain-dependent task, mainly because sentiment expressions can have different meanings in different domains (e.g., "long" in "long waiting time" of a restaurant review is negative, while in "long battery life" of a laptop review is positive). Moreover, the amount of labeled data is highly imbalanced across different domains. Many NLP domains still lack sufficient labeled data for training a high-performance classifier. Therefore, it is crucial to adapt sentiment knowledge from resource-rich domains to low-resource domains. This strand of research is known as domain adaptation (DA). Prior works on domain adaptation typically follow a one-to-one (Jiang and Zhai, 2007) or many-to-one adaptation setting (Zhao et al., 2018), where the model is usually trained on labeled data from the source domain along with unlabeled data from the target domain and is then evaluated on the target domain data. The usage of the unlabeled target data is crucial in DA. Prior works mainly use it for domain-invariant representation learning or model selection. More recently, Wright and Augenstein (2020) have demonstrated that large pretrained language models (PrLMs) are able to achieve promising performance for cross-domain sentiment classification. Subsequent works further improve the adaptation performance through domain adversarial training (Du et al., 2020; Karouzos et al., 2021), iterative pseudo-labeling (Ye et al., 2020), or prompting (Ben-David et al., 2022), where they all adapt the DA setting and assume unlabeled target data is available during model training.

However, in more realistic scenarios, one may be asked to build a text classification model that will be applied to unknown target domains, which implies unlabeled target domain data is unavailable during training or model selection. Domain generalization (Li et al., 2018;Wang et al., 2021b;Wang et al., 2021a) has been proposed to address this problem by learning a universal representation using labeled data from multiple source domains, without access to target domain data. Building such a generalized model enables us to predict sentiment polarity of emerging domains, such as COVID-19 vaccines and pandemic-related medical equipment. Compared to domain adaptation, the major advantage of domain generalization is that one trained model can be used for all target domains, whereas domain adaptation needs to train

[*] Qingyu Tan is under the Joint PhD Program between Alibaba and National University of Singapore.

[†] Corresponding author.

[1]Our code is available at https://github.com/tonytan48/MSCL.

Proceedings of the 29th International Conference on Computational Linguistics, pages 6916–6926
October 12–17, 2022.

a tailored model for each target domain. To the best of our knowledge, even though there are many works that focused on multi-source domain adaptation (Zhao et al., 2018; Guo et al., 2020; Wright and Augenstein, 2020), there is no prior work that tackles domain generalization in the context of text classification.

In this work, we focus on domain generalization for text classification. We aim to build a generalized sentiment classifier using labeled source data from multiple domains. The trained model can be applied to other unseen domains. We train a classifier to learn a joint hypothesis over all source domain data. On this basis, we propose to use supervised contrastive learning (SCL) to better capture the similarity between examples from different domains but belong to the same sentiment class and contrast them with examples belonging to the other class. SCL explicitly pulls the representations from the same class together and repulses the representations from different classes in the embedding space. This objective helps the model better extract the domain invariant features among all source domains, thereby allowing a better classification decision boundary. The usage of SCL helps the classifier achieve better generalization ability, which is especially beneficial to the domain generalization scenario. Although SCL has previously been applied to image classification tasks (Khosla et al., 2020) and in-domain sentence-level classification tasks (Gunel et al., 2021), to the best of our knowledge, we are the first to apply it in the context of domain generalization. It is shown that the performance of SCL is highly affected by batch size as it requires a large number of contrasting examples for computing the contrastive loss (Khosla et al., 2020;Gunel et al., 2021). An optimal batch size requires large memory, making it impractical in many use cases. To apply SCL without excessive memory consumption, we further propose to use a memory bank to store the representations to increase the size of contrasting features, so that the hidden representations for sentiment classification will be reused for computing supervised contrastive loss. In this way, we can significantly improve performance compared to directly applying SCL for text classification.

To examine our proposed method, we conducted domain generalization experiments on two popular Amazon review datasets (one monolingual and the other multilingual) and on the PHEME rumour detection dataset. Following previous works (Chen and Cardie, 2018; Ye et al., 2020; Li et al., 2020; Liu et al., 2021), different languages can be seen as distinct domains based on a shared cross-lingual encoder. Hence, the domain generalization problem can also be extended to language generalization. We conducted experiments in cross-domain (CD), cross-language (CL), and cross-language cross-domain (CLCD) settings. In our experiments, our proposed method is able to outperform the second best domain adaptation method by 0.81% accuracy score in sentiment analysis and 1.68% F1 score in rumour detection.

Our contributions can be summarized as follows:

- We are the first to tackle the domain generalization problem for text classification.
- We proposed a novel memory-based alternative for supervised contrastive learning to improve its performance. Experimental results show that our proposed method consistently outperforms state-of-the-art domain adaptation baselines.

2 Related Work

2.1 Domain Adaptation

Prior works on domain adaptation mainly focus on minimizing the distributional discrepancy between the source domain and the target domain. Kernelized methods, such as Maximum Mean Discrepancy (MMD) (Arbel et al., 2019), spectral feature alignment (Pan et al., 2010), and domain-adversarial training (Ganin et al., 2016) are commonly used for feature alignment from different domains. Another widely explored approach for DA is self-training, where a classifier is first trained on the source domain and later used for predicting pseudo-labels for unlabeled data in the target domain. Ye et al. (2020) have proposed a robust self-training approach to improve the performance of the joint hypothesis between source and target domains and thereby improve performance on the target domain.

