Cross-Domain Sentiment Classification With In-domain Contrastive Learning

Tian Li[*]
Peking University
davidli@pku.edu.cn

Xiang Chen[*]
Peking University
caspar@pku.edu.cn

Shanghang Zhang[†]
UC Berkeley
shz@eecs.berkeley.edu

Zhen Dong[†]
UC Berkeley
zhendong@berkeley.edu

Kurt Keutzer
UC Berkeley
keutzer@berkeley.edu

Abstract

Contrastive learning (CL) has been successful as a powerful representation learning method. In this paper, we propose a contrastive learning framework for cross-domain sentiment classification. Specifically, we introduce in-domain contrastive learning and entropy minimization to enable to same classifier to be optimal across domains. Interestingly, we find through ablation studies that these two techniques behave differently in case of large label distribution shift and thus conclude that the best practice is to use in-domain CL or entropy minimization adaptively according to label distribution shift. The new state-of-the-art results our model achieves on standard benchmarks show the efficacy of the proposed method.

1 Introduction

Domain shift is common in language applications. One is more likely to find "internet" or "PC" in reviews on electronics than those on books, while he or she is more likely to find "writing" or "B.C." in reviews on books than those on electronics. This proposes a fundamental challenge to NLP in that many computational models fail to maintain comparable level of performance across domains. Formally, a distribution shift happens when a model is trained on data from one distribution (source domain), but the goal is to make good predictions on some other distribution (target domain) that shares the label space with the source.

We study unsupervised domain adaptation in this work, where we have fully-labeled data on source domain but no labeled data on target domain. The most prevailing methods in this field aim to learn domain-invariant feature by aligning the source and target domains in the feature space. The pioneering works in this field try to bridge domain gap with discrepancy-based approach. [31] first introduce MMD to measure domain discrepancy in feature space and [24] use its variant MK-MMD as an objective to minimize domain shift. Another line of work [14] introduces a domain classifier and adversarial training to induce domain invariant feature, followed by works using generative models to enhance adversarial training [19]. However, note that both MMD-based approach and adversarial training formulates with a minimax optimization procedure that is widely known as hard to converge to a satisfactory local optimum [13, 11]. Moreover, some recent works [8, 22, 40] have discovered that both of them don't guarantee good adaptation and will introduce inevitable error on target domain under label distribution shift because they may render incorrect distribution matching. For example, thinking of a binary classification task, the source domain has 50% of positive samples

[*]Equal Contribution
[†]Correspondence Author

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

and 50% of negative samples while the target domain has 30% postive and 70% negative. Successfully aligning these distributions in representation space requires the classifier to predict the same fraction of positive and negative on source and target. If one achieves 100% accuracy on the source, then target accuracy will be at most 80%, that is 20% error at best.

Self-supervised representation learning could be a good workaround for this problem because it enforces predictive behaviour matching [22] instead of distribution matching. The main idea is to learn discriminative representation that is able to genenralize across domains. [38, 42, 41, 43] use sentiment-indicating pivot prediction as their auxiliary task for cross-domain sentiment analysis. The method proposed in this paper adopts contrastive learning to extract generalizable discriminative feature. Contrastive learning is a subclass of self-supervised learning that is gaining popularity thanks to recent progress [17, 6, 4, 5]. It utilizes positive and negative samples to form contrast against the queried sample on pretext tasks in order to learn meaningful representations. However, the pretext tasks must be carefully chosen. [33] shows with experiments on computer vision tasks that the transfer performance will suffer under improper pretext tasks like pixel reconstruction.

Therefore, in this paper we explore two classic data augmentation methods in natural language processing—synonym substitution and back translation to define our pretext task. Experiments on two cross-domain sentiment classification benchmarks show the efficacy of the proposed method. We also examine whether in-domain contrastive learning and entropy minimization [34] helps cross-domain sentiment classification under varied label distribution settings. Our main contributions in this work are summarized as follows:

- We are the first to introduce contrastive learning to domain adaptation for NLP to the best of our knowledge.
- We introduce in-domain contrastive learning and entropy minimization for cross-domain sentiment classification and adaptively apply them according to label distribution shift across domains for best accuracy performance.
- We carefully choose pretext tasks for contrastive learning to promote generalizable discriminative representation learning.
- We beat strong baselines on standard benchmarks of domain adaptation for sentiment analysis.

