Unsupervised Energy-based Adversarial Domain Adaptation for Cross-domain Text Classification

Han Zou[*] Microsoft hanzou@microsoft.com
Jianfei Yang[*] Nanyang Technological University yang0478@ntu.edu.sg
Xiaojian Wu[†] Microsoft xiaojian.academy@gmail.com

Abstract

Transferring knowledge from a label-rich domain (source domain) to a label-scarce domain (target domain) for pervasive cross-domain Text Classification (TC) is a non-trivial task. To overcome this issue, we propose EADA, a novel unsupervised energy-based adversarial domain adaptation framework. First, a deep pre-trained language model (e.g. RoBERTa) is leveraged as a shared feature extractor that maps the text sequences from both source and target domains to a feature space. Since the source features maintain good feature discriminability because of the full supervised training, we design a method that encourages target features towards the source ones via adversarial learning. An autoencoder is designed as an energy function that focuses on reconstructing source feature embeddings, while the feature extractor aims to generate source-like target feature embeddings to deceive the autoencoder. In this manner, the target feature embeddings become domain-invariant and inherit great discriminability. Extensive experiments on multi-domain sentiment classification (Amazon review dataset) and Yes/No question-answering classification (BoolQ and MARCO dataset) are conducted. The experimental results validate that EADA largely alleviates the domain discrepancy while maintaining excellent discriminability and achieves state-of-the-art cross-domain TC performance.

1 Introduction

With the booming development of Natural Language Processing (NLP) in recent years, text classification (TC) is playing a vital role in a myriad of services in our daily lives, such as online recommendations, email spam detection, sentiment classification and social media analysis. Large pre-trained language models, e.g. BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019b), achieve outstanding results on challenging NLP benchmarks, i.e. GLUE (Wang et al., 2018), RACE (Lai et al., 2017), and SQuAD (Rajpurkar et al., 2016). These models enable numerous downstream NLP tasks with compelling performance, including TC, where the model is further fine-tuned with annotated data.

TC tasks are usually domain dependent in real-world. Thus, the performance of these powerful deep models is still fluctuated and even degraded when directly implementing them in a unseen domain (target domain), where the task topic or the data distributions are different from the domain during training (source domain). Although their performance can be improved via fine-tuning with full supervision in the target domain, a significant amount of labeled target data is required. Collecting high-quality data is usually difficult and expensive in many real-world domains. Furthermore, the annotating process is extremely time-consuming and labor-intensive. To overcome these issues, unsupervised domain adaptation (UDA), which aims to transfer the knowledge from a label-rich domain (source domain) to a label-scarce or unlabeled domain (target domain) is proposed (Li et al., 2017; Chen et al., 2018; Guo et al., 2018; Zhang et al., 2019).

The intuitive objective of UDA is to align the marginal distribution of features across source and target domains. In general, UDA methods can be classified into two categories. One line of research focuses on reducing the discrepancy by minimizing statistical measurements, e.g. maximum mean discrepancy (Tzeng et al., 2014a). Another category leverages adversarial learning to alleviate the domain shift. Motivated by Generative Adversarial Network (GAN) (Goodfellow et al., 2014), adversarial domain adaptation (ADA) introduces a binary domain discriminator to identify the domain

[*]Equal Contribution.
[†]Work done while at Microsoft.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1208–1218
August 1–6, 2021.

label of the data, while an encoder learns to fool the discriminator. ADA has achieved encouraging results on nontrivial DA problems across various applications, such as image classification (Vu et al., 2019; Yang et al., 2020b), human activity recognition (Zou et al., 2019; Yang et al., 2018; Zou et al., 2018), Internet of Things (Yang et al., 2020a), and also text classification (Li et al., 2017; Zhang et al., 2019). For instance, AMN (Li et al., 2017) trains a sentiment classifier and a domain discriminator to reduce the domain discrepancy. ADAN (Chen et al., 2018) exploits adversarial learning for cross-lingual sentiment classification. HAGAN (Zhang et al., 2019) integrates the hierarchical attention mechanism with ADA to obtain features that are sentiment distinguishable but domain indistinguishable.

