Domain Adaptation with BERT-based Domain Classification and Data Selection

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, Bing Xiang
AWS AI Labs
{xiaofeim, pengx, zhiguow, rnallapa, bxiang}@amazon.com

Abstract

The performance of deep neural models can deteriorate substantially when there is a domain shift between training and test data. For example, the pre-trained BERT model can be easily fine-tuned with just one additional output layer to create a state-of-the-art model for a wide range of tasks. However, the fine-tuned BERT model suffers considerably at zero-shot when applied to a different domain. In this paper, we present a novel two-step domain adaptation framework based on curriculum learning and domain-discriminative data selection. The domain adaptation is conducted in a mostly unsupervised manner using a small target domain validation set for hyper-parameter tuning. We tested the framework on four large public datasets with different domain similarities and task types. Our framework outperforms a popular discrepancy-based domain adaptation method on most transfer tasks while consuming only a fraction of the training budget.

1 Introduction

Modern deep NLP models with millions of parameters are powerful learners in that they can easily adapt to a new learning task and dataset when enough supervision is given. However, they are also very fragile when deployed in the wild since the data distribution and sometimes even the task type can be very different between the training and inference time. Domain adaptation (Csurka, 2017), a prominent approach to mitigate this problem, aims to leverage labeled data in one or more related source domains to learn a classifier for unseen or unlabeled data in a target domain.

Fine-tuning deep neural networks (Chu et al., 2016) is a popular supervised approach for domain adaptation in which a base network is trained with the source data, and then the first n layers of the base network are fixed while the target domain labeled data is used to fine-tune the last few layers of the network. However, this approach requires a significant amount of labeled data from the target domain to be successful.

While classical methods such as instance reweighting and feature transformation (Pan and Yang, 2010) are among the most popular and effective early solutions of domain adaptation for classical machine learning algorithms, deep learning architectures specifically designed for domain adaptation is more promising for deep domain adaptation. The major idea in unsupervised domain adaptation is to learn a domain invariant representation (Wang and Deng, 2018) leveraging both labeled data from the source domains and unlabeled data from the target domain. Various methods and architectures have been proposed which often fall into discrepancy-based or adversarial-based domain adaptation categories. In discrepancy-based methods, domain discrepancy based on maximum mean discrepancy (MMD) (Smola et al., 2006) or Wasserstein Distance (Shen et al., 2017) defined between corresponding activation layers of the two streams of the Siamese architecture is often used as a regularization term to enforce the learning of domain non-discriminative representations. In adversarial-based approaches, which can be either generative or non-generative, the aim is to encourage domain confusion through an adversarial objective. In the generative approach, a Generative Adversarial Network (GAN) is used to generate synthetic target data to pair with synthetic source data to share label information (Liu and Tuzel, 2016). Inspired by GAN, in the non-generative approach, a domain confusion loss produced by the domain discriminator helps to learn the domain-invariant representations. For example, Ganin et al. implemented a domain-adversarial net-

Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo), pages 76–83
Hong Kong, China, November 3, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

work in which unsupervised domain adaptation is achieved by adding a domain classifier. The domain classifier is trained via a gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation. As the training progresses, the approach promotes the emergence of a representation that is discriminative for the main learning task and indiscriminate with respect to the shift between the domains. However, such type of models are usually hard to train since the optimization problem involves a minimization with respect to some parameters, as well as a maximization with respect to the others.

Very early approaches in NLP utilized instance re-weighting (Jiang and Zhai, 2007) and target data co-training (Chen et al., 2011) to achieve domain adaptation. Recently, Denoising Autoencoders (Glorot et al., 2011), domain discrepancy regularization and domain adversarial training (Shah et al., 2019; Shen et al., 2017) have been employed to learn a domain invariant representation for neural network models. Many domain adaptation studies have focused on tasks such as sentiment analysis (Glorot et al., 2011; Shen et al., 2017) , Part-Of-Speech (POS) tagging (Ruder et al., 2017a) and paraphrase detection (Shah et al., 2019), and tested on neural network models such as multilayer perceptron (MLP) and Long Short-term Memory (LSTM). In terms of multiple source domain adaptation, while some of the methods of single-source adaptation can be directly extended to the multiple sources case, models that specially designed for multiple sources domain adaptation such as the mixture of experts and knowledge adaptation (teacher-student network) (Ruder et al., 2017b) are more effective.