2 Related Work

Cross-Domain Sentiment Classification The most prevailing methods for unsupervised domain adaptation for sentiment classification aims to learn domain-invariant feature by aligning the source and target domains in feature space. One line of work in this field derives from [24], using MMD and its variants to measure and minimize domain discrepancy[32, 18]. Another line of work follows [14], using a domain classifier and adversarial training to induce domain invariant feature [7, 23, 15]. However, these methods fail to take care of label shift across domains. This can cause undesired performance degradation on target domain according to our analysis in the introduction section. Another important line of work follows Structure Correspondence Learning [3]. They use pivot prediction as an auxiliary task to help extract domain-invariant knowledge [30, 38, 41, 42, 43]. Since transformer-based language models catch on, [10] designs novel self-supervised "post-training" tasks for BERT [9] along with domain adversarial training to help domain transfer. Please note that although we also use BERT as feature extractor, we're different from it in multiple aspects: We not only use contrastive learning instead of BERT-pretraining-style self-supervised learning as is in [10], but we also get rid of distribution matching with domain adversarial training. Our competitive performance on benchmarks further illustrates the efficacy of our model.

Contrastive Learning Recent developments on contrastive learning [17, 6, 4, 5] have achieved promising results on the standard representation learning benchmarks on computer vision tasks. Although there have been several works applying contrastive learning to NLP, most of them [16, 12, 37, 21, 35] concentrate on single-domain tasks like image caption retrieval, machine translation and those on the GLUE benchmark. To the best of our knowledge, we first adopt contrastive learning as an approach to facilitate domain adaptation in natural language processing.

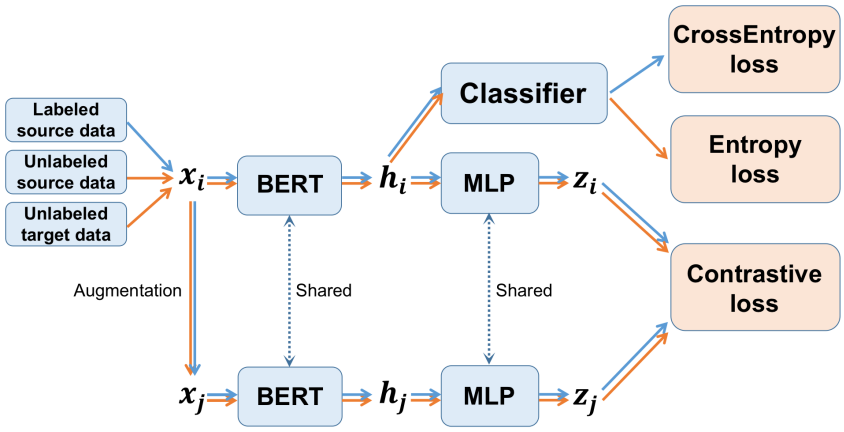


Figure 1: The whole pipeline of the proposed method: for each input document x_i, we first generate its positive sample x_j with augmentation. x_i and x_j are then forwarded to the shared feature extractor BERT to get the hidden features h_i, h_j and a shared MLP projection head is applied subsequently to get the projected representations z_i, z_j. The projecion, along with that of other samples in the minibatch as negatives (omitted here for simplicity), are used to compute the contrastive loss. To induce more discriminative feature, the hidden feature are fed into the sentiment classifier to get the logits. Finally, for the unlabeled instances, they are used to compute the *entropy*, while for the labeled instances, they are used to compute the *cross entropy loss*. The model is jointly trained on the three objectives.

3 Method

3.1 Contrastive Learning Framework

We propose a framework based on [4] but we elaborate it to fit the domain adaptation setting where we have unlabeled data from two domains(source and target) plus labeled data on source. As illustrated in Figure 1, our framework consists of the following components:

- **Positive sample generation:** We explore two classic data augmentation methods in natural language processing: synonym substitution and back translation, respectively, to generate positive sample x_j for each document x_i.
- **Feature extractor:** We use pretrained BERT[9] as our feature extractor because of its remarkable accomplishments in language understanding. Note that it's shared among all data, including their augmentations. It computes on x_i, x_j independently and outputs hidden features h_i, h_j.
- **Projection head:** Following SimCLR [4], we adopt a MLP with one hidden layer applying on h_i, h_j to get the projected representations z_i, z_j respectively. We find that this MLP projection benefits our model in terms of learning better discriminative feature with the contrastive loss, as is found in SimCLR [4].
- **Contrastive loss:** We introduce in-domain contrastive loss [20] based on the InfoNCE loss [4] in this work. Details are discussed in section 3.2.
- **Sentiment classifier:** We use another one-hidden-layer MLP as our sentiment classifier to generate the predictions. The logits are then used to compute the information entropy as the *entropy loss* for the unlabeled instances, while for the labeled instances, they are used to compute the ordinary cross entropy loss w.r.t the ground truth sentiment labels. Intuitions for the *entropy loss* are discussed in section 3.3.

3.2 In-Domain Contrastive Loss

Contrastive loss function is designed to maximize the similarities between positive pairs and minimize the similarities of negative ones. At each iteration of the training, we randomly sample a minibatch

$\{\boldsymbol{x}_i\}_{i=1}^N$ of N samples and generate positive pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}_{i+N})\}_{i=1}^N$ for each of them. As proposed by [4, 16], for each positive pairs, we treat other $2(N-1)$ samples as their negative examples. The contrastive InfoNCE loss function is defined as:

$$\mathcal{L}_{con}(\boldsymbol{z}) = \frac{1}{2N}\sum_{k=1}^{2N}[l(\boldsymbol{z}_{2k-1}, \boldsymbol{z}_{2k}), l(\boldsymbol{z}_{2k}, \boldsymbol{z}_{2k-1})] \tag{1}$$

$$l(\boldsymbol{z}_i, \boldsymbol{z}_j) = -\log \frac{\exp(\textit{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\textit{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} \tag{2}$$

where $\mathbb{1}_{[k\neq i]}$ is an indicator function equaling to 0 iff $k = i$, and $\textit{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\|\|\boldsymbol{z}_j\|)$ denotes the cosine similarity between hidden representations $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. τ is temperature parameter.

However, we noticed that contrastive loss of this form is not suitable to be directly applied to our domain adaptation setting. If we sample a minibatch that contains two instances $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from different domains, optimizing the contrastive loss above will even widen the distance between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. This will back-propagate to the hidden features $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$ thus enlarging the domain discrepancy in hidden space.

Therefore, we proposed to perform contrastive learning in the source and target domain independently, i.e. in-domain contrastive learning. At each iteration we randomly sample N instances $\boldsymbol{x}^{(s)} = \{\boldsymbol{x}_i^{(s)}\}_{i=1}^N$ from source domain and N instances $\boldsymbol{x}^{(t)} = \{\boldsymbol{x}_i^{(t)}\}_{i=1}^N$ from target domain. The in-domain contrastive loss is defined as the sum of contrastive loss from both domains:

$$\mathcal{L}_{con} = \mathcal{L}_{con}(\boldsymbol{z}^{(s)}) + \mathcal{L}_{con}(\boldsymbol{z}^{(t)}) \tag{3}$$

where $\boldsymbol{z}^{(s)}, \boldsymbol{z}^{(t)}$ is the projected representation of $\boldsymbol{x}^{(s)}, \boldsymbol{x}^{(t)}$.

In section 4.2 , we show that in-domain contrastive learning boosts domain transfer performance especially when the target domain has significant inbalanced label distribution.

3.3 Entropy Minimization

Note that in unsupervised domain adaptation we have labels on the source domain. Therefore it's easy to train an optimal classifier for the source domain. Despite lack of label on the target domain, we can still align the optimal classifier for the target domain with the source [1]. To this end, we minimize the entropy of the model's prediction in order to disambiguate the positive and negative instances over the unlabeled data, especially on the target domain [34]. In this way, the margin between positive and negative clusters on target domain will be widened so that there is a larger chance that the optimal decision boundary for the source domain falls within it.

However, our model fails when applying the entropy loss starting from the first epoch. This is probably because the model needs to get a general picture of the source and target domains with contrastive learning in the first place and draw the decision boundary for the source domain. Applying the entropy loss too hastily asks the model to early decide labels for uncertain instances before the model has learnt good representations and the classification boundary on source domain. Therefore, we apply the entropy loss from the second epoch and it works out as expected.