Although these ADA methods achieve good results in certain cross-domain TC tasks, one major issue is the unstable prediction performance in the target domain (Xie et al., 2018; Saito et al., 2018). After the adversarial training achieves convergence, the conventional binary domain discriminator cannot distinguish the domain label of the feature representations, which means these representations obtain good transferability. However, there is no constraint on the discriminability in the target domain. The model can generate trivial but useless target feature representations as long as they can fool the domain discriminator. Thus, this uncertainty in adversarial training deteriorates the discriminability of the target feature representations and ignores the decision boundary learned in the source domain, which leads to unstable and even poor prediction performance in the target domain (Chen et al., 2019a; Cui et al., 2020). Some works aim to adjust the decision boundary of the label classifier (Saito et al., 2018; Shu et al., 2018) or align additional semantic information (Xie et al., 2018) to overcome this issue during adversarial training. However, these additional learning steps either require a sophisticated hyper-parameter tuning process or increase the computational overhead, that limits the generalization capability of the ADA methods for NLP tasks. Therefore, a simple yet efficient solution is urgently desired.

In this paper, we propose EADA, an energy-based adversarial domain adaptation framework that tackles the uncertainty issue during adversarial learning and dedicates for text classification tasks. EADA consists of three modules, a shared feature extractor, a label predictor, and an autoencoder. We employ a deep pre-trained language model (RoBERTa) as a shared feature extractor that maps the text sequences from both source domain and target domain into a latent feature space. With the labeled source data, the feature extractor and the label predictor are fine-tuned under full supervision. Since the source feature representations generated from the feature extractor contain superb discriminability, the innovative goal of EADA is to fix these source features by adding constrains in the objective and only force the target feature distribution to align the source feature distribution through adversarial training so that the target features could remain discriminative, and the label predictor could also perform well in the target domain. Since autoencoder is acknowledged as an energy function that learns to map the observed sample to the low-energy space (LeCun et al., 2006), we design an autoencoder that leverages this property to fix the source features by associating lower energies to it while pushing the target domain to the low-energy space by minimizing the margin loss of the autoencoder. Meanwhile, it can also cluster similar data to form a high-density manifold, which helps to preserve more semantic information. We train the autoencoder to reconstruct the source features and train the feature extractor to generate source-like target features to deceive the autoencoder via a minimax with a margin loss. In summary, we make the following contributions:

- To address the problem of conventional binary domain discriminator that deteriorates the discriminability of the target feature presentation, we propose a novel autoencoder module, which forces the target feature representations to simulate source feature representations such that good discriminability can be inherited.

- As an energy function, the autoencoder maps features from both domains to the low-energy space, which motivates the feature clusters to be tight in an unsupervised manner. It improves the label classification accuracy in the target domain.

- Extensive experiments on public cross-domain TC benchmark datasets, including multi-domain sentiment classification (Amazon review dataset) and cross-domain

Yes/No question-answering (QA) classification (BoolQ and MARCO dataset), are conducted. The experimental results demonstrate that EADA alleviates the uncertainty during adversarial training and enhances the feature discriminability in the target domain. This enables EADA to outperform existing methods and achieve new state-of-the-art ADA results for cross-domain TC tasks without requiring any labeled data in the target domain.

The rest of the paper is organized as follows. Section 2 summarizes the existing domain adaptation methods for TC tasks. The limitation of existing ADA Methods is elaborated in Section 3. Section 4 presents the framework architecture of EADA. In Section 5, we present the experimental results and performance evaluation. We conclude our work in Section 6.

2 Related Work

Domain Adaptation aims to tackle the domain shift issue when the data distribution in the source domain and target domain are different (Ben-David et al., 2010). Unsupervised domain adaptation (UDA) aims to learn a model that is able achieve good classification accuracy without any annotation in the target domain (Tzeng et al., 2017; Zhao et al., 2019). Certain statistical measurements, such as maximum mean discrepancy (MMD) (Tzeng et al., 2014b; Ma et al., 2019), are leveraged to quantify the distribution differences.

Inspired by the recent success of Generative Adversarial Network (GAN) (Goodfellow et al., 2014) for data generation, researchers have proposed adversarial domain adaptation (ADA), that constructs an adversarial loss to accommodate the domain shift. It consists of an encoder and a domain discriminator. The generator aims to fool the discriminator to make the target domain samples look like the source domain ones, while the discriminator tries to identify the domain labels (source or target). ADDA (Tzeng et al., 2017) learns a discriminative representation using the labels in the source domain and then a separate encoder that maps the target data to the same space using an asymmetric mapping learned through a standard GAN loss without weights sharing. CoGAN (Liu and Tuzel, 2016) trains 2 GANs to synthesize both source and target images and achieves a domain invariant feature space by tying the high-level layer parameters of the 2 GAN to solve the domain transfer problem.