BERT model (Devlin et al., 2018) is one of the latest models that leverage heavily on language model pre-training. It has achieved state-of-the-art performance in many natural language understanding tasks ranging from sequence classification and sequence-pair classification to question answering. Although pre-trained BERT can be easily fine-tuned with just one additional output layer on a supervised dataset, sometimes the zero-shot transfer of the fine-tuned model from a source domain is necessary due to the very limited labeled data in the target domain. The performance of the fine-tuned BERT can deteriorate substantially if there is a domain shift between the fine-tuning and the test data (see section 4.3). Due to the complex attention mechanisms and large parameter size, it is hard to train BERT for domain adaptation using the domain-adversarial approach. Our initial experiments demonstrated the unsteadiness of this approach when applied to BERT. Unsupervised language model (LM) fine-tuning method (Howard and Ruder, 2018) consisting of general-domain LM pre-training and target task LM fine-tuning is effective using a AWD-LSTM language model on many text classification tasks such as sentimental analysis, question classification and topic classification. However, due to the unique objective of BERT language model pre-training (masked LM and next sentence prediction) which requires multi-sentences natural language paragraphs, unsupervised fine-tuning of BERT LM does not apply to many sentence-pair classification datasets.

In this work, we propose a novel domain adaptation framework, in which the idea of domain-adversarial training is effectively executed in two separate steps. In the first step, a BERT-based domain classifier is trained on data from different domains with domain labels. In the second step, a small subset of source domain data is selected based on the domain classifier for fine-tuning BERT. The order of presentation of the selected source domain data to the model learner (training curriculum) also plays an important role and is determined by the point-wise domain probability. We demonstrate the effectiveness of our framework by comparing it against an MMD-based domain adaptation method and a naive zero-shot baseline. Our method achieved the best performance on most transfer tasks while only consuming a portion of the training budget.

2 Related Work

Our method is inspired by the work on curriculum learning and recent work on data selection for transfer learning.

Curriculum Learning: Curriculum Learning (Bengio et al., 2009) deals with the question of how to use prior knowledge about the difficulty of the training examples, to boost the rate of learning and the performance of the final model. The ranking or weighting of the training examples is used to guide the order of presentation of examples to the learner. The idea is to build a curriculum of progressively harder samples in order to significantly accelerate a neural network's training. While curriculum learning can leverage label information (loss of the model, training progress) (Weinshall and Amir, 2018) to guide data selection, this work assumes no or few labeled data in the new domain.

Data Selection: Not all the data points from the source domain are equally important for target domain transfer. Irrelevant source data points only add noise and overfit the training model. Recent work from Ruder and Plank, applied Bayesian optimization to learn a scoring function to rank the source data points. Data selection method was also used by Tsvetkov et al. to learn the curriculum for task-specific word representation learning, and by Axelrod et al.; Duh et al. for machine translation using a neural language model.

3 Approach

In this section, we propose a domain adaptation framework based on domain-discriminative data selection. Specifically, instead of training a deep neural network model in a domain-adversarial way, we effectively execute the idea in two separate steps. In the first step, we train a domain classifier with the same model architecture on the data from different domains with domain labels. In the second step, we select a subset of source domain data based on the domain probability from the domain classifier, and train the original model on the selected source data. We further design the training curriculum by presenting first the data points that are most similar to the target domain as ranked by the domain probability. Compared with the integrated training of domain classifier and task classifier based on batch-wise input of source and target data, the advantage of our two-step approach is that all the source data can be ranked at the same time and only the source data that are most similar to the target domain are selected for training the task classifier. We apply this framework to the domain adaptation of the fine-tuned BERT model.

BERT Domain Classifier BERT representations are very discriminative of texts from different domains due to the extensive language model pre-training. A t-SNE plot of BERT embeddings is presented at Figure 3, on which the data points from different domains are grouped into well-separated regions. In order to effectively select source data that is most similar to the target domain distribution, we train a BERT-based domain classifier on mixed data points with domain labels. The probability score from the domain classifier quantifies the domain similarity.

