The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Wasserstein Distance Guided Representation Learning for Domain Adaptation

Jian Shen, Yanru Qu, Weinan Zhang,* Yong Yu
Shanghai Jiao Tong University
{rockyshen, kevinqu, wnzhang, yyu}@apex.sjtu.edu.cn

Abstract

Domain adaptation aims at generalizing a high-performance learner on a target domain via utilizing the knowledge distilled from a source domain which has a different but related data distribution. One solution to domain adaptation is to learn domain invariant feature representations while the learned representations should also be discriminative in prediction. To learn such representations, domain adaptation frameworks usually include a domain invariant representation learning approach to measure and reduce the domain discrepancy, as well as a discriminator for classification. Inspired by Wasserstein GAN, in this paper we propose a novel approach to learn domain invariant feature representations, namely Wasserstein Distance Guided Representation Learning (WDGRL). WDGRL utilizes a neural network, denoted by the domain critic, to estimate empirical Wasserstein distance between the source and target samples and optimizes the feature extractor network to minimize the estimated Wasserstein distance in an adversarial manner. The theoretical advantages of Wasserstein distance for domain adaptation lie in its gradient property and promising generalization bound. Empirical studies on common sentiment and image classification adaptation datasets demonstrate that our proposed WDGRL outperforms the state-of-the-art domain invariant representation learning approaches.

Introduction

Domain adaptation defines the problem when the target domain labeled data is insufficient, while the source domain has much more labeled data. Even though the source and target domains have different marginal distributions (Ben-David et al. 2007; Pan and Yang 2010), domain adaptation aims at utilizing the knowledge distilled from the source domain to help target domain learning. In practice, unsupervised domain adaptation is concerned and studied more commonly since manual annotation is often expensive or time-consuming. Faced with the covariate shift and the lack of annotations, conventional machine learning methods may fail to learn a high-performance model.

To effectively transfer a classifier across different domains, different methods have been proposed, including instance reweighting (Mansour, Mohri, and Rostamizadeh 2009), subsampling (Chen, Chen, and Weinberger 2011), feature mapping (Tzeng et al. 2014) and weight regularization (Rozantsev, Salzmann, and Fua 2016). Among these methods feature mapping has shown great success recently, which projects the data from different domains to a common latent space where the feature representations are domain invariant. Recently, deep neural networks, as a great tool to automatically learn effective data representations, have been leveraged in learning knowledge-transferable feature representations for domain adaptation (Glorot, Bordes, and Bengio 2011; Chen et al. 2012; Zhuang et al. 2015; Long et al. 2015; Ganin et al. 2016).

On the other hand, generative adversarial nets (GANs) (Goodfellow et al. 2014) are heavily studied during recent years, which play a minimax game between two adversarial networks: the discriminator is trained to distinguish real data from the generated data, while the generator learns to generate high-quality data to fool the discriminator. It is intuitive to employ this minimax game for domain adaptation to make the source and target feature representations indistinguishable. These adversarial adaptation methods have become a popular solution to reduce domain discrepancy through an adversarial objective with respect to a domain classifier (Ganin et al. 2016; Tzeng et al. 2017). However, when the domain classifier network can perfectly distinguish target representations from source ones, there will be a gradient vanishing problem. A more reasonable solution would be to replace the domain discrepancy measure with Wasserstein distance, which provides more stable gradients even if two distributions are distant (Arjovsky, Chintala, and Bottou 2017).

In this paper, we propose a domain invariant representation learning approach to reduce domain discrepancy for domain adaptation, namely Wasserstein Distance Guided Representation Learning (WDGRL), inspired by recently proposed Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017). WDGRL trains a domain critic network to estimate the empirical Wasserstein distance between the source and target feature representations. The feature extractor network will then be optimized to minimize the estimated Wasserstein distance in an adversarial manner. By iterative adversarial training, we finally learn feature representations invariant to the covariate shift between domains. Additionally, WDGRL can be easily adopted in existing domain adap-

*Weinan Zhang is the corresponding author.

tation frameworks (Tzeng et al. 2014; Long et al. 2015; Zhuang et al. 2015; Long et al. 2016; Bousmalis et al. 2016) by replacing the representation learning approaches. Empirical studies on common domain adaptation benchmarks demonstrate that WDGRL outperforms the state-of-the-art representation learning approaches for domain adaptation. Furthermore, the visualization of learned representations clearly shows that WDGRL successfully unifies two domain distributions, as well as maintains obvious label discrimination.

Related Works

Domain adaptation is a popular subject in transfer learning (Pan and Yang 2010). It concerns covariate shift between two data distributions, usually labeled source data and unlabeled target data. Solutions to domain adaptation problems can be mainly categorized into three types: i). Instance-based methods, which reweight/subsample the source samples to match the distribution of the target domain, thus training on the reweighted source samples guarantees classifiers with transferability (Huang et al. 2007; Chen, Weinberger, and Blitzer 2011; Chu, De la Torre, and Cohn 2013). ii). Parameter-based methods, which transfer knowledge through shared or regularized parameters of source and target domain learners, or by combining multiple reweighted source learners to form an improved target learner (Duan, Xu, and Chang 2012; Rozantsev, Salzmann, and Fua 2016). iii). The last but the most popular and effective methods are feature-based, which can be further categorized into two groups (Weiss, Khoshgoftaar, and Wang 2016). Asymmetric feature-based methods transform the features of one domain to more closely match another domain (Hoffman et al. 2014; Kandemir 2015; Courty et al. 2017) while symmetric feature-based methods map different domains to a common latent space where the feature distributions are close.