When multiple source domains are present, DA methods should be modified accordingly. Li et al. (2018) have used pairwise-MMD and an adversarial autoencoder to overcome domain discrepancy. Wu and Huang (2016) extended domain-adversarial training from Ganin et al. (2016) to multiple source domains. Domain adversarial training has many popular variants (Zhao et al., 2018;

Liu et al., 2018; Chen et al., 2018), showing that extracting domain-invariant representation is a crucial part for domain adaptation. Another strand of work for multi-source DA is based on mixture of experts (MoE). Chen and Cardie (2018) have explored MoE for multi-source cross-lingual sentiment classification, and MoE encourages the model to learn from more relevant source languages. Guo et al. (2020) proposed a DistanceNet-Bandits approach to tackle multi-source DA. It first measures domain distance with multiple metrics, and then uses a multi-armed bandit mechanism to learn from closer source domains. However, prior works on multi-source domain adaptation rely on distance measurement from source to target domain. Therefore, even though no labeled data from the target domain is used, unlabeled target data must be used to perform domain adaptation. Unlike multi-source domain adaptation, domain generalization has no access to any target domain data during training. In many practical scenarios, we may need to find people's opinions towards emerging and unseen domains, such as pandemic-related medical equipment. Therefore, it is important to study the problem of domain generalization for sentiment and text classification.

2.2 Contrastive Learning

Recently, contrastive learning (CL) has led to major advances in self-supervised representation learning. The common idea in these works is maximizing the agreement score between an anchor and a 'positive' example in the embedding space, and pushing apart the anchor from many 'negative' examples (Chen et al., 2020). The positive example pair is typically a particular image and its augmentation, and the negative pairs are formed by that image with other images within the same batch. However, such approaches require a substantially large batch size. Otherwise the performance of contrastive learning deteriorates significantly. On the other hand, other approaches have been proposed to alleviate resource consumption due to a large batch size. Grill et al. (2020) have shown superior performance by maximizing the agreement of positive pairs by a momentum encoder, without the need of negative samples. Khosla et al. (2020) have extended the idea of contrastive learning to the supervised setting, i.e., supervised contrastive learning (SCL). To leverage label information, SCL considers examples with the same label as positive pairs and examples with different labels as negative pairs, and achieves significant performance gain in image classification tasks. Gunel et al. (2021) extended the application of SCL to finetuning pre-trained language models in natural language understanding tasks. Graf et al. (2021) further analyzed SCL in image classification problem, showing that SCL is able to increase the inter-cluster distance and reduce the intra-cluster distance for each class. We leverage the idea of SCL to explicitly align features of the same class but from different domains. Since the negative sample size is crucial for accurate mutual information estimation, we propose to use an additional memory bank to store representations and progressively reuse the encoded sentence representations, thereby improving the performance of contrastive learning. In this way, we force the domain generalization model to focus on aligning features of the same class and implicitly reduce the domain discrepancy among the multi-source training data.

3 Problem Definition

For the task of domain generalization on text classification, suppose we have labeled data from k source domains $\{\mathcal{D}_{s_i}\}_{i=1}^{k}$. For each source domain, the labeled data D_{s_i} is denoted as $D_{s_i} = \{X_{s_i}, Y_{s_i}\}$. In the training phase, only source domains are available. Then, the labeled dataset from the target domain $D_t = \{X_t, Y_t\}$ is used for evaluation. The problem setup of domain generalization is different from multi-source domain adaptation (MSDA), which requires an additional unlabeled set from the target domain during training (Wu and Huang, 2016; Ding et al., 2019; Zhao et al., 2018; Guo et al., 2020). In contrast, in the domain generalization setup, the model is only trained to obtain a domain-invariant feature from the given source domains, while the target-domain data D_t is only used during evaluation. In this way, the trained model can be used to make predictions on unseen domains.

4 Model Description

Since the domain generalization problem is to train an algorithm based on multiple source domains, the key challenge of this classification problem is to learn an ideal joint hypothesis of the source domains. That is, we aim to separate the data points by their labels as much as possible, thereby minimizing domain discrepancy within each class in

the feature embedding space. Our supervised contrastive learning model not only widens the margin of decision boundaries, but also enforces the distribution within the same class to be more uniform, hence minimizing the source domain distances for each class.

We first shuffle all the source domain data and divide the joint dataset into mini-batches. This is mainly for enforcing a stable domain distribution for each mini-batch. For one sampled mini-batch of size N, we have $S = \{x_i, y_i\}_{i=1}^{N}$, where x_i represents the input text, y_i represents the sentiment label of that example. We first adopt a pre-trained language model (PrLM) as encoder. We use the hidden state of the last layer's [CLS] token as document representation, and denote it as $\mathbf{h}$. We then use a feed-forward neural network (followed by a $\tanh$ activation function and layer norm) f for dimension reduction. Specifically, we have $\mathbf{z} = f(\mathbf{h})$. Feature $\mathbf{z}$ will later be fed into a classifier g for downstream tasks:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(g(\mathbf{z}_i)\big) \tag{1}$$

where g is a fully connected classifier and $\mathcal{L}_{CE}$ is the cross-entropy classification loss.

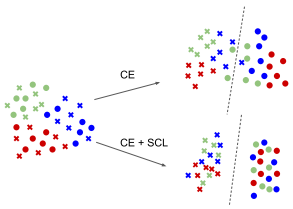


Figure 1: Illustration of SCL. ◯ refers to a positive review, × refers to a negative review, and the different colors of data points represent their domains.