Learning Curriculum As demonstrated in many curriculum learning papers, the order of training data presented to the learning algorithm plays an important role in convergence rate and final model performance. The idea is to build a curriculum of progressively harder samples so that a neural network can learn from easy samples first and gradually adjust its parameters. As part of the proposed domain adaptation framework, we propose a learning curriculum based on the domain probability from the domain classifier. In the context of domain adaptation, an "easy" source sample is a sample very similar to the target domain data, while a "hard" sample is a sample very different from the target domain data.

Domain Regularization Method We compare our framework with a popular domain adaptation method: MMD-based domain regularization. Specifically, we enforce domain regularization by minimizing the maximum mean discrepancy (MMD) in the BERT latent space between the source and target domains. Formally, the squared MMD between the probability distributions P and Q in the reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS) with kernel k is defined as:

$$d_k^2(P,Q) := \|\mathbf{E}_P[x] - \mathbf{E}_Q[x]\|^2_{\mathcal{H}_k}$$

With that, we have the following domain regularized training objective for the BERT model:

$$\min_\theta \frac{1}{S} \Sigma_{(x_i,y_i)\in S}\ \mathscr{L}(x_i, y_i; \theta) + \lambda \cdot d_k^2(D_s, D_t; \theta)$$

where S is the collection of labeled source domain data, and λ is the regularization parameter. We choose rational quadratic kernel of the form:

$$k(x,x') = \sigma^2 (1 + \frac{(x-x')^2}{2al^2})^{-\alpha}$$

as the characteristic kernel in the experiment. The lengthscale l determines the length of the "wiggles" in the function. The parameter α determines the relative weighting of large-scale and small-scale variations.

4 Experiment

In this section, we conduct both qualitative and quantitative studies of the proposed method, and compare its performance against the MMD-based domain regularization method and naive zero-shot transfer from the source domain. In the experiments, in order to determine the optimal number of data points selected from the source domain, we set aside a small target domain dataset for validation. Starting from only a hundred examples, we double the training data size every time we observe a significant change in transfer performance evaluated on the validation set.

4.1 Datasets

We tested our framework on four large public datasets across three task categories: natural language inference (SNLI and MNLI), answer sentence selection (QNLI) and paraphrase detection (Quora). Large datasets usually have a much smaller variance in evaluation metrics compared with smaller datasets. We used the pre-processed datasets from GLUE natural language understanding benchmark (Wang et al., 2018). A summary of the dataset statistics and the details of the experiment setup are presented in Table 1.

Task Category	Dataset	Train Size	Dev Size
Natural Language Inference	SNLI	510,711	9,831
Natural Language Inference	MNLI	392,702	9,815
Answer Sentence Selection	QNLI	108,436	5,732
Paraphrase Detection	Quora	363,847	40,430

Table 1: Summary of the datasets

SNLI The Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) is a collection of 570k human-written English sentence pairs supporting the task of natural language inference. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). In order to make the label set the same across all the datasets, we convert the original three-label classification task into a binary classification task with “entailment” as the positive label, and “contradiction” and “neutral” as negative.

MNLI The Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2017) is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus but differs in that it covers a range of genres including transcribed speech, fiction, and government reports, and supports a distinctive cross-genre generalization evaluation. We used the training data from GLUE but evaluate only on the matched (in-domain) section. Similar as in SNLI, we convert the three-label classification task into a binary classification task.

QNLI The Question-answering Natural Language Inference (QNLI) is a dataset converted from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Although its name contains “natural language inference”, the text domain and task type of QNLI are fundamentally different from those of SNLI and MNLI. The original SQuAD dataset consists of question-paragraph pairs, where one of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the corresponding question (written by an annotator). GLUE converts the task into sentence pair classification by forming a pair between each question and each sentence in the corresponding context and filtering out pairs with low lexical overlap between the question and the context sentence. The task is to determine whether the context sentence contains the correct answer to the question.

QQP The Quora Question Pairs (QQP) dataset is a collection of question pairs from the comunity question-answering website Quora (Wang et al., 2017). The task is to determine whether a pair of questions are semantically equivalent. One source of negative examples are pairs of “related questions” which, although pertaining to similar topics, are not truly semantically equivalent. Due to community nature, the ground-truth labels contain some amount of noise.

4.2 Experiment Setup

BERT Domain Classifier The setup for training the BERT domain classifier is shown in Figure 1. Basically, the setup is similar to that for fine-tuning BERT on sequence-pair classification task (since we test our method on sequence-pair classification tasks). We take the hidden state of the [CLS] token of the input sequence pair, and feed it into a two-layer feedforward neural network with hidden units of 100 and 10 in each layer and ReLU as the activation function. The label for each data point is the domain that the data point belongs to.