Recently, deep learning has been regarded as a powerful way to learn feature representations for domain adaptation. Symmetric feature-based methods are more widely studied since it can be easily incorporated into deep neural networks (Chen et al. 2012; Zhuang et al. 2015; Long et al. 2015; Ganin et al. 2016; Bousmalis et al. 2016; Luo et al. 2017). Among symmetric feature-based methods, minimizing the maximum mean discrepancy (MMD) (Gretton et al. 2012) metric is effective to minimize the divergence of two distributions. MMD is a nonparametric metric that measures the distribution divergence between the mean embeddings of two distributions in reproducing kernel Hilbert space (RKHS). The deep domain confusion (DDC) method (Tzeng et al. 2014) utilized MMD metric in the last fully connected layer in addition to the regular classification loss to learn representations that are both domain invariant and discriminative. Deep adaptation network (DAN) (Long et al. 2015) was proposed to enhance the feature transferability by minimizing multi-kernel MMD in several task-specific layers. On the other hand, correlation alignment (CORAL) method (Sun, Feng, and Saenko 2016) was proposed to align the second-order statistics of the source and target distributions with a linear transformation and (Sun and Saenko 2016) extended CORAL and proposed Deep CORAL to learn a nonlinear transformation that aligns correlations of layer activations in deep neural networks.

Another class of symmetric feature-based methods uses an adversarial objective to reduce domain discrepancy. Motivated by theory in (Ben-David et al. 2007; 2010) suggesting that a good cross-domain representation contains no discriminative information about the origin (i.e. domain) of the input, domain adversarial neural network (DANN) (Ajakan et al. 2014; Ganin et al. 2016) was proposed to learn domain invariant features by a minimax game between the domain classifier and the feature extractor. In order to back-propagate the gradients computed from the domain classifier, DANN employs a gradient reversal layer (GRL). On the other hand, (Tzeng et al. 2017) proposed a general framework for adversarial adaptation by choosing adversarial loss type with respect to the domain classifier and the weight sharing strategy. Our proposed WDGRL can also be viewed as an adversarial adaptation method since it evaluates and minimizes the empirical Wasserstein distance in an adversarial manner. Our WDGRL differs from previous adversarial methods: i). WDGRL adopts an iterative adversarial training strategy, ii). WDGRL adopts Wasserstein distance as the adversarial loss which has gradient superiority.

Another related work for domain adaptation is optimal transport (Courty, Flamary, and Tuia 2014; Courty et al. 2017), which is equivalent to Wasserstein distance. And (Redko, Habrard, and Sebban 2016) gave a theoretical analysis that Wasserstein distance can guarantee generalization for domain adaptation. Though these works utilized Wasserstein distance in domain adaptation, there are distinct differences between WDGRL and the previous ones: these works are asymmetric feature-based methods which design a transformation from source representations to target ones based on optimal transport while WDGRL is a symmetric method that projects both domains to a common latent space to learn domain invariant features. And WDGRL can be integrated into other symmetric feature-based adaptation frameworks.

Besides learning shared representations, domain separation network (DSN) (Bousmalis et al. 2016) was proposed to explicitly separate private representations for each domain and shared ones between the source and target domains. The private representations were learned by defining a difference loss via a soft orthogonality constraint between the shared and private representations while the shared representations were learned by DANN or MMD mentioned above. With the help of reconstruction through private and shared representations together, the classifier trained on the shared representations can better generalize across domains. Since our work focuses on learning the shared representations, it can also be integrated into DSN easily.

Wasserstein Metric

Before we introduce our domain invariant feature representation learning approach, we first give a brief introduction of the Wasserstein metric. The Wasserstein metric is a distance measure between probability distributions on a given metric space (M, ρ), where $\rho(x, y)$ is a distance function for two

instances x and y in the set M. The p-th Wasserstein distance between two Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$W_p(\mathbb{P},\mathbb{Q}) = \Big(\inf_{\mu\in\Gamma(\mathbb{P},\mathbb{Q})}\int \rho(x,y)^p d\mu(x,y)\Big)^{1/p}, \quad (1)$$

where $\mathbb{P},\mathbb{Q} \in \{\mathbb{P} : \int \rho(x,y)^p d\mathbb{P}(x) < \infty, \forall y \in M\}$ are two probability measures on M with finite p-th moment and $\Gamma(\mathbb{P},\mathbb{Q})$ is the set of all measures on $M \times M$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Wasserstein metric arises in the problem of optimal transport: $\mu(x,y)$ can be viewed as a randomized policy for transporting a unit quantity of some material from a random location x to another location y while satisfying the marginal constraint $x \sim \mathbb{P}$ and $y \sim \mathbb{Q}$. If the cost of transporting a unit of material from $x \in \mathbb{P}$ to $y \in \mathbb{Q}$ is given by $\rho(x,y)^p$, then $W_p(\mathbb{P},\mathbb{Q})$ is the minimum expected transport cost.