4.1 Supervised Contrastive Loss for Domain Generalization

As illustrated in Figure 1, supervised contrastive learning explicitly pulls the representations of the same class together and repulses representations from different classes, which increases the discriminative capability of hidden representations and benefits hard negative mining (Khosla et al., 2020). This objective suits our goal of learning a joint hypothesis for different source domains, and enables the model to learn a more uniform distribution for each label class. Specifically, for a given mini-batch S of size N, the supervised contrastive loss is computed as follows:

$$\mathcal{L}_{SCL} = -\sum_{\mathbf{z}_i \in S} \frac{1}{N} \sum_{\mathbf{z}_p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{\mathbf{z}_a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)} \tag{2}$$

where $\mathbf{z}$ is the vector representation. For a given anchor representation $\mathbf{z}_i$, $P(i) \equiv \{\mathbf{z}_j \in S, y_j = y_i\}$ refers to the set of positive examples, $A(i) \equiv \{\mathbf{z}_j \in S, j \neq i\}$ refers to the union of positive examples and negative examples, S refers to the set of the mini-batch. τ is a scaling hyper-parameter, also known as temperature. Then, we have our combined loss function as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SCL} \tag{3}$$

However, in our preliminary experiments, directly applying supervised contrastive learning has marginal effect for domain generalization. This is because the performance of supervised contrastive learning is heavily related to batch size, as a larger batch size is better in representing the mixed distribution of multiple domains. However, increasing the batch size inevitably introduces high computation and memory costs. In order to improve the performance of domain generalization while reducing memory consumption, we propose to use an additional memory bank to reuse the previously encoded sentence representations. The details for our memory bank are described in the following subsection.

4.2 Memory-Based Supervised Contrastive Learning

To increase the number of contrasting examples while limiting memory consumption, we propose to use a memory bank Q to store the sentence representations and their labels of each batch. The purpose of introducing this memory bank is to progressively reuse the encoded sentence to compute $\mathcal{L}_{SCL}$. The memory bank Q stores the sentence representations $\mathbf{z}$ and their corresponding labels y, $S^{'} = \{\mathbf{z}_i, y_i\}_{i=1}^{N}$, during computation of one batch. The maximum size of Q is denoted as M, which indicates that when the number of examples in Q exceeds M, only the last M examples will be kept and previous examples will be discarded. To avoid repeated computation, for each batch S, all the examples of the current batch S are deemed as anchor features. The computation of SCL with memory bank Q still follows Equation 2, except that the number of contrasting features is increased. Let S_C denote the set of contrasting features, which

is the union of the current batch S and memory bank Q, i.e., $S_C = S \cup Q$. Then the set of positive examples becomes $P(i)' \equiv \{\mathbf{z}_j \in S_C, y_j = y_i\}$ and the union of positive and negative examples becomes $A(i)' \equiv \{\mathbf{z}_j \in S_C, j \neq i\}$. Given batch size N and memory bank size M, the number of anchor features is N, whereas the number of contrasting features becomes $N + M$, since the set of contrasting features S_C is the union of the current batch S and the memory bank Q. We provide the pseudo-code for our approach in Algorithm 1.

Algorithm 1: Algorithm for supervised contrastive learning with memory bank

Input: Batch size N, encoder f, classifier g, memory bank Q

for $t \leq T_{max}$ **do**
- Sample minibatch $S = \{x_i, y_i\}_{i=1}^{N}$
- $\mathbf{z} = f(PrLM(x))$;
- $Q = None$;
- $\mathcal{L}_{CE} = \frac{1}{N}\sum_{i=1}^{N} -y_i \log\big(g(\mathbf{z}_i)\big)$;
- $S_C = S \cup Q$;
- $P(i) \equiv \{\mathbf{z}_j \in S_C, y_j = y_i\}$;
- $A(i) \equiv \{\mathbf{z}_j \in S_C, j \neq i\}$;
- $\mathcal{L}_{SCL} = \sum_{\mathbf{z}_i \in S} \frac{-1}{N} \sum_{\mathbf{z}_p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{\mathbf{z}_a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}$;
- $EnqueueAndDequeue(Q, \{\mathbf{z}_i, y_i\}_{i=1}^{N})$;
- $\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{SCL}$;
- update network by combined loss $\mathcal{L}$

end

5 Experiments

5.1 Dataset Statistics

We have conducted experiments on two benchmarks for cross-domain and cross-lingual sentiment classification and one benchmark for rumour detection.

Multi-Domain Sentiment Dataset (Blitzer et al., 2007) This dataset contains 8,000 Amazon product reviews, equally distributed from four domains: books (B), DVDs (D), kitchen and housewares (K), and electronics (E). In each domain, there are 1,000 positive and 1,000 negative reviews. We follow the split of prior works (Ganin et al., 2016; Du et al., 2020; Guo et al., 2020) for fair comparison, resulting in 1,600 training examples and 400 test examples for each domain. Since we do not have access to target domain data, training and model selection are all based on the mixture distribution of source domains.

Cross-Lingual Sentiment Dataset (Prettenhofer and Stein, 2010) This is a multi-lingual and multi-domain Amazon review dataset. It contains four languages: English (En), German (De), French (Fr), and Japanese (Jp). For all languages, there are three domains: books (B), DVDs (D), and music (M). For each domain, there are 2,000 training examples and 2,000 test examples, where each set contains 1,000 positive and 1,000 negative reviews. In summary, there are twelve language-domain combinations in this dataset.

PHEME Rumour Detection Dataset There are 5,802 annotated tweets from 5 different events ((C)harlie(H)ebdo, (F)erguson, (G)erman(W)ings, (O)ttawa(S)hooting, and (S)ydneySiege) labeled as rumour or non-rumour (1,972 rumours, 3,830 non-rumours).

5.2 Experimental Setup

We conducted experiments in cross-domain (CD), cross-language (CL), and cross-language cross-domain (CLCD) settings. For CD experiments, the model is trained on three source domains and evaluated on the target domain. Model selection is based on validation performance on the combined test set of the source domains. Using the multi-domain Amazon review dataset as an example, since there are three source domains in total, for each target domain experiment, there are 4,800 training examples.

In CL experiments, the training and test domains are the same while the source languages are different from the target language. For example, for target German-DVD, the training data will be English, French, and Japanese DVD, and the total number of training examples is 6,000. Since English is a high-resource language, we did not conduct experiments with English as the target language.