4 Experiments

4.1 Experiment Setting

Cross-domain Sentiment Classification To demonstrate the efficacy of our model, we conduct experiments on the Amazon-review[3] dataset [2] and Airlines dataset[4]. We follow [42] to introduce the Airlines dataset because the label distribution on this domain is relatively more balanced compared to the domains in the Amazon-review dataset. As is shown in Table 1, this dataset contains reviews for four kinds of products, books, DVD, electronics, and kitchen, each defining a domain. Each domain has 1000 sample labeled as positive and negative respectively, thus the 2000 labeled data. But the

[3]Dataset can be found at http://www.cs.jhu.edu/ mdredze/datasets/sentiment/index2.html

[4]Dataset and process procedures can be found at https://github.com/quankiquanki/skytrax-reviews-dataset

number of unlabeled data and its label distribution on each domain is different, as is shown in the last two columns of the table [42]. Note that the ratio of positive over negative on the domains in this dataset is significantly large. However, in the airline domain there are only 1.15 positive reviews for every negative review. Therefore we follow [42] to introduce this dataset to demonstrate our model's performance when the target domain's label distribution varies across a wide range. To align the setting with the Amazon-review dataset, we randomly sample 1000 positive review and 1000 negative review from the airlines dataset as labeled data, and the rest of data are removed of labels as unlabeled data.

It's important to clarify that in terms of label distribution shift in this paper we care about the label distribution of *labeled* data on source domain and that of *unlabeled* data on target domain. This is because the optimal classifier is trained with only labeled data on source domain but we want the unlabeled data on target domain to share the same optimal classifier. Also note that the labeled data of the domains involved in the experiments of this paper all have balanced label distribution, so the label distribution shift is only determined by target domain.

Domains	**labeled**	**unlabeled**	**pos:neg**
Books	2000	6000	6.43:1
DVD	2000	34741	7.39:1
Electronics	2000	13153	3.65:1
Kitchen	2000	16785	4.61:1
Airlines	2000	39396	1.15:1

Table 1: The statistics of domains on the Amazon-review dataset and the Airlines dataset reported by [42]. "pos:neg" denotes the ratio of unlabeled positive samples over unlabeled negative samples on that domain.

Baselines We mainly compare our model with the state-of-the-art method BERT-DAAT [10] on this benchmark. We also train BERT on the source labeled data and directly test on the target labeled data, constituting a second strong baseline called BERT-base. Results of two non-BERT methods HATN [23] and IATN [39] are also included.

Augmentation methods We use synonym substitution and back translation as augmentation methods to generate positives in this paper. For synonym substitution, we use the method provided by python nlpaug library [26] based on WordNet [28]. For back translation, we use the method provided by nlpaug library based on fairseq [29]. We do random synonym substitution online, meaning that it's different among repeated training runs. But We do back translation as offline preprocessing since it's slower. To be specific, we translate the English texts to German and then back.

Hyperparameter tuning We adopt the pretrained BERT-base-uncased model from huggingface [36] in this paper. We use ReLU as the activation function for both projection head and sentiment classifier. On the neural network training, we use the AdamW [25] optimizer with learning rate $2e-5$, linear learning rate scheduler, linear learning rate warm up, warm up steps 0.1 of total training steps and weight decay 0.01. We train the model for 4 epochs and set the temperature parameter of the contrastive loss to 0.05. For synonym substitution, we set the augmentation rate to 0.3 and remove max limit to the number of augmented words. For back translation, we set the beam parameter to 1.

4.2 Experiment Results

Comparison with baseline Table 2 shows our model's performance on the benchmarks. We find that back translation is generally better than synonym substitution as an augmentation method for contrastive learning. Although we don't beat BERT-DAAT in all settings, both of our models are able to surpass it on a average basis. Note that our strong baseline model BERT-base has comparable accuracy with BERT-DAAT on average.

Note that we adaptively choose techniques for our model according to label distribution shift across the source and target domains. For B→E and B→A settings we use entropy minimization and ordinary InfoNCE loss instead of in-domain contrastive loss because the label distribution shift is

relatively small. For K→D setting we apply in-domain contrastive loss but not entropy minimization because the label distribution shift is larger. The intuitions for the choice are discussed at length in the ablation studies.