MMD-based Domain Regularization The goal of domain regularization is to train the BERT model on the source domain but learn domain-invariant latent representations. The computation pipeline of training BERT model using MMD-based domain regularization is presented in Fig-

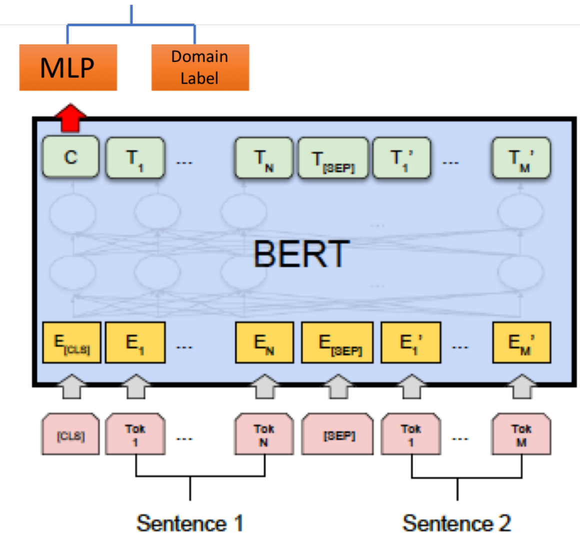

Figure 1: Setup for training a BERT domain classifier. Picture adapted from (Devlin et al., 2018)

ure 2. Basically, we feed both labeled source domain data and unlabelled target domain data to the model, a classification loss is calculated based on source labels and model prediction, and an MMD domain loss is calculated from the BERT representations of source domain data and target domain data. We combine the two losses as the training objective. It is straightforward to optimize this objective using stochastic gradient methods.

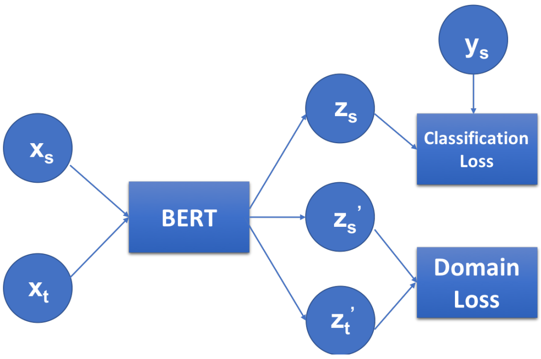

Figure 2: Setup for BERT domain adaptation with MMD-based domain regularization.

Experiment Details The experiments were conducted in three phases. In the first phase, a BERT-based domain classifier is trained to distinguish samples from a pair of datasets. In the second phase, all source domain training samples are ranked based on the output from the BERT domain classifier, and a subset of data points is selected from the source domain training set. The selected subset of source data and their ground truth labels are then used to fine-tune a BERT model in the final phase.

We train one binary domain classifier for each pair of source-target datasets. For each dataset, $5,000$ data points were randomly selected to make up the training set, and another $1,000$ data points were sampled as the test set. We train the BERT domain classifier for a fixed step of 100, using a small learning rate of $2e-6$ and batch size of 64. Due to the domain discriminative nature of pre-trained BERT representations, the BERT domain classifier can easily achieve an accuracy $> 99\%$ domain classification performance on the holdout test dataset.

The trained domain classifier is then used to predict the target domain probability for each data point from the source domain. Source data points with the highest target domain probability are selected for fine-tuning BERT for domain adaptation. For each target domain, we set aside a small validation set (1 percent of the target training set) for tuning the hyper-parameters such as batch size. We incrementally increase the size of the selected source data. For each batch size and number of selected source data combination, we fine-tuned the BERT model for 10 epochs, and record the best performance for each configuration.

4.3 Results

BERT Representations The fact that pre-trained BERT can be easily fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks suggests that BERT representations are potential universal text embeddings. In order to visualize BERT representations, we randomly select $5,000$ training samples from each dataset and extract the BERT embeddings of them. Figure 3 presents the t-SNE plot of the BERT representations. As we can see from the figure, data points from different datasets are grouped into well-separated regions. This shows that BERT is extremely effective at mapping text from different domains to different locations within its representation space.