ADA has been adopted for cross-domain NLP tasks as well (Peng et al., 2018; Li et al., 2017; Shah et al., 2018; Chen and Cardie, 2018; Cai and Wan, 2019; Wang et al., 2019). AMN (Li et al., 2017) is an end-to-end adversarial memory network for cross-domain sentiment classification, which is the pioneering work for ADA in NLP. An adversarial deep averaging network is proposed in (Chen et al., 2018) for cross-lingual sentiment classification. A dedicated ADA framework for machine reading comprehension is proposed in (Wang et al., 2019). (Chen and Cardie, 2018) designed an ADA model that learns domain invariant representation across multiple domains for text classification. Target domain-specific information is being exploited in (Peng et al., 2018) to further improve the DA performance, while labeled data in the target domain is required.

Large deep pre-trained language models pioneered by BERT (Devlin et al., 2019), have been employed as feature encoders to embed text sequences into a latent feature space. Then, the encoder is further fine-tuned with the discriminator via adversarial learning using the labeled source data and unlabeled target data (Lee et al., 2019; Ma et al., 2019). For instance, BERT and ADA were adopted in (Lee et al., 2019) for domain-agnostic question-answering. A similar framework that integrates BERT and MMD is proposed in (Ma et al., 2019) for cross-domain sentiment classification. However, all these approaches leverage the binary domain discriminator which has failed to consider the discriminative features during feature learning. This leads to severe performance degradation since the decision boundary of the label predictor trained with source data is no longer valid in the target domain due to the domain shift.

3 Limitation of Existing ADA Methods

In this section, we analyze the learning process of conventional ADA methods and reveal their limitations. In common UDA setup, N_s labeled examples from a source domain $\mathcal{D}_S = \{\mathbf{x}_i^s, y_i^s\}$ and N_t unlabeled examples from a target domain $\mathcal{D}_T = \{x_i^t\}$ are available. The distributions of $\mathcal{D}_S$ and $\mathcal{D}_T$ are different due to the domain discrepancy. UDA aims to build up a model that provides good class prediction in both source and target domain. Discriminability of the feature representation is the clustering capacity in the feature manifold, that controls the easiness of class category separation

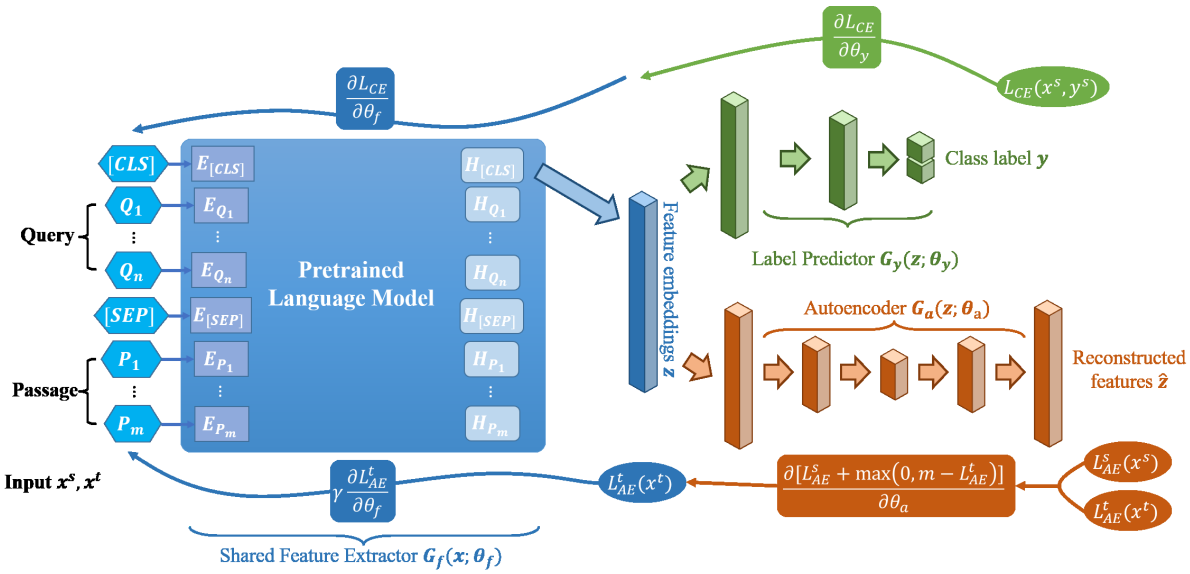

Figure 1: EADA constitutes: a pre-trained language model as shared feature extractor G_f, a label predictor G_y and an autoencoder G_a. In addition to the full supervised learning of G_f and G_y with the labeled source data, the autoencoder G_a serves as a domain classifier to learn reconstructing the source feature representations and push the target feature representations away. The feature extractor G_f aims to generate source-like target feature representations to deceive the autoencoder. This objective is realized by forcing the target feature representations towards the those from the source domain in the feature space via adversarial learning.