The Kantorovich-Rubinstein theorem shows that when M is separable, the dual representation of the first Wasserstein distance (Earth-Mover distance) can be written as a form of integral probability metric (Villani 2008)

$$W_1(\mathbb{P},\mathbb{Q}) = \sup_{\|f\|_L\le 1} \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x\sim\mathbb{Q}}[f(x)], \quad (2)$$

where the Lipschitz semi-norm is defined as $\|f\|_L = \sup|f(x)-f(y)|/\rho(x,y)$. In this paper, for simplicity, Wasserstein distance represents the first Wasserstein distance.

Wasserstein Distance Guided Reprensentation Learning

Problem Definition

In unsupervised domain adaptation problem, we have a labeled source dataset $X^s = \{(x_i^s, y_i^s)\}_{i=1}^{n^s}$ of n^s samples from the source domain $\mathcal{D}_s$ which is assumed sufficient to train an accurate classifier, and an unlabeled target dataset $X^t = \{x_j^t\}_{j=1}^{n^t}$ of n^t samples from the target domain $\mathcal{D}_t$. It is assumed that the two domains share the same feature space but follow different marginal data distributions, $\mathbb{P}_{x^s}$ and $\mathbb{P}_{x^t}$ respectively. The goal is to learn a transferable classifier $\eta(x)$ to minimize target risk $\epsilon_t = \Pr_{(x,y)\sim\mathcal{D}_t}[\eta(x) \neq y]$ using all the given data.

Domain Invariant Representation Learning

The challenge of unsupervised domain adaptation mainly lies in the fact that two domains have different data distributions. Thus the model trained with source domain data may be highly biased in the target domain. To solve this problem, we propose a new approach to learn feature representations invariant to the change of domains by minimizing empirical Wasserstein distance between the source and target representations through adversarial training.

In our adversarial representation learning approach, there is a feature extractor which can be implemented by a neural network. The feature extractor is supposed to learn the domain invariant feature representations from both domains. Given an instance $x \in \mathbb{R}^m$ from either domain, the feature extractor learns a function $f_g : \mathbb{R}^m \to \mathbb{R}^d$ that maps the instance to a d-dimensional representation with corresponding network parameter θ_g. And then in order to reduce the discrepancy between the source and target domains, we use the domain critic, as suggested in (Arjovsky, Chintala, and Bottou 2017), whose goal is to estimate the Wasserstein distance between the source and target representation distributions. Given a feature representation $h = f_g(x)$ computed by the feature extractor, the domain critic learns a function $f_w : \mathbb{R}^d \to \mathbb{R}$ that maps the feature representation to a real number with parameter θ_w. Then the Wasserstein distance between two representation distributions $\mathbb{P}_{h^s}$ and $\mathbb{P}_{h^t}$, where $h^s = f_g(x^s)$ and $h^t = f_g(x^t)$, can be computed according to Eq. (2)

$$\begin{aligned} W_1(\mathbb{P}_{h^s},\mathbb{P}_{h^t}) &= \sup_{\|f_w\|_L\le 1} \mathbb{E}_{\mathbb{P}_{h^s}}[f_w(h)] - \mathbb{E}_{\mathbb{P}_{h^t}}[f_w(h)] \\ &= \sup_{\|f_w\|_L\le 1} \mathbb{E}_{\mathbb{P}_{x^s}}[f_w(f_g(x))] - \mathbb{E}_{\mathbb{P}_{x^t}}[f_w(f_g(x))]. \end{aligned} \quad (3)$$

If the parameterized family of domain critic functions $\{f_w\}$ are all 1-Lipschitz, then we can approximate the empirical Wasserstein distance by maximizing the domain critic loss $\mathcal{L}_{wd}$ with respect to parameter θ_w

$$\mathcal{L}_{wd}(x^s,x^t) = \frac{1}{n^s}\sum_{x^s\in X^s} f_w(f_g(x^s)) - \frac{1}{n^t}\sum_{x^t\in X^t} f_w(f_g(x^t)). \quad (4)$$

Here comes the question of enforcing the Lipschitz constraint. (Arjovsky, Chintala, and Bottou 2017) proposed to clip the weights of domain critic within a compact space $[-c,c]$ after each gradient update. However (Gulrajani et al. 2017) pointed out that weight clipping will cause capacity underuse and gradient vanishing or exploding problems. As suggested in (Gulrajani et al. 2017), a more reasonable way is to enforce gradient penalty $\mathcal{L}_{grad}$ for the domain critic parameter θ_w

$$\mathcal{L}_{grad}(\hat{h}) = (\|\nabla_{\hat{h}} f_w(\hat{h})\|_2 - 1)^2, \quad (5)$$

where the feature representations $\hat{h}$ at which to penalize the gradients are defined not only at the source and target representations but also at the random points along the straight line between source and target representation pairs. So we can finally estimate the empirical Wasserstein distance by solving the problem

$$\max_{\theta_w}\{\mathcal{L}_{wd} - \gamma\mathcal{L}_{grad}\} \quad (6)$$

where γ is the balancing coefficient.