For CLCD experiments, the training data for a given language-domain combination come from both a different language and a different domain. For example, for the same target German-DVD, the training data will still be in English, French, and Japanese, while the domains for the training data will be books and music. For a fair comparison with CL experiments, the training data for CLCD experiments will be down-sampled to half of its original size, with 6,000 reviews in total.

In the monolingual domain generalization ex-

	D	E	K	B	Avg Acc	CH	F	GW	OS	S	Avg μF1
Guo et al. (2018)	87.70	89.50	90.50	87.90	88.90	-	-	-	-	-	-
Wright and Augenstein (2020)	88.90	90.30	90.80	90.00	90.00	67.90	45.40	74.50	62.60	64.70	63.02
DistilBERT											
Baseline	$89.30_{0.3}$	$89.80_{0.2}$	$89.98_{0.2}$	$89.24_{0.2}$	89.58	$64.78_{1.3}$	$43.03_{1.5}$	$69.87_{1.9}$	$60.42_{0.8}$	$62.02_{1.4}$	60.02
MMD	$89.00_{0.2}$	$89.86_{0.2}$	$89.64_{0.2}$	$89.38_{0.4}$	89.47	$63.80_{1.0}$	$43.44_{1.1}$	$69.04_{2.1}$	$63.97_{1.1}$	$63.27_{0.7}$	60.70
MoE	$89.20_{0.3}$	$89.92_{0}.3$	$90.26_{0.3}$	$89.88_{0.1}$	89.82	$65.84_{2.2}$	$43.61_{1.1}$	$72.23_{1.2}$	$61.63_{1.0}$	$64.25_{1.4}$	61.51
Intra	$88.46_{0.6}$	$89.80_{0.3}$	$90.06_{0.2}$	$89.22_{0.4}$	89.39	$64.14_{1.4}$	$42.89_{1.2}$	$70.77_{1.6}$	$61.84_{1.5}$	$62.41_{0.7}$	60.41
Adv	$88.40_{0.4}$	$89.60_{0.2}$	$90.00_{0.2}$	$89.04_{0.2}$	89.30	$64.83_{1.5}$	$42.23_{1.2}$	$65.94_{1.0}$	$61.47_{0.9}$	$62.81_{1.6}$	59.45
SCL	$89.35_{0.1}$	$89.85_{0.1}$	$90.25_{0.2}$	$89.50_{0.2}$	89.74	$65.57_{0.9}$	$43.22_{1.9}$	$73.03_{1.1}$	$63.50_{1.5}$	$63.52_{1.4}$	61.77
SCL+M=64	$\mathbf{90.10}_{0.2}$	$90.26_{0.2}$	$\mathbf{90.80}_{0.2}$	$\mathbf{89.98}_{0.2}$	**90.28**	$65.88_{0.8}$	$43.64_{1.2}$	$74.54_{1.1}$	$\mathbf{67.25}_{0.5}$	$\mathbf{65.99}_{1.7}$	63.46
SCL+M=128	$89.73_{0.3}$	$\mathbf{90.30}_{0.1}$	$90.50_{0.1}$	$90.04_{0.3}$	90.14	$\mathbf{68.08}_{1.0}$	$\mathbf{44.55}_{1.9}$	$\mathbf{75.41}_{0.8}$	$66.52_{2.0}$	$65.19_{0.8}$	**63.95**
SCL+M=256	$89.43_{0.5}$	$89.97_{0.3}$	$90.00_{0.2}$	$89.47_{0.3}$	89.72	$67.41_{0.8}$	$42.83_{0.9}$	$74.64_{0.7}$	$65.73_{2.0}$	$64.06_{1.5}$	62.93
Roberta-Large											
Basline	$90.00_{0.4}$	$93.95_{0.2}$	$93.40_{0.5}$	$92.65_{0.7}$	92.50	$67.12_{1.2}$	$\mathbf{43.97}_{2.4}$	$70.78_{3.2}$	$65.69_{2.9}$	$62.66_{1.7}$	62.04
MMD	$89.85_{0.4}$	$94.15_{0.3}$	$93.70_{0.5}$	$92.55_{0.5}$	92.56	$64.88_{1.7}$	$43.13_{1.2}$	$69.86_{2.9}$	$66.60_{0.7}$	$63.17_{2.9}$	61.53
MoE	$90.25_{0.3}$	$94.04_{0.4}$	$93.99_{0.2}$	$92.50_{0.2}$	92.69	$67.24_{1.8}$	$43.63_{2.2}$	$73.60_{1.8}$	$68.77_{2.4}$	$63.59_{1.2}$	63.38
Intra	$90.06_{0.3}$	$94.00_{0.3}$	$94.06_{0.2}$	$92.75_{0.2}$	92.72	$66.87_{2.2}$	$43.63_{2.2}$	$73.73_{1.6}$	$69.28_{1.8}$	$63.72_{1.0}$	63.44
Adv	$90.25_{0.7}$	$94.45_{0.5}$	$94.60_{0.3}$	$92.85_{0.6}$	93.04	$64.71_{2.4}$	$42.92_{1.0}$	$71.01_{1.6}$	$66.69_{2.2}$	$63.84_{3.7}$	61.83
SCL	$89.95_{0.3}$	$94.25_{0.7}$	$93.10_{0.6}$	$93.45_{0.5}$	92.69	$65.83_{2.0}$	$42.65_{1.5}$	$71.41_{1.7}$	$65.92_{1.5}$	$62.33_{2.0}$	61.63
SCL+M=64	$91.40_{0.7}$	$\mathbf{95.10}_{0.5}$	$95.05_{0.4}$	$93.25_{0.4}$	93.70	$67.44_{0.6}$	$43.53_{1.2}$	$\mathbf{75.23}_{0.5}$	$\mathbf{72.22}_{3.2}$	$\mathbf{67.18}_{2.0}$	**65.12**
SCL+M=128	$\mathbf{91.45}_{0.5}$	$\mathbf{95.10}_{0.4}$	$\mathbf{95.10}_{0.5}$	$\mathbf{93.7}_{0.5}$	**93.85**	$\mathbf{68.32}_{0.8}$	$43.39_{2.8}$	$73.89_{1.2}$	$72.01_{2.4}$	$66.70_{1.3}$	64.86
SCL+M=256	$91.10_{0.3}$	$95.05_{0.5}$	$94.90_{0.5}$	$93.40_{0.6}$	93.62	$66.81_{0.8}$	$40.91_{1.1}$	$73.52_{1.0}$	$72.21_{2.0}$	$65.59_{1.3}$	63.81