(Chen et al., 2019b). Excellent discriminability can be achieved for the source features due to the full supervision learning in the source domain. The objective of UDA is to transfer and ensure the model maintains this discriminability in the target domain.

ADA method, as one category of UDA, which is pioneered by Domain-Adversarial Training of Neural Networks (DANN) (Ganin et al., 2016) and Adversarial Memory Network (AMN) (Li et al., 2017) have shown promising performance in numerous NLP tasks in recent years (Chen and Cardie, 2018; Shah et al., 2018; Cai and Wan, 2019; Wang et al., 2019). It usually consists of a shared feature extractor $\mathbf{f} = G_f(x)$, a label predictor $\mathbf{y} = G_y(x)$ and a domain discriminator $\mathbf{d} = G_d(x)$. In addition to the standard full supervision learning process in the source domain, a minimax game is designed between $\mathbf{f}$ and $\mathbf{d}$. The domain discriminator $\mathbf{d}$ aims to distinguish the domain label between source and target, meanwhile the feature extractor $\mathbf{f}$ is trained to deceive $\mathbf{d}$. This adversarial training process can be formulated as

$$\min_{G_f, G_y} \mathcal{L}_y(\mathbf{X}_s, Y_s) - \gamma \mathcal{L}_f(\mathbf{X}_s, \mathbf{X}_t), \quad (1)$$

$$\min_{G_d} \mathcal{L}_d(\mathbf{X}_s, \mathbf{X}_t), \quad (2)$$

where $\mathcal{L}_y$ is the cross-entropy classification loss. In this manner, the model can learn domain-invariant features and transfer them across domains when the Nash Equilibrium is achieved (Zhao et al., 2017). The hyper-parameter γ controls the significance of adversarial training that improves transferability. As shown in Eq(1), the training process of feature extractor $\mathbf{f}$ of conventional ADA methods aims to achieve two tasks: (1) learn source representations with good discriminability; (2) train representations that are indistinguishable to the domain discriminator $\mathbf{d}$. Since both source domain and target domain data are involved in the adversarial feature learning as presented in the second term of Eq(1), the objective is equivalent to move two domains closer in the feature space to deceive $\mathbf{d}$. However, this process does not impose any constraint on the discriminability in the target domain. The feature extractor $\mathbf{f}$ can generate trivial but useless target representations as long as they can fool the discriminator $\mathbf{d}$. Therefore, these ADA methods cannot guarantee that the good decision boundary learned via full supervision in the source domain can still separate the categorical clusters in the target domain (Chen et al., 2019b; Liu et al., 2019a). This degradation of discriminability in the target domain is the major reason that hinders the performance of existing ADA methods.

4 Energy-based Adversarial Domain Adaptation

It is not a trivial task to maintain the source manifolds during adversarial training. Our solution is to decouple the adversarial training process of source and target feature representations. To be specific, we fix the source representation in the feature space and only encourage the target representations align to the source representations. Therefore, the superb discriminability learned in the source domain can be preserved and a label predictor that performs well in both source and target domain can be obtained.

To achieve this goal, we propose Energy-based Adversarial Domain Adaptation (EADA), which innovatively utilizes an autoencoder structure as a domain discriminator during adversarial training. Figure 1 demonstrates the model structure of EADA. It consists of three modules, a pre-trained language model as a shared feature extractor G_f parameterized by θ_f to embed input sample to feature embedding z. After that, a label predictor G_y parameterized by θ_y, which consists of several fully connected layers, further maps the feature embedding z to the predicted label $\hat{y}$. Another module is an autoencoder G_a parameterized by θ_a, that reconstructs a feature embedding z to $\hat{z}$. The detailed functionality of each module is elaborated as follows.

4.1 Shared Feature Extractor

Large pre-trained language models (e.g. BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019b)) have achieved a series of state-of-the-art results on NLP benchmarks. These powerful pre-trained language models are built up on bidirectional transformer architecture, and pre-trained on large corpora with a masked language model, that enable various downstream NLP tasks, including text classification.