Transfer Performance The transfer performance of different methods is presented in Table 2. As the first baseline, we evaluate the performance of naive zero-shot transfer of fine-tuned BERT models. The results are presented in the column “NZS”. Each fine-tuned BERT model is trained to convergence using all the source domain data, and zero-shot transferred to the target do-

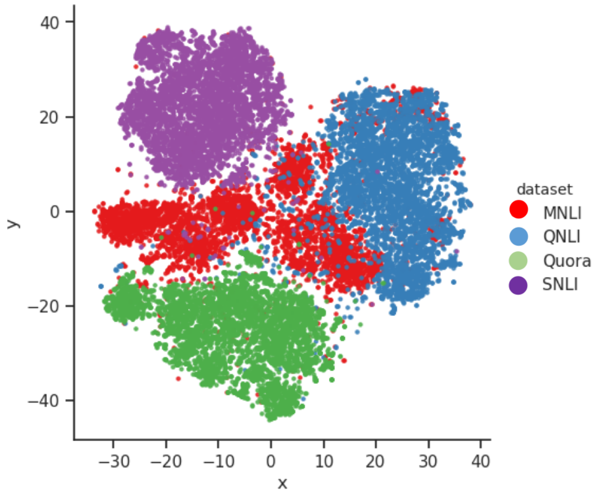

Figure 3: t-SNE plot of [CLS] token activations from the second-to-last encoder of BERT for four datasets in this paper. Second-to-last layer is used since the last layer embeddings may be biased to the target of BERT pre-training tasks.

Source	Target	IFT	NZS	MMD	DDS	% Data
MNLI	QNLI	85.3	49.8	58.0	**58.5**	0.1 %
MNLI	Quora	89.3	73.7	71.5	**73.9**	26.1 %
MNLI	SNLI	92.9	87.4	87.6	**88.3**	26.1 %
QNLI	MNLI	88.3	63.7	66.0	**67.2**	0.4 %
QNLI	Quora	89.3	61.8	66.1	**67.6**	1.5 %
QNLI	SNLI	92.9	56.1	65.3	**66.6**	0.4 %
Quora	MNLI	88.3	71.0	70.6	**83.6**	3.5 %
Quora	QNLI	85.3	50.8	58.8	**59.1**	1.8 %
Quora	SNLI	92.9	69.1	**72.4**	71.6	1.8 %
SNLI	MNLI	88.3	77.0	**82.2**	80.2	5.0 %
SNLI	QNLI	85.3	49.0	54.9	**56.7**	0.1 %
SNLI	Quora	89.3	67.0	70.8	**70.9**	1.3 %

Table 2: Transfer performance (accuracy) of different domain adaptation methods. “IFT”: in-domain fine-tuning. “NZS”: naive zero-shot. “MMD”: MMD-based domain regularization. “DDS”: discriminative data selection. “% Data”: percentage of source domain data selected in DDS method.

main. While in-domain fine-tuned BERT models usually achieve state-of-the-art performance, their zero-shot performance on the target domain can be significantly degraded. For transfers between dissimilar domains such as SNLI to QNLI, naive zero-shot can lead to more than 40% loss in accuracy compared with in-domain supervised training. By learning a domain invariant representation, the MMD-based domain adaptation method (column “MMD”) significantly outperforms the naive zero-shot baseline in almost all the transfer tasks. However, our discriminative data selection method (column “DDS”) achieves the best transfer performance in 10 out of the 12 pairwise transfer tasks while training on only a small fraction of source domain data (column “% Data”). The relative improvement is as large as 18% over the naive zero-shot and 3.3% over the MMD-based domain regularization. Even though we doubled the training data size every time we observe an increase in transfer performance, the cumulative training time is still much smaller than fine-tuning on the whole source dataset. Compared with the batch-wise iterative adaptation or regularization techniques, our method ranks all the source domain data at the same time, and the learner is trained on the most target-domain-similar data first. This difference is critical since early stage updates usually play an important role in the final model performance.