S→ T	Previous Models		BERT			
	HATN	IATN	BERT-base	BERT-DAAT	BERT-DANN	BERT-CDA
B→ E	85.70	86.50	90.50	89.57	91.67	**91.98**
K→ B			88.50	87.98	89.38	**89.80**
K→ D	84.50	84.10	87.90	88.81	88.89	**89.09**
E→ K			94.20	93.18	**94.54**	94.39
K→ E			93.34	91.72	93.15	**93.65**
B→ K			92.46	90.75	92.86	**93.17**
Average	85.10	85.30	91.15	90.34	91.75	**92.01**
B→ A	–	–	86.18	–	86.66	**86.86**

Table 2: Sentiment classification accuracy on the Amazon-review dataset and Airlines dataset. We only evaluate on three domain settings as an initial trial. BERT-base denotes BERT trained on source and direcly tested on target, BERT-DAAT is the method proposed in [10]. BERT-CL^{bt} is our model with back translation as augmentation. BERT-CL^{ss} is our model with synonym substitution as augmentation. B, D, E, K denotes the 4 domains in Amazon-review dataset respectively and A denotes the airlines domain.

Ablation studies Table 3 shows the results of our ablation study on in-domain contrastive learning and entropy minimization. Interestingly, we find that in-domain contrastive learning gain much benefit on K→D domains but doesn't help on B→E domain setting.

In-domain contrastive learning promotes contrast between positive and negative clusters intra-domain, thus widens the decision margin and helps classification on target domain as well as source domain. On the contrary, ordinary contrastive learning additionally forms contrast inter-domain and thus increases domain discrepancy. This is better explained with visualization. Look at Figure 2b. The red and yellow dots respectively represent the feature of positive and negative samples on the source domain, and the blue and green dots represent those on the target domain. In-domain contrastive learning focuses on pushing red points away from yellow and green points away from blue, which desirably enlarge the margin between positive and negative. On the other hand, although ordinary contrastive learning does the same thing as above, it additionally pushes blue away from red and green away from yellow, which will potentially entangle the positive and negative clusters on the target domain.

However, the tradeoff here is that in-domain contrastive learning prevents the divergence of source and target domains in feature space at the cost of not fully exploiting the training data. Insufficient utilizing data will consistently cause performance loss. But the benefit brought by in-domain contrastive learning is larger as the label distribution shift across domains is larger. The underlying fact is that divergence between source and target domains will potentially push points near the decision boundary to cross it. In particular, when the label distribution shift is significant such as 1:1 to 7.39:1 in the K→D setting, there are more positive points near the decision boundary. Divergence between the

S→T	**in-domain**	**entropy loss**	**accuracy**
B→E			90.49
	✓		90.05
		✓	**91.48**
	✓	✓	91.36

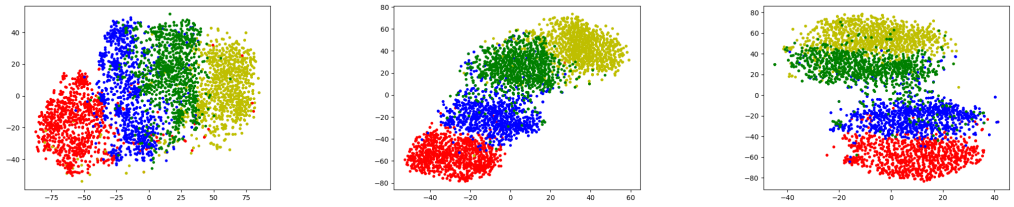
S→T	**in-domain**	**entropy loss**	**accuracy**
K→D			88.39
	✓		**88.99**
		✓	87.25
	✓	✓	86.36

Table 3: Ablation study of in-domain contrastive learning (section 3.2) and entropy minimization (section 3.3) on the Amazon-review dataset. Note that all experiments in this table use back translation as default augmentation method.

kitchen and DVD domains will make instances of the positive instances on DVD domain more prone to cross the optimal decision boundary for kitchen domain. But when the label distribution shift is not so large, this effect will not be so salient. The benefit of preventing divergence is small so as to be offset by the defect of insufficient utilization of training data. That's why we find in-domain contrastive learning doesn't work well on B→E setting.