Since the Wasserstein distance is continuous and differentiable almost everywhere, we can first train the domain critic to optimality. Then by fixing the optimal parameter of domain critic and minimizing the estimator of Wasserstein distance, the feature extractor network can learn feature representations with domain discrepancy reduced. Up to now the representation learning can be achieved by solving the minimax problem

$$\min_{\theta_g}\max_{\theta_w}\{\mathcal{L}_{wd} - \gamma\mathcal{L}_{grad}\} \quad (7)$$

where γ should be set 0 when optimizing the minimum operation since the gradient penalty should not guide the representation learning process. By iteratively learning feature

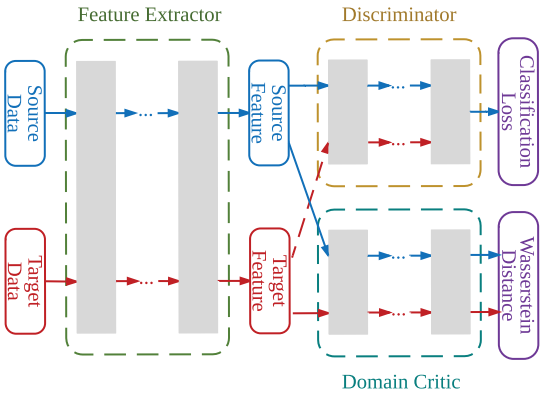


Figure 1: WDGRL Combining with Discriminator.

representations with lower Wasserstein distance, the adversarial objective can finally learn domain invariant feature representations.

Combining with Discriminator

As mentioned above, our final goal is to learn a high-performance classifier for the target domain. However, the process of WDGRL is in an unsupervised setting, which may result in that the learned domain invariant representations are not discriminative enough. Hence it is necessary to incorporate the supervision signals of source domain data into the representation learning process as in DANN (Ganin et al. 2016). Next we further introduce the combination of the representation learning approaches and a discriminator, of which the overview framework is given by Figure 1. A detailed algorithm of the combination is given in Algorithm 1.

We further add several layers as the discriminator after the feature extractor network. Since WDGRL guarantees transferability of the learned representations, the shared discriminator can be directly applied to target domain prediction when training finished. The objective of the discriminator $f_c : \mathbb{R}^d \to \mathbb{R}^l$ is to compute the softmax prediction with parameter θ_c where l is the number of classes. The discriminator loss function is defined as the cross-entropy between the predicted probabilistic distribution and the one-hot encoding of the class labels given the labeled source data:

$$\mathcal{L}_c(x^s, y^s) = -\frac{1}{n^s}\sum_{i=1}^{n^s}\sum_{k=1}^{l} 1(y_i^s = k) \cdot \log f_c(f_g(x_i^s))_k, \tag{8}$$

where $1(y_i^s = k)$ is the indicator function and $f_c(f_g(x_i^s))_k$ corresponds to the k-th dimension value of the distribution $f_c(f_g(x_i^s))$. By combining the discriminator loss, we attain our final objective function

$$\min_{\theta_g,\theta_c}\left\{\mathcal{L}_c + \lambda \max_{\theta_w}\left[\mathcal{L}_{wd} - \gamma\mathcal{L}_{grad}\right]\right\}, \tag{9}$$

where λ is the coefficient that controls the balance between discriminative and transferable feature learning and γ should be set 0 when optimizing the minimum operator.

Note that this algorithm can be trained by the standard back-propagation with two iterative steps. In a mini-batch

Algorithm 1 Wasserstein Distance Guided Representation Learning Combining with Discriminator

Require: source data X^s; target data X^t; minibatch size m; critic training step n; coefficient γ, λ; learning rate for domain critic α_1; learning rate for classification and feature learning α_2

Initialize feature extractor, domain critic, discriminator with random weights $\theta_g, \theta_w, \theta_c$
repeat
  Sample minibatch $\{x_i^s, y_i^s\}_{i=1}^m$, $\{x_i^t\}_{i=1}^m$ from X^s and X^t
  for $t = 1, ..., n$ **do**
    $h^s \leftarrow f_g(x^s)$, $h^t \leftarrow f_g(x^t)$
    Sample h as the random points along straight lines between h^s and h^t pairs
    $\hat{h} \leftarrow \{h^s, h^t, h\}$
    $\theta_w \leftarrow \theta_w + \alpha_1 \nabla_{\theta_w}[\mathcal{L}_{wd}(x^s, x^t) - \gamma\mathcal{L}_{grad}(\hat{h})]$
  end for
  $\theta_c \leftarrow \theta_c - \alpha_2 \nabla_{\theta_c}\mathcal{L}_c(x^s, y^s)$
  $\theta_g \leftarrow \theta_g - \alpha_2 \nabla_{\theta_g}[\mathcal{L}_c(x^s, y^s) + \mathcal{L}_{wd}(x^s, x^t)]$
until $\theta_g, \theta_w, \theta_c$ converge

containing labeled source data and unlabeled target data, we first train the domain critic network to optimality by optimizing the max operator via gradient ascent and then update the feature extractor by minimizing the classification loss computed by labeled source data and the estimated Wasserstein distance simultaneously. The learned representations can be domain invariant and target discriminative since the parameter θ_g receives the gradients from both the domain critic and the discriminator loss.

Theoretical Analysis

In this section, we give some theoretical analysis about the advantages of using Wasserstein distance for domain adaptation.