The Effect of Learning Curriculum In order to evaluate the effectiveness of our learning curriculum, we designed experiments to compare the learning curves of five learning curricula. The five learning curricula are described as the following: “Most Similar”: is the curriculum adopted in this paper, in which all the source training samples are ranked based on the target domain probability. A subset of source data is selected and presented to the model learner according to the curriculum that the samples with the highest target domain probability are trained first. “Most Dissimilar”: the curriculum ranks the source data reversely according to the target domain probability, selects and trains the most dissimilar samples first. In “Random Sample” curriculum, a subset of source samples are randomly selected and fed into the training model. In “Random Order within Selected” case, the subset is selected first based on target domain probability. However, the order of presentation during training is random. In “Reverse Order within Selected” scenario, the subset is selected based on target domain probability, and the order of presentation during training is based on the reverse order of target domain probability. As we can see from the figure, both the data selection strategy and learning curriculum have a clear effect on the transfer performance. “Most Similar” curriculum enjoys the highest convergence rate when trained on a small amount of source data, while “Most Dissimilar” curriculum has the lowest convergence rate. The transfer performance of all the learning curricula benefits initially from adding more training data and eventually saturates. Overall, “Most Similar” curriculum converges to the best performance among other curricula. The observation demonstrates the effectiveness of using target domain probability as a measure of learning "hardness".

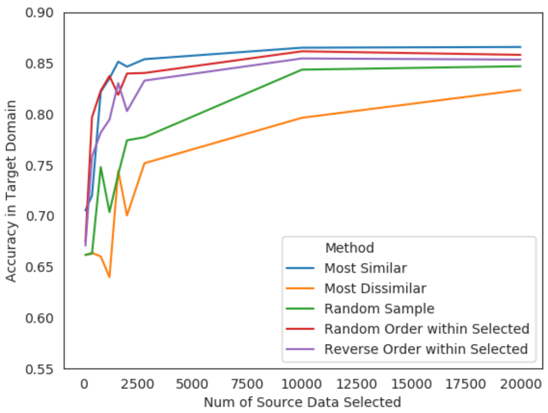

Figure 4: Transfer performance from MNLI to SNLI for five learning curricula. "Most Similar" curriculum achieves the best convergence rate and transfer performance.

5 Conclusion

In conclusion, we propose a novel domain adaptation framework for fine-tuned BERT models through a two-step domain-discriminative data selection and curriculum learning. Our approach significantly outperforms the baseline models on four large datasets, which demonstrates the effectiveness of both the data selection strategy and curriculum design. The method can be readily extended to multi-source domain adaptation, or applied to few-shot learning scenarios in which the selected source domain data can be used to augment the limited target domain training data.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain Adaptation Via Pseudo In-Domain Data Selection. *EMNLP (Empirical Methods in Natural Language Processing)*, pages 355–362.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Minmin Chen, Kilian Q Weinberger, and John Blitzer. 2011. Co-training for domain adaptation. *Advances in neural information processing systems*, pages 2456–2464.

Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. 2016. Best Practices for Fine-Tuning Visual Classifiers to New Domains. *ECCV Workshops*, pages 435–442.

Gabriela Csurka. 2017. A comprehensive survey on domain adaptation for visual applications. In *Advances in Computer Vision and Pattern Recognition*, 9783319583464, pages 1–35.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation. In *ACL-2013: 51st Annual Meeting of the Association for Computational Linguistics*, 1, pages 678–683.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2015. Domain-Adversarial Training of Neural Networks. 17:1–35.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. *Proceedings of the 28th International Conference on Machine Learning*, (1):513–520.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification.

Jing Jiang and Chengxiang Zhai. 2007. Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, (October):264–271.

Ming-yu Liu and Oncel Tuzel. 2016. Coupled Generative Adversarial Networks. (Nips).

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. (ii).

Sebastian Ruder, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017a. Learning what to share between loosely related tasks.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2017b. Knowledge Adaptation: Teaching to Adapt.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian Optimization.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2019. Adversarial Domain Adaptation for Duplicate Question Detection. pages 1056–1063.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2017. Wasserstein Distance Guided Representation Learning for Domain Adaptation.

A. J. Smola, H.-P. Kriegel, M. J. Rasch, B. Scholkopf, K. M. Borgwardt, and A. Gretton. 2006. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):e49–e57.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the Curriculum with Bayesian Optimization for Task-Specific Word Representation Learning. pages 130–139.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *IJCAI International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Daphna Weinshall and Dan Amir. 2018. Theory of Curriculum Learning, with Convex Loss Functions. pages 1–18.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. pages 1112–1122.