In this work, we employ RoBERTa as the shared feature extractor G_f (highlighted in blue in Figure 1) that embeds both labeled text data ($\mathbf{X}_s$) from the source domain, as well as the unlabeled text data ($\mathbf{X}_t$) from the target domain into a latent feature space. To be specific, as a QnA classification problem, the input for the G_f are sequence pairs <query $\mathbf{Q}$, passage $\mathbf{P}$> as depicted in Figure 1 in the format of [CLS] <s> Q </s> <s> P </s>, where [CLS] is a dummy token for classification and <s> </s> are separator tokens. We leverage the roberta.base architecture (12-layer, 768-hidden, 12-heads, 125M parameters) (Liu et al., 2019b) as the shared feature extractor G_f. Since our objective is text classification, the last hidden representation of the [CLS] token, $H_{[CLS]} \in \mathbf{R}^{768\times 1}$ (feature embedding z) serves as the output of G_f. These embeddings $\mathbf{z}$ are utilized by both classifier G_y and autoencoder G_a.

4.2 Class Label Predictor

The class label predictor G_y consists of several fully connected layers that map the feature embedding z to the predicted label $\hat{y}$. Since the source domain is label-rich by default, we assume that n labeled samples $\mathcal{D}_S = \{\mathbf{x}_i^s, y_i^s\}$ are available from the source domain for finetuning of the shared feature extractor (language model) G_f (blue part in Figure 1) and the label predictor G_y (green part in Figure 1). The good classification accuracy of G_y is achieved by minimizing the cross-entropy loss via back-propagation under full supervision:

$$\min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) = - \mathbb{E}_{(\mathbf{x}_s, y_s)\sim(\mathbf{X}_s, Y_s)} \sum_{n=1}^{N_s} [\mathbb{I}_{[l=y_s]} \log G_y(G_f(\mathbf{x}_s))]. \tag{3}$$

4.3 Autoencoder as Domain Discriminator

After obtaining the source feature representation with good discriminability, the next task is to learn transferable features with k unlabeled samples from a target domain $\mathcal{D}_T = \{x_i^t\}$. To ensure both transferability and discriminability of the feature representation, we design an autoencoder G_a with a margin Mean Squared Error (MSE) loss to replace the conventional binary domain discriminator. The MSE loss of the autoencoder is defined as:

$$\mathcal{L}_{AE}(\mathbf{x}_i) = ||G_a(G_f(\mathbf{x}; \theta_f); \theta_a) - \mathbf{x}_i||_2^2, \tag{4}$$

where $||\cdot||_2^2$ denotes the squared L_2-norm. Since the source embeddings z_s always contain superb discriminability due to full supervision during the training of the classifier, z_s should be fixed to preserve the good decision boundary, while the target embeddings z_t should be encouraged to align with the distribution of z_s. To achieve this goal, the autoencoder G_a is designed to be able to only reconstruct features in the source domain but not features in the target domain. Namely, when two

domains distribute similarly, the autoencoder will incur the same reconstruction loss in both domains. The training process of the autoencoder is formulated as:

$$\min_{G_a} \mathcal{L}_{AE}(\mathbf{X_s}) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X_t})), \quad (5)$$

where m is the margin between the representations from the source domain and the target domain. The autoencoder G_a can be considered as an energy function that associates lower energies to the observed samples in a binary classification problem (LeCun et al., 2006). With the inspiration of Energy-based GAN which theoretically proves that using an energy function in GAN, the true distribution can be simulated by the generator at *Nash Equilibrium* (Zhao et al., 2017). In EADA, the autoencoder module G_a provides similar functionality that associates low energies to the source features (focuses on reconstructing source embeddings z_s). As presented in Eq(5), the training goal of the autoencoder is to have $\mathcal{L}_{AE}(\mathbf{X_s}) = 0$ and $\mathcal{L}_{AE}(\mathbf{X_t}) = m$. It behaves proportionally similar to a binary domain discriminator. But G_a includes more domain information and can transfer it during adversarial training.

4.4 The Learning Framework

The adversarial training objective of three modules forms a minimax game, that is defined by:

$$\begin{aligned} &\min_{G_f, G_y} \mathcal{L}_{CE}(\mathbf{X}_s, Y_s) + \gamma \mathcal{L}_{AE}(\mathbf{X_t}), \\ &\min_{G_a} \mathcal{L}_{AE}(\mathbf{X_s}) + \max(0, m - \mathcal{L}_{AE}(\mathbf{X_t})), \end{aligned} \quad (6)$$

where γ is a hyper-parameter to control the effectiveness of G_a. The shared feature extractor G_f maps both labeled source data $\mathbf{X_s}$ and unlabeled target data $\mathbf{X_t}$ to a latent feature space. Both G_f and the label predictor G_y are trained with full supervision using the labeled data in the source domain. Another key role of the feature extractor G_f is to deceive the autoencoder G_a by generating source-like features for unlabeled target samples. Therefore, we only incorporate the $\mathcal{L}_{AE}(\mathbf{X_t})$ term into the training of G_f. The adversarial training of G_f is formulated by:

$$\min_{G_f} \mathcal{L}_{AE}(\mathbf{X_t}). \quad (7)$$

In the minimax game, the autoencoder G_a aims to maximize the domain divergence by pushing two domains away from a margin m, while the objective of the feature extractor G_f is to minimize the domain divergence by deceiving the autoencoder. When the model achieves convergence, the target feature representations inherit excellent discriminability from the source domain so that the generalization capability of the label predictor G_y is improved and performs well not only in the source domain but also in the target domain.

