DUAL ADVERSARIAL TRAINING FOR UNSUPERVISED DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks obtain remarkable achievements in diverse real-world applications. However, their success relies on the availability of large amounts of labeled data. A trained model may fail to generalize well on a domain whose distribution differs from the training data distribution. Collecting abundant labeled data for all domains of interest are expensive and time-consuming, sometimes even impossible. Domain adaptation sets out to address this problem, aiming to leverage labeled data in the source domain to learn a good predictive model for the target domain whose labels are scarce or unavailable. A mainstream approach is adversarial domain adaptation, which learns domain invariant-features by performing alignment across different distributions. Most domain adaptation methods focus on reducing the divergence between two domains to make the improvement. A prerequisite of domain adaptation is the adaptability, which is measured by the expected error of the ideal joint hypothesis on the source and target domains, should be kept at a small value in the process of domain alignment. However, adversarial learning may degrade the adaptability, since it distorts the original distributions by suppressing the domain-specific information. In this paper, we propose an approach, which focuses on strengthening the model's adaptability, for domain adaptation. Our proposed dual adversarial training (DAT) method introduces class-invariant features to enhance the discriminability of the latent space without sacrificing the transferability. The class-invariant features, extracted from the source domain, can play a positive role in the classification on the target domain. We demonstrate the effectiveness of our method by yielding state-of-the-art results on several benchmarks.

1 INTRODUCTION

Deep learning has achieved great success in many machine learning tasks, such as image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012), and natural language processing (Wu & Guo, 2019). These tasks generally assume that there exists sufficient data to train a good predictive model in the domain of interest. Unfortunately, in many real-world cases, it is always difficult to collect sufficient labeled data for each domain of interest. Thus, it is of great importance to explore how to apply knowledge learned from a label-dense domain, referred as the source domain, to a label-scarce domain, referred as the target domain. Even though the source and target domains share the same label space, due to the existence of domain shift, deep neural network based models trained on the source domain are inclined to make spurious predictions on the target domain (Pan & Yang, 2009). In order to mitigate the harmful effects of domain shift, domain adaptation is proposed to learn transferable knowledge between domains such that a model trained on the source domain can simultaneously perform well on the target domain. In this work, we focus on a more challenging setting: unsupervised domain adaptation, where there exists no labeled data in the target domain.

Motivated by the domain adaptation theory (Ben-David et al., 2007; 2010), which suggests that the expected error rate on the target domain is bounded by three elements: the expected error rate on the source domain, the divergence between the two domains, and the adaptability which is quantified as the error rate of the ideal joint hypothesis on both source and target domains. Most domain adaptation methods attempt to deal with domain shift by reducing the divergence between domains (Pan & Yang, 2009). Early unsupervised domain adaptation methods reweighed the source instances

based on their associations to the target domain with regard to the human engineered features (Gong et al., 2013). Recent methods are largely based on deep neural networks, which can automatically extract features from massive data. One possible way is to minimize some measures of domain distance such as maximum mean discrepancy (MMD) (Yan et al., 2017) and correlation distances (Sun & Saenko, 2016). On par with these distance minimizing methods, adversarial domain adaptation introduces adversarial learning (Goodfellow et al., 2014) to impose domain-invariant constraint to the latent space. The learned domain-invariant features are supposed to contain task discriminative information which can be applied to conduct classification. These adversarial domain adaptation methods (Long et al., 2018b; Wu et al., 2020) have yielded remarkable performance gains.