Table 1: Experimental results for cross-domain text classification. The reported metric is accuracy for sentiment analysis and micro-F1 for rumour detection. The experimental results are averaged over five runs and each subscript indicates the standard deviation of five runs.

Batch Size	**8**	**16**	**32**	**64**	**128**	**256**	**16+M=128**
Memory (GB)	10.5	14.1	21.3	35.6	64.3	121.7	14.7

Table 2: Memory consumption for training Roberta-large model for sentiment analysis. The memory sizes used above batch size of 32 are interpolated.

periments, we adopt Distil-Bert-base (Sanh et al., 2019) and RoBERTa-large (Liu et al., 2019) pre-trained models as encoders. In the cross-lingual experiments, we use XLM-R-large (Conneau et al., 2020) as encoder. We pre-process each review to 180 sentencepiece tokens. For all encoders, we set the dimension for classifier representation z to be 256. We train all the models with Adam Optimizer with learning rate of 1e-5. The batch size we use for training DistilBert/RoBERTa baselines is 16. We use batch size of 2 with 8 gradient accumulation steps for training XLM-R due to GPU memory constraint. All our experiments are conducted on Nvidia 16GB V100 GPU. We conduct grid-search for $M \in \{16, 32, 64, 128, 256, 512\}$ and $\tau \in \{0.1, 0.2, 0.5, 0.7, 0.8, 1.0\}$ in D,K,E-B transfer (i.e., D, K, E as source domains and B as the target domain) and select $M = 128$ and $\tau = 0.2$. We provide the memory consumption for training the Roberta model with different batch sizes in Table 2. We use interpolation estimate for memory usage above batch size of 32. From Table 2, we observe that our proposed method significantly saves memory consumption compared to directly using large batch sizes.

5.3 Compared Methods

We compare our methods with several state-of-the-art domain adaptation methods. However, previous works are trained with unlabeled target domain data, which is a more relaxed setup compared to ours. For fair comparison, we also employed state-of-the-art DA models in the same domain generalization setting as our baselines:

MoE (Guo et al., 2018; Wright and Augenstein, 2020): Mixture-of-Experts (MoE) models are the current state-of-the-art method for multi-source domain adaptation. It can also be applied to domain generalization. The MoE models consist of multiple models. For K source domains $\{S_i\}_1^K$, each source domain will be treated as a meta-target domain during training and each of them will have a dedicated model. The labeled data for each meta-target will be excluded during training of its model. There is an additional global encoder that is trained on labeled data from all source domains. Hence, there are $K + 1$ models in total for the MoE structure. During inference, the ensemble predictions of $K + 1$ models are aggregated.

CFd (Ye et al., 2020): This one-to-one domain adaptation approach is based on self-training. The model is first trained on a source domain, and high-confidence pseudo-labeled data in the target domain are generated for bootstrapping. We also compare to the optimal one-to-one transfer pair for this baseline.

CLDFA (Xu and Yang, 2017): This method, cross-lingual knowledge distillation on parallel corpora