In the meantime, entropy minimization also behaves differently for the B→E domain and the K→D domain. Although it helps greatly on B→E setting, it hurts the model's accuracy on K→D where the label distribution on the target domain is significantly unbalanced. This is in line with findings in [34] that entropy minimization may cause the model to exploit over-represented features, i.e. it will bias to the dominant domain and the dominant class in the domain. The induced bias is likely to undermine the model's performance in case of large label distribution shift.

Given that in-domain contrastive learning and entropy minimization behaves oppositely in case of large label distribution shift, we conclude that the best practice is to use one of them adaptively according to the scale of label distribution shift. Use entropy minimization when it's small but use in-domain contrastive learning when it's large.



(a) Baseline

(b) In-domain contrastive learning without entropy minimization

(c) In-domain contrastive learning with entropy minimization

Figure 2: t-SNE [27] projection of (a) BERT-base hidden feature, (b) hidden feature of BERT-CL without entropy minimization, (c) hidden feature of BERT-CL full model. The red and yellow dots respectively represent the feature of positive and negative samples on the source domain, and the blue and green dots represents those on the target domain. Note that the margin between positive cluster and negative cluster on target domain becomes clearer from left to right.

Visualization Figure 2 shows the hidden features of our model trained on the books as source and electronics as target, projected to 2D space with t-SNE [27]. Figure 2a shows that BERT-base can only learn discriminative feature on source domain, while Figure 2b shows that our BERT with in-domain contrastive learning is able to learn better and approximately align the optimal decision boundary on source and target domain. Finally, Figure 2c demonstrates that, further with entropy minimization, our model is able to disambiguate on the target domain and widen the gap between target positive and negative clusters.

5 Conclusion and Broader Impact

In this paper, we propose in-domain contrastive learning with entropy minimization to promote domain transfer. Our proposed model beats strong baselines and visualization results also show the efficacy of our model. Extensive ablation studies unveil how label distribution shift may interact with our model. It remains open question how to address label shift across domains. Besides, although synonym substitution works better than back translation in this paper, it's still interesting to explore more data augmentation methods and summarize what serve best for contrastive learning for cross-domain sentiment classification.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. Invariant risk minimization games, 2020.

[2] John Blitzer, Mark Dredze, and Fernando Pereira. 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, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[7] Stéphane Clinchant, Gabriela Csurka, and Boris Chidlovskii. A domain adaptation regularization for denoising autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 26–31, Berlin, Germany, August 2016. Association for Computational Linguistics.

[8] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, 2020.

[11] John C Duchi, John Lafferty, Yuancheng Zhu, et al. Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems*, pages 3423–3431, 2016.

[12] Hongchao Fang and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

[13] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

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

[15] Deepanway Ghosal, Devamanyu Hazarika, Navonil Majumder, Abhinaba Roy, Soujanya Poria, and Rada Mihalcea. Kingdom: Knowledge-guided domain adaptation for sentiment analysis. *arXiv preprint arXiv:2005.00791*, 2020.

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

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[18] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Adaptive semi-supervised learning for cross-domain sentiment classification. *arXiv preprint arXiv:1809.00530*, 2018.

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[20] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A. Plummer, Stan Sclaroff, and Kate Saenko. Cross-domain self-supervised learning for domain adaptation with few source labels, 2020.

[21] Tassilo Klein and Moin Nabi. Contrastive self-supervised learning for commonsense reasoning, 2020.

[22] Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

[23] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. volume 37 of *Proceedings of Machine Learning Research*, pages 97–105, Lille, France, 07–09 Jul 2015. PMLR.

[25] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[26] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[28] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[29] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

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

[31] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[32] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation, 2015.

[33] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[34] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Fully test-time adaptation by entropy minimization, 2020.

[35] Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages, 2020.

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, page arXiv:1910.03771, October 2019.

[37] Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] Jianfei Yu and Jing Jiang. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas, November 2016. Association for Computational Linguistics.

[39] Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5773–5780, 2019.

[40] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

[41] Yftah Ziser and Roi Reichart. Neural structural correspondence learning for domain adaptation, 2017.

[42] Yftah Ziser and Roi Reichart. 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: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[43] Yftah Ziser and Roi Reichart. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5895–5906, Florence, Italy, July 2019. Association for Computational Linguistics.