For the domain adaptation methods, an essential prerequisite is the adaptability should remain at a small value in the domain alignment process. When the adaptability is poor, a good domain adaptation model can not be obtained. Since domain alignment will inevitably distort the original feature distributions and enlarge the value of adaptability, adversarial domain adaptation is risky in this regard. In order to address this issue, we propose a novel dual adversarial training (DAT) framework, which introduces class-invariant features in domain adaptation. Considering the importance of domain-invariant features in domain adaptation, it is intuitive to explore the feasibility of applying class-invariant features in domain adaptation. The generation of class-invariant feature can be formulated into a two-player mini-max game which is similar to that of domain-invariant feature. The class-invariant features are supposed to contain domain-specific information. Our proposed DAT utilizes class-invariant features drawn from the source domain to adapt the classifier from the source to the target with guaranteed adaptability. This work aims to provide an alternative for the mainstream domain adaptation methods which focus on reducing the domain divergence. Our method demonstrates that we can improve the generalization ability of the model on the target domain via optimizing the adaptability and the class-invariant features, which contain domain-specific information of the source domain, can benefit the classification on the target domain. The empirical experimental results on four benchmarks show the promise of our approach by yielding state-of-the-art classification results. The contributions of our paper are summarized as follows:

- We propose a dual adversarial training framework, which introduces class-invariant features of the source domain in adversarial domain adaptation to guarantee a good adaptability, for unsupervised domain adaptation.
- The proposed approach demonstrates that we can train a good domain adaptation model by optimizing the generalization error rate of the ideal joint labeling function on the source and target domains. This method can be regarded as an alternative for the mainstream domain adaptation methods which focus on reducing domain divergence.
- We empirically validate the efficacy of our method on four unsupervised domain adaptation benchmarks, our proposed DAT can yield state-of-the-art results.

2 Related Work

The main objective of domain adaptation is to transfer the knowledge learned from a label-abundant source domain to a label-scarce target domain. Unsupervised domain adaptation tackles a more challenging scenario where there is no direct access to the label information of the target domain. Domain adaptation methods based on deep neural networks have achieved impressive success in recent years. Some methods directly minimize domain discrepancy measured by certain metrics, such as: maximum mean discrepancy (MMD) (Yan et al., 2017; Li et al., 2018; Long et al., 2018a) and correlation distances (Sun & Saenko, 2016). The deep domain confusion (DDC) utilized MMD in the last fully-connected layer to learn features containing transferability and discriminability (Gretton et al., 2007). The deep adaptation network (DAN) applied MMD to layers embedded in a reproducing kernel hilbert space, effectively matching higher order statistics of the two distributions (Long et al., 2015). The joint adaptation network (JAN) learned a learner by aligning the joint distributions of multiple domain-specific layers across different domains based on a joint MMD criterion (Long et al., 2017). The deep correlation alignment (CORAL) proposed to match the means and covariances of two distributions (Sun & Saenko, 2016).

Adversarial learning was pioneered by generative adversarial networks (GANs) (Goodfellow et al., 2014), which was first introduced for image generation. It plays a two player mini-max game between a generator and a discriminator. The discriminator aims to distinguish real images from

generated images, while the generator tries to deceive the discriminator. When these two networks reach an equilibrium, the generated images can not be identified by the discriminator. With insights from both the theory (Ben-David et al., 2010) and the adversarial learning (Goodfellow et al., 2014), (Ganin et al., 2016) proposed a domain discriminative neural network (DANN) which can learn domain-invariant features by exploiting adversarial learning between a domain discriminator and a feature extractor. It mapped two distributions into a shared latent space and performed domain alignment. Most recent domain adaptation methods followed this line and enhanced the domain-invariant constraint to the latent space to make the improvement. The adversarial discriminative domain adaptation (ADDA) used asymmetric feature extractors for the two domains to conduct the alignment (Tzeng et al., 2017). The multi-adversarial domain adaptation (MADA) captured multi-mode structures by re-weighting features with category predictions (Pei et al., 2018). The generate-to-adapt method (GTA) generated source-like images using source features and target-like images using target features to train the method which can minimize the distance between the generated image distributions (Sankaranarayanan et al., 2018). The cycle-consistent adversarial domain adaptation (CyCADA) implemented domain adaptation at both pixel-level and feature-level by using cycle-consistent adversarial training (Hoffman et al., 2018). The conditional adversarial domain adaptation (CDAN) conditioned the domain discriminator on discriminative information by multiplicative interactions between feature representations and predictions (Long et al., 2018b).