	German			Avg	French			Avg	Japanese			Avg
	Book	DVD	Music		Book	DVD	Music		Book	DVD	Music	
Cross-language: *With unlabeled target data, results taken from original papers*												
CLDFA	83.95	83.14	79.02	82.04	83.37	82.56	83.31	83.08	77.36	80.52	76.46	78.11
MAN-MoE	82.40	78.80	77.15	79.45	81.10	84.25	80.90	82.08	62.78	69.10	72.60	68.16
CFd	93.95	91.69	93.89	93.18	94.25	93.79	93.39	93.81	89.41	88.68	89.54	89.21
Cross-language: *Without unlabeled target data*												
Baseline	$93.94_{0.3}$	$91.37_{0.4}$	$93.74_{0.4}$	93.02	$93.76_{0.5}$	$93.09_{0.4}$	$93.28_{0.4}$	93.38	$89.61_{0.4}$	$89.20_{0.3}$	$89.73_{0.6}$	89.51
Intra	$94.56_{0.4}$	$91.99_{0.5}$	$94.21_{0.6}$	93.59	$94.68_{0.2}$	$93.60_{0.2}$	$93.59_{0.2}$	93.96	$90.26_{0.2}$	$90.22_{0.4}$	$91.09_{0.3}$	90.52
Adv	$94.61_{0.4}$	$92.19_{0.3}$	$94.36_{0.3}$	93.72	$94.67_{0.3}$	$93.74_{0.3}$	$93.63_{0.5}$	94.01	$90.16_{0.3}$	$90.25_{0.3}$	$90.78_{0.5}$	90.40
SCL	$93.78_{0.5}$	$91.67_{0.5}$	$93.91_{0.2}$	93.12	$94.25_{0.7}$	$93.26_{0.2}$	$93.37_{0.3}$	93.63	$89.41_{0.3}$	$89.33_{0.8}$	$89.76_{1.1}$	89.50
SCL+M=64	$94.55_{0.4}$	$92.32_{0.3}$	$\mathbf{94.42}_{0.3}$	93.76	$94.58_{0.3}$	$\mathbf{94.05}_{0.2}$	$93.85_{0.2}$	94.16	$\mathbf{90.74}_{0.4}$	$90.48_{0.3}$	$\mathbf{91.34}_{0.4}$	**90.85**
SCL+M=128	$94.69_{0.4}$	$\mathbf{92.42}_{0.3}$	$94.20_{0.2}$	**93.77**	$\mathbf{94.70}_{0.3}$	$93.85_{0.3}$	$\mathbf{94.02}_{0.2}$	**94.19**	$90.52_{0.4}$	$\mathbf{90.67}_{0.4}$	$91.21_{0.3}$	90.80
SCL+M=256	$\mathbf{94.70}_{0.5}$	$92.19_{0.6}$	$93.88_{0.2}$	93.59	$94.54_{0.4}$	$93.66_{0.3}$	$93.87_{0.3}$	94.02	$90.55_{0.3}$	$90.53_{0.2}$	$90.97_{0.4}$	90.68
Cross-language and cross-domain *Without unlabeled target data*												
Baseline	$93.76_{0.3}$	$90.78_{0.3}$	$93.95_{0.4}$	92.83	$93.67_{0.3}$	$93.25_{0.2}$	$92.98_{0.2}$	93.30	$89.49_{0.5}$	$89.16_{0.4}$	$89.73_{0.5}$	89.46
Intra	$94.23_{0.3}$	$91.55_{0.3}$	$93.78_{0.2}$	93.19	$93.9_{0.3}$	$93.39_{0.3}$	$93.16_{0.2}$	93.48	$89.75_{0.4}$	$89.90_{0.4}$	$90.66_{0.3}$	90.1
Adv	$94.13_{0.3}$	$91.51_{0.3}$	$93.64_{0.3}$	93.09	$93.31_{0.4}$	$93.44_{0.3}$	$93.15_{0.2}$	93.3	$89.7_{0.4}$	$89.72_{0.4}$	$90.74_{0.3}$	90.05
SCL	$93.83_{0.4}$	$90.98_{0.3}$	$93.86_{0.2}$	92.89	$93.86_{0.3}$	$93.59_{0.3}$	$93.16_{0.2}$	93.54	$89.5_{0.3}$	$89.34_{0.4}$	$89.5_{0.3}$	89.45
SCL+M=64	$94.05_{0.3}$	$91.26_{0.2}$	$93.83_{0.4}$	93.04	$94.17_{0.2}$	$93.94_{0.3}$	$93.67_{0.2}$	93.93	$89.85_{0.3}$	$89.83_{0.4}$	$90.45_{0.2}$	90.05
SCL+M=128	$\mathbf{94.46}_{0.4}$	$\mathbf{91.90}_{0.3}$	$93.97_{0.2}$	**93.41**	$\mathbf{94.24}_{0.5}$	$93.79_{0.3}$	$\mathbf{93.95}_{0.4}$	**93.99**	$89.83_{0.6}$	$\mathbf{90.27}_{0.4}$	$\mathbf{91.02}_{0.4}$	**90.37**
SCL+M=256	$94.23_{0.5}$	$91.13_{0.2}$	$\mathbf{94.07}_{0.3}$	93.14	$94.06_{0.3}$	$\mathbf{93.96}_{0.4}$	$93.82_{0.2}$	93.95	$\mathbf{90.15}_{0.4}$	$89.86_{0.3}$	$90.25_{0.2}$	90.07

Table 3: Experimental results for Multilingual Amazon benchmark. Experiments are conducted in both cross-language (CL) and cross-language cross-domain (CLCD) settings. The reported metric is average accuracy and each subscript indicates the standard deviation of five runs.

for cross-lingual transfer learning, leverages translated Amazon reviews as a parallel corpus.

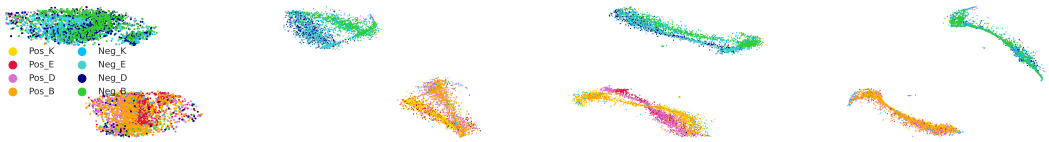
MAN-MoE (Chen and Cardie, 2018): This model uses a multinomial adversarial network to extract language-invariant features for sentiment classification. It studies cross-lingual transfer with multiple source languages. Besides, it also leverages MoE to focus on more transferable source languages.

Baseline fine-tunes pretrained language models (DistilBERT/Roberta/XLM-R) on labeled data from source domains and directly tests on the target domain. **MMD**: Following Li et al. (2018), pairwise Maximum Mean Discrepancy (MMD) losses among three source domains are added to cross-entropy loss. **Intra** refers to center loss used in Wen et al. (2016) and Ye et al. (2020), which maximizes the agreement between each example and its class center. **Adv** refers to the widely studied domain adversarial neural network (Ganin et al., 2016), where a gradient reversal layer is used to reverse the gradients calculated by the domain classification task. **SCL** adopts supervised contrastive loss from Gunel et al. (2021). It refers to directly applying supervised contrastive learning with a small batch size. Our model that enhances supervised contrastive learning (SCL) with memory bank is denoted as **SCL+M**. We provide experimental results for $M \in \{64, 128, 256\}$.