5 Experiments

We evaluate the domain adaptation performance of EADA on two public real-world cross-domain text classification benchmarks: 1) sentiment classification (Amazon reviews dataset); 2) Natural QA Yes/No classification (BoolQ $\Leftrightarrow$ MS Marco), and compared it with state-of-the-art baselines.

5.1 Evaluation on Sentiment Classification

Amazon reviews dataset (Pan et al., 2010) is the standard and well-known benchmark for sentiment classification domain adaptation. It contains reviews on four domains: Books (B), DVDs (D), Electronics (E), and Kitchen (K). Each domain contains 1000 positive reviews (higher than 3 stars) and 1000 negative reviews (3 stars or lower). 12 cross-domain sentiment classification tasks: D→B, E→B, K→B, K→E, D→E, B→E, B→D, K→D, E→D, B→K, D→K, E→K, where the letter before the arrow represents the source domain and the letter after the arrow indicates the target domain by following (Li et al., 2017). For each pair of domain adaptation, 800 labeled positive (Pos) and 800 labeled negative (Neg) reviews from the source domain (src), together with 1600 unlabeled reviews from the target domain (tgt) are randomly selected for training. The rest of 200 positive and 200 negative reviews from the target domain are used for testing.

We configured the feature extractor module G_f as RoBERTa.base for single sequences task since each review is one sequence passage. The input is tokenized as [CLS] <s> Review </s>. The maximum input sequence length is set to 256 tokens. The autoencoder module G_a consists of 5 fully connected layers (768-384-96-384-768). The entire EADA framework is implemented in PyTorch. The Adam optimizer with the constant learning rate $\mu = 1e^{-5}$ with a batch size of 24 was adopted and we used 5-fold cross-validation to tune the hyperparameter $m = 4$ and $\gamma = 1e^{-2}$ during the training.

Table 1: Cross-domain classification accuracy of different methods on Amazon review dataset.

Tasks	Source-only RoBERTa	AMN	ADAN	HAGAN	HAGAN-C	MoE	ADA RoBERTa	EADA
D→B	85.55	81.52	81.70	81.22	81.69	81.70	86.80	**88.10**
E→B	81.95	77.80	78.55	79.05	79.23	79.65	82.10	**85.25**
K→B	72.65	79.37	79.25	78.52	78.99	80.94	77.60	**81.20**
B→D	81.95	81.32	82.30	82.07	82.38	84.60	82.47	**86.05**
E→D	82.50	77.51	79.70	81.00	80.65	84.20	83.70	**85.35**
K→D	80.00	80.03	80.45	80.83	80.91	**81.22**	80.58	80.35
B→E	85.95	80.07	77.60	79.87	80.12	84.66	86.00	**89.35**
D→E	82.75	80.00	79.70	80.57	80.99	84.10	82.99	**87.15**
K→E	82.45	81.97	86.85	85.94	85.23	**85.81**	84.32	85.11
B→K	86.80	81.00	76.10	81.25	82.00	86.32	87.00	**89.65**
D→K	85.75	83.88	77.35	81.73	81.50	86.33	86.55	**89.20**
E→K	86.95	87.10	83.95	84.30	84.99	85.40	87.10	**90.50**
Average	*82.94*	*80.96*	*80.29*	*81.36*	*81.56*	*83.74*	83.93	***86.44***

The performance of EADA is compared with the following state-of-the-art baselines: *Source-only RoBERTa* directly applies the finetuned RoBERTa model from source domain in target domain; *AMN* (Li et al., 2017) adopts attention mechanism and ADA for cross-domain sentiment classification; *ADAN*(Chen et al., 2018) employs ADA to transfer the knowledge of resource-rich source language to low-resource language; *HAGAN* and *HAGAN-C* (Zhang et al., 2019) incorporate hierarchical attention into ADA for cross-domain sentiment classification; *MoE* (Guo et al., 2018) utilizes a mixture of experts from different source domains to further enhance the DA performance in target domain; *ADA RoBERTa* uses RoBERTa as a shared feature extractor and a conventional binary domain discriminator as the domain classifier for ADA.