Different from the above methods, our proposed DAT approach extracts class-invariant features from the source domain by playing a two-player mini-max game between a feature extractor and a multinomial task discriminator, and uses the class-invariant features to optimize the adaptability during the domain alignment to improve the system performance. In conventional adversarial domain adaptation, class-invariant features, which contain domain-specific information, are regarded as noises such that they are often ignored in domain adaptation since they can deteriorate the domain-invariance of the learned features. The domain separation network (DSN) introduced domain-specific spaces for both source and target domains and demonstrated that domain-specific information can only be beneficial to its own domain (Bousmalis et al., 2016). The usage of domain-specific information of DSN in domain adaptation is limited. Our method demonstrates that the source domain-specific information can be applied in classification on the target domain and improve the accuracy. Our method aims to boost the system performance via optimizing the adaptability instead of further reducing the domain discrepancy.

3 METHOD

In this work, we consider unsupervised domain adaptation in the following setting. There exists abundant labeled instances in the source domain, $D^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with $\mathbf{x}_i^s \in \mathcal{X}$ and $y_i^s \in \mathcal{Y}$, and a set of unlabeled instances in the target domain, $D^t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ with $\mathbf{x}_i^t \in \mathcal{X}$. The data in the two domains are drawn from different distributions $\mathcal{S}$ and $\mathcal{T}$, but share the same label space. The main objective is to learn a prediction model $h : \mathcal{X} \to \mathcal{Y}$ that has a good capacity of generalizing on both source and target domains.

3.1 ADVERSARIAL DOMAIN ADAPTATION

The key idea of adversarial domain adaptation is to learn domain-invariant features that can be generalized across domains. Starting from the domain adversarial neural network (DANN) (Ganin et al., 2016), adversarial domain adaptation methods have deployed the adversarial learning strategy to learn feature representations to bridge the domain divergence. Considering DANN as an example, the base network is composed of three components: a feature extractor F, a task-specific classifier C and a binary domain discriminator D. The feature extractor $F : \mathcal{X} \to \mathbb{R}^m$ can map any input instance $\mathbf{x} \in \mathcal{X}$ from the input space $\mathcal{X}$ into a learned representation space $F(\mathbf{x}) \in \mathbb{R}^m$. The task-specific classifier $C : \mathbb{R}^m \to \mathcal{Y}$ can transform a feature vector in the latent space to the label space $\mathcal{Y}$. The domain discriminator $D : \mathbb{R}^m \to [0, 1]$ separates the source features (with domain index 0) from the target ones (with domain index 1) in the latent space. The adversarial learning is formulated as a two-player mini-max game between the feature extractor F and the domain discriminator D: the domain discriminator aims to distinguish features between the source and target domains, while the feature extractor tries to confuse the domain discriminator. The intuition is that if a strong

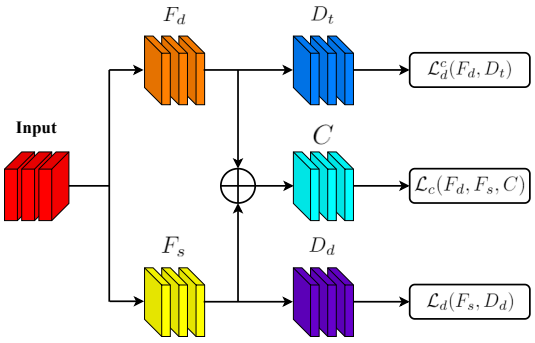


Figure 1: The architecture of the proposed dual adversarial training (DAT) method. Our DAT consists of five components: the domain-related feature extractor F_d is used to capture class-invariant features, the class-related feature extractor F_s aims to learn domain-invariant features, the multinomial task discriminator D_t identifies the label of the input feature, the binary domain discriminator D_d differentiates between the source features and the target features, and the classifier C is used to conduct the classification. $\mathcal{L}_c(F_d, F_s, C)$ is the cross-entropy loss, $\mathcal{L}_d(F_s, D_d)$ and $\mathcal{L}_d^c(F_d, D_t)$ are adversarial losses that guide the domain-invariant feature generation and the class-invariant feature generation, respectively.