5.4 Experimental Results

The experimental results of CD text classification are shown in Table 1. We compare our methods with previous SOTA methods on multi-source domain adaptation and strong baselines of RoBERTa variants. Our findings are as follows. Firstly, directly applying supervised contrastive loss has limited improvement over the baseline performance of directly fine-tuning the pretrained language model, and the performance of **SCL** does not exceed previous domain adaptation methods, such as Mixture-of-Experts (**MoE**) and intra-class loss (**Intra**). Secondly, increasing the number of contrasting examples M significantly improves performance compared to directly using SCL. Our proposed method **SCL+M=128** achieves the best performance among the compared methods, exceeding the Roberta baseline by 1.35% in cross-domain sentiment analysis and by 2.82% in cross-domain rumour detection. Finally, the domain generalization method **MMD** performs poorly in the CD setting. This may be because the data distribution of computer vision tasks is different from that of sentiment analysis.

The CL and CLCD experimental results are shown in Table 3. In the CL and CLCD settings, we do not include the baseline result for MMD, as we observe high variance validation loss during training, and sometimes training diverges. We also do not include MoE for CL experiments, as the encoder for CL experiments is significantly larger than CD experiments and the MoE structure exceeds our hardware limit. From Table 3, our XLM-R baseline does not exceed the one-to-one self-training based DA approach **CFd**, showing



(a) RoBERTa (b) MoE (c) SCL (d) SCL+M=128

Figure 2: T-SNE visualization of the 256-dimensional sentence embeddings $\mathbf{z}$ for each model. The models are trained with B,K,E as source domains and D as the target domain. We sample 1,000 examples for each domain.

that knowledge of unlabeled target data is more important than increasing out-of-language training examples. In addition, in CL and CLCD settings, the performance of directly applying **SCL** is worse than intra-class loss (**Intra**) and adversarial training (**Adv**). We believe this is due to the small batch size (i.e., 2) during training of XLM-R. Hence, it is important to increase the contrasting examples for SCL in the small-batch setting. With increasing contrasting examples for SCL, we are able to achieve significant performance gain over our competitive baselines, and our best model **SCL+M=128** achieves state-of-the-art performance for both CL and CLCD settings. We observe performance drop when M is larger than 128. This is primarily due to the trade-off between the number of contrasting examples and their quality. Even though increasing the size of memory bank will benefit the lower bound of mutual information estimation, using an excessive number of prior examples will introduce noise for contrastive learning, since the text encoder has already been updated for many steps.

5.5 Analysis on Domain Divergence

To further analyze the performance of our model, we provide both intuitive visualization and quantitative analysis of domain discrepancy. Following Du et al. (2020) and Ben-David et al. (2010), we use $\mathcal{A}$-distance as the measurement for domain distance. To calculate $\mathcal{A}$-distance, we freeze the fine-tuned language model and the feed-forward layer f as encoder. Since the 256-dimensional $\mathbf{z}$ is used for the downstream classification task, we analyze domain discrepancy in this feature space. We sample two balanced sets of source examples and target examples with binary domain labels, i.e., source and target. Since we have multiple source domains, examples from each source domain will be down-sampled when calculating $\mathcal{A}$-distance, so that the total number of source examples and the number of target examples are balanced. This mixed dataset with binary domain labels will be split into two equal-size subsets, one for training and the other for testing. We then train a linear classifier with the first subset to distinguish source and target domain features. The error rate ϵ for this domain distinguishing classifier is calculated on the second subset, and we have the $\mathcal{A}$-distance as $d_{\mathcal{A}} = 2(1 - 2\epsilon)$.

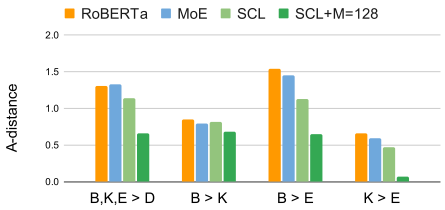


Figure 3: $\mathcal{A}$-distance of B,K,E to D generalization transfer.

We compare $d_{\mathcal{A}}$ of RoBERTa baseline, **MoE**, **SCL**, and **SCL+M=128** with B,K,E to D transfer, as well as $d_{\mathcal{A}}$ of source domain pairs. Results are shown in Figure 3. We see that the MoE model has little impact on reducing domain divergence for the backbone encoder. In contrast, the **SCL** and **SCL+M=128** models are able to reduce domain divergence and the latter achieves the lowest domain divergence compared to all other baselines.

To intuitively understand how our models overcome domain discrepancy, we also plot the t-SNE (Van der Maaten and Hinton, 2008) visualization of the features from different domains, as shown in Figure 2. For the RoBERTa baseline (Fig. 2a), we observe clear domain discrepancy within each sentiment cluster, and the sentiment cluster is relatively dispersed. Similarly, for the MoE model (Fig. 2b), we also observe domain discrepancy in the sentiment clusters, showing that the MoE objective does not improve the encoders' capability for producing domain-invariant representation. For the SCL baseline (Fig. 2c), we observe that the sentiment clusters are more concentrated and domain discrepancy is significantly reduced, but there is still some heterogeneity within the positive and negative review clusters. From Fig. 2d, we observe that increasing the contrasting features is able to further reduce domain divergence, which

is in line with the quantitative analysis of Fig. 3. We believe this is because **SCL+M** is trained with a large number of contrasting examples. By minimizing the intra-cluster distance in each batch, domain divergence within each sentiment cluster is reduced.

6 Conclusions and Future Work

In this paper, we study the under-explored domain generalization problem for text classification. We show that for cross-domain text classification, generalization performance from multiple source domains can exceed the best performance of one-to-one domain adaptation, even if the target domain is unknown during training. To this end, domain generalization is more practical and easier to deploy in realistic scenarios. To further improve the performance of cross-domain text classification, we propose an effective and memory-saving approach based on supervised contrastive learning for the domain generalization problem. We conduct extensive experiments in CD, CL, and CLCD settings. Experimental results have shown that our framework consistently outperforms strong baselines and the previous state of the art in all three experimental settings.