Gradient Superiority In domain adaptation, to minimize the divergence between the data distributions $\mathbb{P}_{x^s}$ and $\mathbb{P}_{x^t}$, the symmetric feature-based methods learn a transformation function to map the data from the original space to a common latent space with a distance measure. There are two situations after the mapping: i). The two mapped feature distributions have supports that lie on low dimensional manifolds (Narayanan and Mitter 2010) in the latent space. In such situation, there will be a gradient vanishing problem if adopting the domain classifier to make data indistinguishable while Wasserstein distance could provide reliable gradients (Arjovsky, Chintala, and Bottou 2017). ii). The feature representations may fill in the whole space since the feature mapping usually reduces dimensionality. However, if a data point lies in the regions where the probability of one distribution could be ignored compared with the other distribution, it makes no contributions to the gradients with traditional cross-entropy loss since the gradient computed by this data point is almost 0. If we adopt Wasserstein distance as the distance measure, stable gradients can be provided wherever. The detailed analysis is provided in the supplementary material[1]. So theoretically

[1]https://arxiv.org/abs/1707.01217

in either situation, WDGRL can perform better than previous adversarial adaptation methods (Ganin et al. 2016; Tzeng et al. 2017).

Genralization Bound (Redko, Habrard, and Sebban 2016) proved that the target error can be bounded by the Wasserstein distance for empirical measures. However, the generalization bound exists when assuming the hypothesis class is a unit ball in RKHS and the transport cost function is RKHS distance. In this paper we prove the generalization bound in terms of the Kantorovich-Rubinstein dual formulation under a different assumption.

We first formalize some notations that will be used in the following statements. Let $\mathcal{X}$ be an instance set and $\{0, 1\}$ be the label set for binary classification. We denote by μ_s the distribution of source instances on $\mathcal{X}$ and use μ_t for the target domain. We denote that two domains have the same labeling function $f : \mathcal{X} \to [0, 1]$ which is always assumed to hold in domain adaptation problem. A hypothesis class H is a set of predictor functions, $\forall h \in H, h : \mathcal{X} \to [0, 1]$. The probability according to the distribution μ_s that a hypothesis h disagrees with the labeling function f (which can also be a hypothesis) is defined as $\epsilon_s(h, f) = \mathbb{E}_{x\in\mu_s}[|h(x) - f(x)|]$. We use the shorthand $\epsilon_s(h) = \epsilon_s(h, f)$ and $\epsilon_t(h)$ is defined the same. We now present the Lemma that introduces Wasserstein distance to relate the source and target errors.

Lemma 1. *Let $\mu_s, \mu_t \in \mathcal{P}(\mathcal{X})$ be two probability measures. Assume the hypotheses $h \in H$ are all K-Lipschitz continuous for some K. Then the following holds*

$$\epsilon_t(h, h') \le \epsilon_s(h, h') + 2KW_1(\mu_s, \mu_t) \tag{10}$$

for every hypothesis $h, h' \in H$.

Proof. We first prove that for every K-Lipschitz continuous hypotheses $h, h' \in H$, $|h - h'|$ is $2K$-Lipschitz continuous. Using the triangle inequality, we have

$$\begin{aligned}|h(x)-h'(x)| &\le |h(x)-h(y)|+|h(y)-h'(x)| \\ &\le |h(x)-h(y)|+|h(y)-h'(y)|+|h'(x)-h'(y)|\end{aligned} \tag{11}$$

and thus for every $x, y \in \mathcal{X}$,

$$\begin{aligned}\frac{|h(x)-h'(x)|-|h(y)-h'(y)|}{\rho(x,y)} &\le \frac{|h(x)-h(y)|+|h'(x)-h'(y)|}{\rho(x,y)} \\ &\le 2K.\end{aligned} \tag{12}$$

Then for every hypothesis h, h', we have

$$\begin{aligned}\epsilon_t(h,h')-\epsilon_s(h,h') &= \mathbb{E}_{\mu_t}[|h(x)-h'(x)|]-\mathbb{E}_{\mu_s}[|h(x)-h'(x)|] \\ &\le \sup_{\|f\|_L\le 2K} \mathbb{E}_{\mu_t}[f(x)]-\mathbb{E}_{\mu_s}[f(x)] \\ &= 2KW_1(\mu_s,\mu_t)\end{aligned} \tag{13}$$

□

Theorem 1. *Under the assumption of Lemma 1, for every $h \in H$ the following holds*

$$\epsilon_t(h) \le \epsilon_s(h) + 2KW_1(\mu_s, \mu_t) + \lambda \tag{14}$$

where λ is the combined error of the ideal hypothesis h^* that minimizes the combined error $\epsilon_s(h) + \epsilon_t(h)$.

Proof.

$$\begin{aligned}\epsilon_t(h) &\le \epsilon_t(h^*) + \epsilon_t(h^*, h) \\ &= \epsilon_t(h^*) + \epsilon_s(h, h^*) + \epsilon_t(h^*, h) - \epsilon_s(h, h^*) \\ &\le \epsilon_t(h^*) + \epsilon_s(h, h^*) + 2KW_1(\mu_s, \mu_t) \\ &\le \epsilon_t(h^*) + \epsilon_s(h) + \epsilon_s(h^*) + 2KW_1(\mu_s, \mu_t) \\ &= \epsilon_s(h) + 2KW_1(\mu_s, \mu_t) + \lambda\end{aligned} \tag{15}$$

□

Thus the generalization bound of applying Wasserstein distance between domain distributions has been proved, while the proof of using empirical measures on the source and target domain samples can be further proved according to Theorem 2.1 in (Bolley, Guillin, and Villani 2007) as the same way in (Redko, Habrard, and Sebban 2016) and this proof is provided in the supplementary material.