domain discriminator can not identify the origin of the feature, the learned features can be regarded as domain-invariant. Formally, DANN can be formulated as:

$$\min_{F,C} \max_{D} \mathcal{L}_c(F, C) + \lambda_d \mathcal{L}_d(F, D) \tag{1}$$

$$\mathcal{L}_c(F, C) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}} \ell(C(F(\mathbf{x}^s)), y^s) \tag{2}$$

$$\mathcal{L}_d(F, D) = \mathbb{E}_{\mathbf{x}^s\sim\mathcal{S}} \log[D(F(\mathbf{x}^s))] + \mathbb{E}_{\mathbf{x}^t\sim\mathcal{T}} \log[1 - D(F(\mathbf{x}^t))] \tag{3}$$

Where $\ell(\cdot, \cdot)$ is the canonical cross-entropy loss, and λ_d is a trade-off hyperparameter.

3.2 Class-invariant Feature

In adversarial domain adaptation, the domain-invariant features, which capture the task discriminative information, play an important role. It is natural to raise the following questions: **(a)** How to extract class-invariant features. **(b)** Whether it is possible to apply class-invariant features in domain adaptation. In unsupervised domain adaptation, at training time, we have an access to a set of training instances $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ from both the source and target domains. We denote with d_i the binary variable (domain label) for each instance, which indicates whether $\mathbf{x}_i$ comes from the source domain ($d_i = 0$) or from the target domain ($d_i = 1$). These instances and their domain labels can be used to obtain domain-invariant features via adversarial learning. By replacing the binary domain discriminator D with a multinomial task discriminator D_t, we can use adversarial learning to obtain class-invariant features by processing the instances and their corresponding labels. Since we have no access to the label information of the target data, the class-invariant features can only be learned from the source domain, which can be encoded as follows:

$$\min_{F} \max_{D_t} \mathcal{L}_d^c(F, D_t) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}} \ell(D_t(F(\mathbf{x}^s)), y^s) \tag{4}$$

Considering that the class-invariant features contain domain-specific information, which has been demonstrated to be only beneficial to its own domain (Bousmalis et al., 2016). It is challenging to apply the source class-invariant features in classification on the target data.

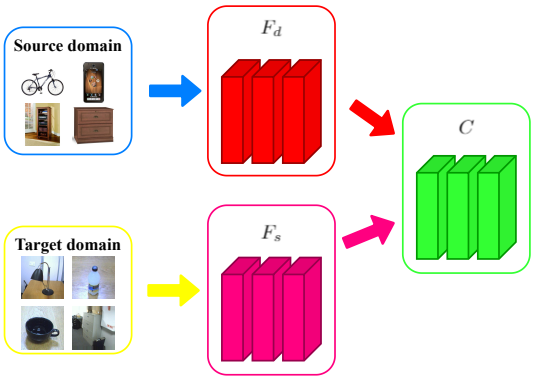


Figure 2: The evaluating process of the DAT method.

Assume an image is composed of two types of information: (1) task discriminative information, which carries the class-specific information of the image; (2) domain discriminative information, which corresponds to the context information of the image. These two types of information are supposed to represent each individual characteristic of an image. For a source instance and a target instance that share the same label, we can obtain their domain-invariant features and class-invariant features. If these features contain no noise, which means that the domain-invariant features contain no domain-related information and the class-invariant features carry no task-related information. When we combine the class-invariant features of the source instance and the domain-invariant features of the target instance, the combination is expected to contain all the essential information of the source instance. In other words, the combination of the class-invariant features of the source instance and the domain-invariant features of the target instance should represent identical characteristics of the combination of the class-invariant features and domain-invariant features of the source instance. If we feed the combination of the class-invariant features of the source instance and the domain-invariant features of the target instance to a classifier which is trained on the source domain, it is supposed to yield the accurate prediction. Therefore, it is possible to apply source class-invariant features for domain adaptation.