7 Acknowledgements

We would like to thank the reviewers for their insightful comments.

References

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. 2019. Maximum mean discrepancy gradient flow. In *Proceedings of NeurIPS*.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *TACL*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of NeurIPS*.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of NAACL*.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

X. Ding, Q. Shi, B. Cai, T. Liu, Y. Zhao, and Q. Ye. 2019. Learning multi-domain adversarial neural networks for text classification. *IEEE Access*.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of ACL*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *JMLR*.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised constrastive learning. In *Proceedings of ICML*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of NeurIPS*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of ICLR*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *Proceedings of AAAI*.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of EMNLP*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of ACL*.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. UDALM: Unsupervised domain adaptation through language modeling. In *Proceedings of NAACL*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of NeurIPS*.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. 2018. Domain generalization with adversarial feature learning. In *Proceedings of CVPR*.

Juntao Li, Ruidan He, Hai Ye, Hwee Tou Ng, Lidong Bing, and Rui Yan. 2020. Unsupervised domain adaptation of a pretrained cross-lingual language model. In *Proceedings of IJCAI*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of ACL*.

Qi Liu, Yue Zhang, and Jiangming Liu. 2018. Learning domain representation for multi-domain sentiment classification. In *Proceedings of NAACL*.

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*.

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 WWW*.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of ACL*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR*.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021a. Meta-learning for domain generalization in semantic parsing. In *Proceedings of NAACL*.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. 2021b. Generalizing to unseen domains: A survey on domain generalization.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of ECCV*.

Dustin Wright and Isabelle Augenstein. 2020. Transformer based multi-source domain adaptation. In *Proceedings of EMNLP*.

Fangzhao Wu and Yongfeng Huang. 2016. Sentiment domain adaptation with multiple sources. In *Proceedings of ACL*.

Ruochen Xu and Yiming Yang. 2017. Cross-lingual distillation for text classification. In *Proceedings of ACL*.

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 EMNLP*.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. In *Proceedings of NeurIPS*.

A Further Ablation Studies

	N=60	N=150	N=300	Full Data
RoBERTa	64.05 ± 9.83	76.10 ± 12.34	86.95 ± 4.16	93.70 ± 0.37
SCL	63.85 ± 8.33	76.40 ± 6.59	85.65 ± 4.47	94.25 ± 0.40
SCL+M=128	**78.75** ± 4.91	**83.65** ± 4.26	**90.95** ± 4.13	**95.10** ± 0.25

Table 4: Ablation study of few-shot cross-domain sentiment classification for B,K,D to E transfer.

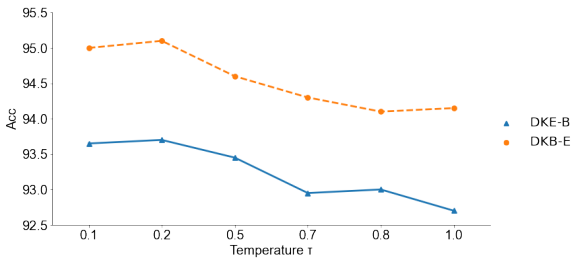


Figure 4: Ablation study of the effect of temperature τ. The reported accuracy is the average of 5 runs.

SCL in Few-Shot Setting. We conduct a few-shot learning experiment with B,K,D to E transfer. As shown in Table 4, SCL+M=128 achieves the most performance gain when the number of training examples is limited to 60, while conventional cross-entropy fails to converge. We believe that in the case of few-shot setting, the memory bank resembles a way for data augmentation, as the previously encoded sentences are reused as contrasting examples.

Effect of Temperature τ. We also provide hyper-parameter analysis for the scaling factor τ. We found that the performance of our model is negatively correlated to τ, and $\tau = 0.2$ works well empirically. We also found that when τ is small, the scale of supervised contrastive loss and cross-entropy validation loss is similar, leading to coherent optimization.

B Hyper-Parameters for the Baselines

We mainly conduct hyper-parameter tuning on the validation set of Multi-domain Sentiment Dataset. For model training, we tune the learning rates in {5e-6, 1e-5, 2e-5, 3e-5}, and batch size in {8, 16, 32}. The final learning rate is 1e-5 and scheduled linearly with training steps. We train our models for 10 epochs for cross-domain (CD), cross-language (CL), and cross-language cross-domain (CLCD) experiments. The batch size for CD experiments is 16. In CL experiments, the batch size is 2 with gradient accumulation step of 8 due to GPU memory constraint. For the experiment for CD sentiment analysis, the combined training size is 4,800. For the baseline experiments, we follow the methodology of Ye et al. (2020), that is adding a λ-weighted loss term to the cross-entropy loss. We show our choices for balancing parameters for losses as follows:

- **Intra-Class Loss** The weight for λ is tuned in {1, 0.5, 0.2, 0.1, 0.05}. We set λ to be 0.2 for CD experiments and 0.1 for CL experiments.
- **MoE** (Wright and Augenstein, 2020) We follow the MoE-Avg method in the original paper. For Roberta-large MoE models, we use 4 Nvidia V100 GPUs for training, as this model requires multiple encoders during training.
- **Adversarial Loss** Following the practice of prior works (Wright and Augenstein, 2020; Guo et al., 2018), the weight for adversarial training λ is 0.003.
- **MMD** We implemented a pair-wise MMD domain generalization approach with RBF kernel, following the practice of Li et al. (2018). The weight of MMD loss is tuned in {1.0, 0.5, 0.2, 0.1, 0.05} and set to 0.2.