The assumption made here is to specify the hypothesis class is K-Lipschitz continuous for some K. While it may seem too restrictive, in fact the hypotheses are always implemented by neural networks where the basic linear mapping functions and the activation functions such as sigmoid and relu are all Lipschitz continuous, so the assumption is not that strong and can be fulfilled. And the weights in neural networks are always regularized to avoid overfitting which means the constant K will not be too large. Compared with the proof in (Redko, Habrard, and Sebban 2016) the assumptions are different and can be used for different cases.

Application to Adaptation Frameworks

WDGRL can be integrated into existing feature-based domain adaptation frameworks (Tzeng et al. 2014; Long et al. 2015; Zhuang et al. 2015; Long et al. 2016; Bousmalis et al. 2016). These frameworks are all symmetric feature-based and aim to learn domain invariant feature representations for adaptation using divergence measures such as MMD and DANN. We provide a promising alternative WDGRL to learn domain invariant representations, which can replace the MMD or DANN. We should point out that although WDGRL has gradient advantage over DANN, it takes more time to estimate the Wasserstein distance. Although we only apply WDGRL on one hidden layer, it can also be applied on multilayer structures as implemented in (Long et al. 2015).

Experiments

In this section, we evaluate the efficacy of our approach on sentiment and image classification adaptation datasets. Compared with other domain invariant representation learning approaches, WDGRL achieves better performance on average. More experimental results including synthetic experiment are provided in the supplementary material.

Datasets

Amazon review benchmark dataset. The Amazon review dataset[2] (Blitzer et al. 2007) is one of the most widely used benchmarks for domain adaptation and sentiment analysis. It is collected from product reviews from Amazon.com and

[2]https://www.cs.jhu.edu/~mdredze/datasets/sentiment/

contains four types (domains), namely books (B), DVDs (D), electronics (E) and kitchen appliances (K). For each domain, there are 2,000 labeled reviews and approximately 4,000 unlabeled reviews (varying slightly across domains) and the classes are balanced. In our experiments, for easy computation, we follow (Chen et al. 2012) to use the 5,000 most frequent terms of unigrams and bigrams as the input and totally $A_4^2 = 12$ adaptation tasks are constructed.

Office-Caltech object recognition dataset. The Office-Caltech dataset[3] released by (Gong et al. 2012) is comprised of 10 common categories shared by the Office-31 and Caltech-256 datasets. In our experiments, we construct 12 tasks across 4 domains: Amazon (A), Webcam (W), DSLR (D) and Caltech (C), with 958, 295, 157 and 1,123 image samples respectively. In our experiments, Decaf features are used as the input. Decaf features (Donahue et al. 2014) are the 4096-dimensional FC7-layer hidden activations extracted by the deep convolutional neural network AlexNet.

Compared Approaches

We mainly compare our proposed approach with domain adversarial neural network (DANN) (Ganin et al. 2016), maximum mean discrepancy metric (MMD) (Gretton et al. 2012) and deep correlation alignment (CORAL) (Sun and Saenko 2016) since these approaches and our proposed WDGRL all aim at learning the domain invariant feature representations, which are crucial to reduce the domain discrepancy. Other domain adaptation frameworks (Bousmalis et al. 2016; Tzeng et al. 2014; Long et al. 2015; 2016; Zhuang et al. 2015) are not included in the comparison, because these frameworks focus on adaptation architecture design and all compared approaches can be easily integrated into these frameworks.

S-only: As an empirical lower bound, we train a model using the labeled source data only, and test it on the target test data directly.

MMD: The MMD metric is a measurement of the divergence between two probability distributions from their samples by computing the distance of mean embeddings in RKHS.

DANN: DANN is an adversarial representation learning approach that a domain classifier aims at distinguishing the learned source/target features while the feature extractor tries to confuse the domain classifier.

CORAL: Deep correlation alignment minimizes domain discrepancy by aligning the second-order statistics of the source and target distributions and can be applied to the layer activations in neural networks.

Implementation Details

We implement all our experiments[4] using TensorFlow and the models are all trained with Adam optimizer. We follow the evaluation protocol in (Long et al. 2013) and evaluate all compared approaches through grid search on the hyperparameter space, and report the best results of each approach. For each approach we use a batch size of 64 samples in total with 32 samples from each domain, and a fixed learning rate 10^{-4}. All compared approaches are combined with a discriminator to learn both domain invariant and discriminative representations and to conduct the classification task.

We use standard multi-layer perceptron (MLP) as the basic network architecture. MLP is sufficient to handle all the problems in our experiments. For Amazon review dataset the network is designed with one hidden layer of 500 nodes, relu activation function and softmax output function, while the network for Office-Caltech dataset has two hidden layers of 500 and 100 nodes. For each dataset the same network architecture is used for all compared approaches and these approaches are all applied on the last hidden layer.