Table 1 presents the mean accuracy with 5 runs of each method on the 12 DA tasks. One observation is that the accuracy of source-only RoBERTa (82.94%) is even worse than MoE (83.74%). This indicates that the problem of DA cannot be solved by just solely using large pre-trained language models. It can be observed that EADA provides 86.44% classification accuracy on average, which outperforms all the baselines. It achieves the best DA performance in 10 out of the 12 tasks. Although ADA-RoBERTa and EADA both adopted RoBERTa as the feature extractor, the accuracy of EADA is still 2.5% higher than ADA-RoBERTa, which validates the advantage of the proposed energy-based ADA method. Moreover, the variance of EADA is the smallest among all the methods, which indicates its performance is more stable in general. By learning a source-like representation for the target feature embeddings, EADA successfully performs cross-domain sentiment classification without any annotated data in target domain.

5.2 Evaluation on Yes/No QA classification

We also validated the performance of EADA on cross-domain naturally occurring yes/no questions between BoolQ dataset (Clark et al., 2019) and Marco dataset (Nguyen et al., 2016). Each example is a triplet of (query, passage, answer). Thus, feature extractor module G_f (RoBERTa) is configured for sequence pairs task, which means the input is tokenized in the format as [CLS] <s> query </s> <s> passage </s>. BoolQ dataset contains 5874 Yes and 3553 No samples for training and 2033 Yes and 1237 No samples for evaluation from Wikipedia. Samples in Marco dataset are web snippets from Bing Search. There are 17339 Yes and 10550 No samples for training and 2033 Yes and 1237 No samples for testing. The data distributions at both domain level and categorical level are imbalanced, which is a common situation in many real-world applications. The number of training samples in BoolQ is only 33.8% of those in Marco, indicating that the data are imbalanced across domains. Moreover, the data categorical distribution is imbalanced because the number of No-samples is at least 39.1% less than the number of Yes-samples in both domains. Since the samples from the two domains are collected from different sources, there is a huge domain shift between the two datasets.

The domain adaptation performance is reported in Table 2. The 2^{nd} column shows the accuracies

Table 2: Cross-domain adaptation for Yes/No QA classification between BoolQ and Marco datasets.

Tasks	Source only	AMN	HAGAN	ADA-RoBERTa	MoE	EADA	Target full supervision
BoolQ → Marco	67.63	72.95	73.11	73.86	74.01	**78.38**	82.51
Marco → BoolQ	69.30	74.34	74.73	75.11	75.23	**79.51**	84.31

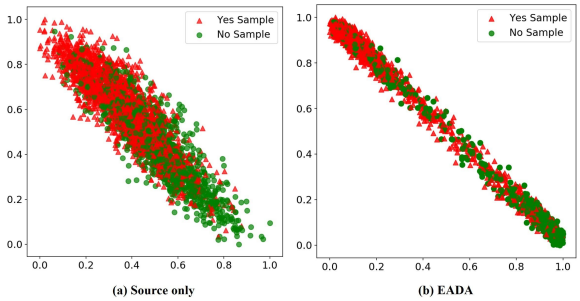

Figure 2: The t-SNE visualization of features embedded using distinct feature extractor in target domain. (Marco → BoolQ).

when the non-adapted source feature extractors and classifiers are directly applied in the target domain, which serves as the lower-baseline. The last column reports the accuracies when the feature extractors and classifiers are trained with full-supervision that all the target training data are labeled (as the upper-baseline). As shown in Table 2, the source-only classifiers can only provide 68% accuracy, which verifies that the domain shift hurts the classification accuracy even when a powerful deep language model is adopted. On the other hand, it can be easily observed from Table 2 that EADA enhances the accuracy in both adaptation directions by at least 15% compared to the lower-baseline in an unsupervised manner, and outperforms all the state-of-the-art baselines. It elevates the performance closer to the target full supervision as well.

We leverage t-Distributed Stochastic Neighbor Embedding (t-SNE) to map the embedded feature representations through different feature extractors to a 2-D space for better visualization and analysis. Figure 2(a) and Figure 2(b) depict the embedded features using the non-adapted source feature extractor and the EADA's feature extractor learned, respectively (Yes sample - red, No sample - Green). If we directly apply the non-adapted source feature extractor in the target domain, as shown in Figure 2(a), a large amount samples with different categorical labels overlap with each other, which leads to corresponding huge misclassification as presented in the 2^{nd} column of Table 2. After employing EADA, the common confusions are further separated in the latent feature space and two clusters are formulated as depicted in Figure 2(b). These observations further validate that the feature embeddings constructed via EADA are not only domain-invariant but also preserve excellent discriminability in both the source domain and the target domain.