3.3 Dual Adversarial Training

In this work, we propose a dual adversarial training (DAT) framework for unsupervised domain adaptation. As illustrated in Figure 1, our model consists of five components: a domain-related feature extractor F_d, a multinomial task discriminator D_t, a class-related feature extractor F_s, a binary domain discriminator D_d, and a classifier C. The domain-related feature extractor F_d learns to capture the class-invariant features, while the class-related feature extractor F_s learns to capture domain-invariant features. There exists two two-player mini-max games in our approach. The first one is played between F_d and D_t, aiming to learn class-invariant features. The second one is played between F_s and D_d, aiming to learn domain-invariant features. The input data from the source domain should be involved in both games while the target data only participates in the second one. Since we need to use the combination of a class-invariant feature and a domain-invariant feature to conduct classification, we redefine the classification loss as:

$$\mathcal{L}_c(F_d, F_s, C) = \mathbb{E}_{(\mathbf{x}^s, y^s)\sim\mathcal{S}} \ell(C([F_d(\mathbf{x}^s), F_s(\mathbf{x}^s)]), y^s) \tag{5}$$

Where $[\cdot, \cdot]$ indicates the concatenation of two vectors. Formally, our proposed DAT can be formulated as:

$$\min_{F_d, F_s, C} \max_{D_t, D_d} \mathcal{L}_c(F_d, F_s, C) + \lambda_d \mathcal{L}_d(F_s, D_d) + \lambda_c \mathcal{L}_d^c(F_d, D_t) \tag{6}$$

Where λ_d and λ_c are hyperparameters that trade-off different loss functions.

3.4 TRAINING PROCEDURE

The training algorithm of DAT, which uses mini-batch stochastic gradient descent, is presented in Algorithm 1. In each iteration, the source and target samples are fed into the model to generate class-invariant features and domain-invariant features for the consequent classification. λ_d and λ_c are hyperparameters that balance different losses. When evaluating on the target domain, we need to use some source data to provide class-invariant features for classification. In our experiments, we draw samples from the source training set to provide domain-specific information when evaluating. The evaluating procedure is illustrated in Figure 2.

Algorithm 1 Stochastic gradient descent training algorithm of DAT

1: **Input:** Source domain: D^s, target domain: D^t and batch size: N.
2: **Output:** Configurations of DAT
3: **Initialize** λ_d and λ_c
4: **for** number of training iterations **do**
5: $(\mathbf{x}^s, y^s) \leftarrow$ RANDOMSAMPLE(D^s, N)
6: $(\mathbf{x}^t) \leftarrow$ RANDOMSAMPLE(D^t, N)
7: Calculate $l_D = \lambda_d \mathcal{L}_d(F_s, D_d) + \lambda_c \mathcal{L}_d^c(F_d, D_t)$;
Update D_d and D_t by ascending along gradients ∇l_D.
8: Calculate $loss = \mathcal{L}_c(F_d, F_s, C) + \lambda_d \mathcal{L}_d(F_s, D_d) + \lambda_t \mathcal{L}_d^c(F_d, D_t)$;
Update F_s, F_d and C by descending along gradients $\nabla loss$.
9: **end for**

3.5 THEORY UNDERSTANDING

Most domain adaptation methods are motivated by the theory (Ben-David et al., 2010).