For the MMD experiments we follow the suggestions of (Bousmalis et al. 2016) and use a linear combination of 19 RBF kernels with the standard deviation parameters ranging from 10^{-6} to 10^{6}. As for DANN implementation, we add a gradient reversal layer (GRL) and then a domain classifier with one hidden layer of 100 nodes. And the CORAL approach computes a distance between the second-order statistics (covariances) of the source and target features and the distance is defined as the squared Frobenius norm. For each approach, the corresponding loss term is added to the classification loss with a coefficient for the trade-off. And the coefficients are tuned different to achieve the best results for each approach.

Our approach is easy to implement according to Algorithm 1. In our experiments, the domain critic network is designed with a hidden layer of 100 nodes. The training steps n is 5 which is chosen for fast computation and sufficient optimization guarantee for the domain critic, and the learning rate for the domain critic is 10^{-4}. We penalize the gradients not only at source/target representations but also at the random points along the straight line between the source and target pairs and the coefficient γ is set to 10 as suggested in (Gulrajani et al. 2017).

Results and Discussion

Amazon review benchmark dataset. The challenge of cross domain sentiment analysis lies in the distribution shift as different words are used in different domains. Table 1

Table 1: Performance (accuracy %) on Amazon review dataset.

	S-only	MMD	DANN	CORAL	WDGRL
B → D	81.09	82.57	82.07	82.74	**83.05**
B → E	75.23	80.95	78.98	82.93	**83.28**
B → K	77.78	83.55	82.76	84.81	**85.45**
D → B	76.46	79.93	79.35	**80.81**	80.72
D → E	76.24	82.59	81.64	83.49	**83.58**
D → K	79.68	84.15	83.41	85.35	**86.24**
E → B	73.37	75.72	75.95	76.91	**77.22**
E → D	73.79	77.69	77.58	78.08	**78.28**
E → K	86.64	87.37	86.63	87.87	**88.16**
K → B	72.12	75.83	75.81	76.95	**77.16**
K → D	75.79	78.05	78.53	79.11	**79.89**
K → E	85.92	86.27	86.11	**86.83**	86.29
AVG	77.84	81.22	80.74	82.16	**82.43**

[3]https://cs.stanford.edu/~jhoffman/domainadapt/

[4]Experiment code: https://github.com/RockySJ/WDGRL.

shows the detailed comparison results of these approaches in 12 transfer tasks. As we can see, our proposed WDGRL outperforms all other compared approaches in 10 out of 12 domain adaptation tasks, and it achieves the second highest scores in the remaining 2 tasks. We find that as adversarial adaptation approaches, WDGRL outperforms DANN, which is consistent with our theoretical analysis that WDGRL has more reliable gradients. MMD and CORAL are both non-parametric and have lower computational cost than WDGRL, while their classification performances are also lower than WDGRL.

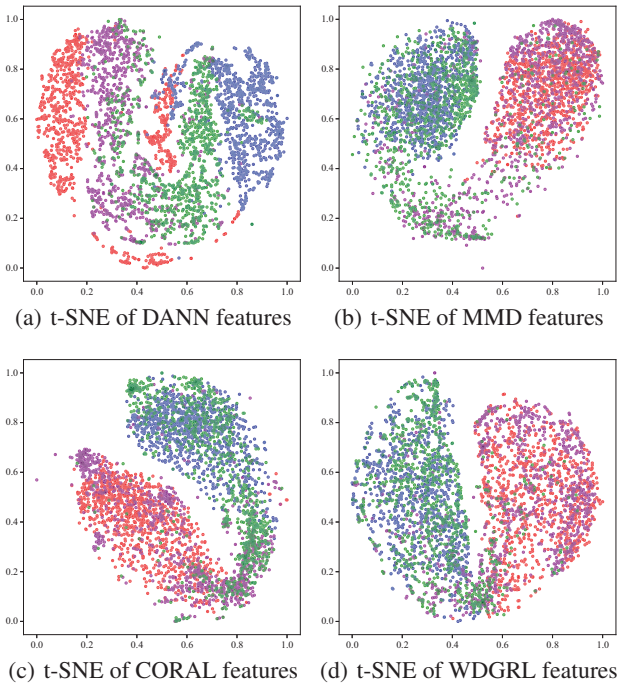
Office-Caltech object recognition dataset. Table 2 shows the results of our experiments on Office-Caltech dataset. We observe that our approach achieves better performance than other compared approaches on most tasks. Office-Caltech dataset is small since there are only hundreds of images in one domain and it is a 10-class classification problem. Thus we can draw a conclusion that the empirical Wasserstein distance can also be applied to small-scale datasets adaptation effectively. We note that CORAL performs better than MMD in Amazon review dataset while it performs worse than MMD in Office-Caltech dataset. A possible reason is that the reasonable covariance alignment approach requires large samples. On the other hand, we can see that these different approaches have different performances on different adaptation tasks.

Feature Visualization

We randomly choose the D→E domain adaptation task of Amazon review dataset and plot in Figure 2 the t-SNE visualization following (Donahue et al. 2014; Long et al. 2016) to visualize the learned feature representations. In these figures, red and blue points represent positive and negative samples of the source domain, purple and green points represent positive and negative samples of the target domain. A transferable feature mapping should cluster red (blue) and purple (green) points together, and meanwhile classification can be easily conducted between purple and green points. We can see that almost all approaches learn discriminative and domain invariant feature representations to some extent. And representations learned by WDGRL are more transferable since the classes between the source and target domains align better and the region where purple and green points mix together is smaller.