We also conducted a sensitivity study of the two hyperparameters: m and γ in Eq(6). We evaluated the impact of margin m with different values from 0 to 10 in both experiments. EADA's accuracy increases while m is increasing. The reason is that the degree of transferability is limited when m is small. The performance becomes stable when $m \geqslant 4$ in both experiments. In general, the objective of γ is to control the weight of the adversarial loss during feature learning of ADA as shown in Eq(1). γ in EADA aims to control the weight of autoencoder reconstruction loss for target domain samples during feature learning as presented in Eq(6)). We evaluated the impact of γ with different values (0-1) for EADA and ADA-RoBERTa. As γ increases, the accuracy of EADA increases and becomes stable when $\gamma \geqslant 0.01$. The accuracy of ADA-RoBERTa fluctuates when γ is increased and decreased when $\gamma \geqslant 0.5$. Thus, EADA provides a more stable training procedure compared to conventional ADA methods, which makes it easier for generalization. We recommend using m=4 and γ=0.01 as the default setup for other tasks.

6 Conclusion

In this paper, we proposed EADA, a novel unsupervised energy-based adversarial domain adaptation method for cross-domain text classification tasks. First, a deep pre-trained language model is leveraged as a shared feature extractor to map the text sequences from both source and target domains to a feature space. The feature extractor and a label predictor are trained with labeled source data. Since the source feature representations are obtained under full supervision, they preserve great feature discriminability. To ensure that the label predictor also provides good label prediction in the target domain, the target feature representations should be

encouraged to align with the source during adversarial training. Thus, we designed an autoencoder that focuses on reconstructing the source feature representations, while the feature extractor aims to generate source-like target feature embeddings to fool the autoencoder. Extensive experiments on public cross-domain TC benchmarks are conducted and demonstrate that EADA not only alleviates the domain discrepancy but also enhances the feature discriminability in the target domain, which leads to compelling cross-domain TC performance without requiring any labeled data in the target domain.

References

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*, 79(1-2):151–175.

Yitao Cai and Xiaojun Wan. 2019. Multi-domain sentiment classification based on domain-aware embedding and attention. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4904–4910. AAAI Press.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. 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 1226–1240.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. 2019a. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. 2019b. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1081–1090, Long Beach, California, USA. PMLR.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Seanie Lee, Donggyu Kim, and Jangwon Park. 2019. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 196–202.

Zheng Li, Yun Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *IJCAI*, pages 2237–2243.

Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. 2019a. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022.

Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

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

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuan-Jing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063.

Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. 2018. A dirt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014a. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014b. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Huazheng Wang, Zhe Gan, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Hongning Wang. 2019. Adversarial domain adaptation for machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2510–2520.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5423–5432, Stockholmsmässan, Stockholm Sweden. PMLR.

Jianfei Yang, Han Zou, Shuxin Cao, Zhenghua Chen, and Lihua Xie. 2020a. Mobileda: Toward edge-domain adaptation. *IEEE Internet of Things Journal*, 7(8):6909–6918.

Jianfei Yang, Han Zou, Hao Jiang, and Lihua Xie. 2018. Carefi: Sedentary behavior monitoring system via commodity wifi infrastructures. *IEEE Transactions on Vehicular Technology*, 67(8):7620–7629.

Jianfei Yang, Han Zou, Yuxun Zhou, Zhaoyang Zeng, and Lihua Xie. 2020b. Mind the discriminability: Asymmetric adversarial domain adaptation. In *European Conference on Computer Vision*, pages 589–606. Springer.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yuebing Zhang, Duoqian Miao, and Jiaqi Wang. 2019. Hierarchical attention generative adversarial networks for cross-domain sentiment classification. *arXiv preprint arXiv:1903.11334*.

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

Junbo Zhao, Michael Mathieu, and Yann LeCun. 2017. Energy-based generative adversarial network. In *Proc. 5th International Conference on Learning Representations*.

Han Zou, Jianfei Yang, Yuxun Zhou, Lihua Xie, and Costas J Spanos. 2018. Robust wifi-enabled device-free gesture recognition via unsupervised adversarial domain adaptation. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–8. IEEE.

Han Zou, Yuxun Zhou, Jianfei Yang, Huihan Liu, Hari Prasanna Das, and Costas J Spanos. 2019. Consensus adversarial domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5997–6004.