Table 2: Performance (accuracy %) on Office-Caltech dataset with Decaf features.

	S-only	MMD	DANN	CORAL	WDGRL
A → C	84.55	**88.62**	87.80	86.18	86.99
A → D	81.05	90.53	82.46	91.23	**93.68**
A → W	75.59	**91.58**	77.81	90.53	89.47
W → A	79.82	92.22	82.98	88.39	**93.67**
W → D	98.25	**100**	**100**	**100**	**100**
W → C	79.67	88.62	81.30	88.62	**89.43**
D → A	84.56	90.11	84.70	85.75	**91.69**
D → W	96.84	**98.95**	**98.95**	97.89	97.89
D → C	80.49	87.80	82.11	85.37	**90.24**
C → A	92.35	93.14	93.27	93.01	**93.54**
C → W	84.21	91.58	89.47	**92.63**	91.58
C → D	87.72	91.23	91.23	89.47	**94.74**
AVG	85.44	92.03	87.67	90.76	**92.74**



(a) t-SNE of DANN features (b) t-SNE of MMD features

(c) t-SNE of CORAL features (d) t-SNE of WDGRL features

Figure 2: Feature visualization of the D→E task in Amazon review dataset.

Conclusions

In this paper, we propose a new adversarial approach WDGRL to learn domain invariant feature representations for domain adaptation. WDGRL can effectively reduce the domain discrepancy taking advantage of the gradient property of Wasserstein distance and the transferability is guaranteed by the generalization bound. Our proposed approach could be further integrated into other domain adaptation frameworks (Bousmalis et al. 2016; Tzeng et al. 2014; Long et al. 2015; 2016; Zhuang et al. 2015) to attain better transferability. Empirical results on sentiment and image classification domain adaptation datasets demonstrate that WDGRL outperforms the state-of-the-art domain invariant feature learning approaches. In future work, we will investigate more sophisticated architectures for tasks on image data as well as integrate WDGRL into existing adaptation frameworks.

Acknowledgement

This work is financially supported by NSFC (61702327) and Shanghai Sailing Program (17YF1428200).

References

Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; and Marchand, M. 2014. Domain-adversarial neural networks. *arXiv:1412.4446*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv:1701.07875*.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *NIPS*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning*.

Blitzer, J.; Dredze, M.; Pereira, F.; et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*.

Bolley, F.; Guillin, A.; and Villani, C. 2007. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *NIPS*.

Chen, M.; Xu, Z.; Weinberger, K.; and Sha, F. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv:1206.4683*.

Chen, M.; Chen, Y.; and Weinberger, K. Q. 2011. Automatic feature decomposition for single view co-training. In *ICML*.

Chen, M.; Weinberger, K. Q.; and Blitzer, J. 2011. Co-training for domain adaptation. In *NIPS*.

Chu, W.-S.; De la Torre, F.; and Cohn, J. F. 2013. Selective transfer machine for personalized facial action unit detection. In *CVPR*.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2017. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*.

Courty, N.; Flamary, R.; and Tuia, D. 2014. Domain adaptation with regularized optimal transport. In *ECML/PKDD*.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*.

Duan, L.; Xu, D.; and Chang, S.-F. 2012. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*. IEEE.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *JMLR*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved training of wasserstein gans. *arXiv:1704.00028*.

Hoffman, J.; Rodner, E.; Donahue, J.; Kulis, B.; and Saenko, K. 2014. Asymmetric and category invariant feature transformations for domain adaptation. *IJCV*.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; Schölkopf, B.; et al. 2007. Correcting sample selection bias by unlabeled data. *NIPS*.

Kandemir, M. 2015. Asymmetric transfer learning with deep gaussian processes. In *ICML*.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer feature learning with joint distribution adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

Long, M.; Wang, J.; Cao, Y.; Sun, J.; and Philip, S. Y. 2016. Deep learning of transferable representation for scalable domain adaptation. *TKDE*.

Luo, L.; Wang, X.; Hu, S.; Wang, C.; Tang, Y.; and Chen, L. 2017. Close yet distinctive domain adaptation. *arXiv:1704.04235*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation with multiple sources. In *NIPS*.

Narayanan, H., and Mitter, S. 2010. Sample complexity of testing the manifold hypothesis. In *NIPS*.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*.

Redko, I.; Habrard, A.; and Sebban, M. 2016. Theoretical analysis of domain adaptation with optimal transport. *arXiv:1610.04420*.

Rozantsev, A.; Salzmann, M.; and Fua, P. 2016. Beyond sharing weights for deep domain adaptation. *arXiv:1603.06432*.

Sun, B., and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV 2016 Workshops*. Springer.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI*.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. *arXiv:1702.05464*.

Villani, C. 2008. *Optimal transport: old and new*.

Weiss, K.; Khoshgoftaar, T. M.; and Wang, D. 2016. A survey of transfer learning. *Journal of Big Data*.

Zhuang, F.; Cheng, X.; Luo, P.; Pan, S. J.; and He, Q. 2015. Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI*.