Theorem 1. *(Ben-David et al., 2010) Let $\mathcal{H}$ be the hypothesis space, given two domains $\mathcal{S}$ and $\mathcal{T}$, for any $h \in \mathcal{H}$, we have*

$$\epsilon_{\mathcal{T}}(h) \leq \epsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda \tag{7}$$

where $\epsilon_{\mathcal{T}}(h)$ is the expected error on the target domain, $\epsilon_{\mathcal{S}}(h)$ is the expected error on the source domain, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ is the $\mathcal{H}\Delta\mathcal{H}$-distance between $\mathcal{S}$ and $\mathcal{T}$, which can be defined as:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) = 2 \sup_{h, h' \in \mathcal{H}} |\mathbb{E}_{\mathbf{x}\sim\mathcal{S}}[h(\mathbf{x}) \neq h'(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{T}}[h(\mathbf{x}) \neq h'(\mathbf{x})]| \tag{8}$$

and λ is the adaptability which is measured by the error of the ideal joint hypothesis h^*, defined as $h^* = \arg\min_{h\in\mathcal{H}} \epsilon_{\mathcal{S}}(h) + \epsilon_{\mathcal{T}}(h)$, on the source and target domains, such that

$$\lambda = \epsilon_{\mathcal{S}}(h^*) + \epsilon_{\mathcal{T}}(h^*) \tag{9}$$

The $\mathcal{H}\Delta\mathcal{H}$-distance measures the divergence between the source and target feature distributions. In typical adversarial domain adaptation, methods focus on minimizing the $\mathcal{H}\Delta\mathcal{H}$-distance to enhance domain-invariance of the latent space, while the classifier is simultaneously trained on the source labeled data to reduce the source error. In most cases, the adaptability λ is treated as a constant, which is expected to remain at a small value to guarantee the feasibility of domain adaptation. However, (Liu et al., 2019) demonstrated that in the process of domain alignment, diminishing domain-specific information will inevitably breaks the class discriminative structures of the original representations, which will lead to a poor λ. In this paper, we introduce class-invariant features to complement domain-invariant features by providing domain-specific information. Our proposed method focuses on optimizing adaptability to implement domain adaptation and can be regarded as an alternative for the mainstream domain adaptation approaches which focus on minimizing the $\mathcal{H}\Delta\mathcal{H}$-distance.

4 EXPERIMENTS

In this section, we evaluate our proposed DAT framework on four unsupervised domain adaptation benchmarks: Office-31, ImageCLEF-DA, VisDA-2017 and Digits.

Table 1: Accuracy (%) on Office-31.

Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50 (He et al., 2016)	68.4±0.2	96.7±0.1	99.3±0.1	68.9±0.2	62.5±0.3	60.7±0.3	76.1
DAN (Long et al., 2015)	80.5±0.4	97.1±0.2	99.6±0.1	78.6±0.2	63.6±0.3	62.8±0.2	80.4
DANN (Ganin & Lempitsky, 2015)	82.0±0.4	96.9±0.2	99.1±0.1	79.7±0.4	68.2±0.4	67.4±0.5	82.2
JAN (Long et al., 2017)	85.4±0.3	97.4±0.2	99.8±0.2	84.7±0.3	68.6±0.3	70.0±0.4	84.3
GTA (Sankaranarayanan et al., 2018)	89.5±0.5	97.9±0.3	99.8±0.4	87.7±0.5	**72.8**±0.3	**71.4**±0.4	86.5
MADA (Pei et al., 2018)	90.0±0.1	97.4±0.1	99.6±0.1	87.8±0.2	70.3±0.3	66.4±0.3	85.2
CDAN (Long et al., 2018b)	**93.1**±0.2	98.2±0.2	**100.0**±0.0	89.8±0.3	70.1±0.4	68.0±0.4	86.6
DAT (Proposed)	90.7±0.3	**99.0**±0.1	**100.0**±0.0	**90.0**±0.5	71.0±0.3	69.9±0.3	**86.8**

4.1 DATASET

Office-31 (Saenko et al., 2010) is a standard domain adaptation dataset. It contains images among 31 classes from 3 domains: Amazon (A) with 2817 images, Webcam (W) with 795 images and DSLR (D) with 498 images. We conduct evaluation on all 6 tasks: A→W, D→W, W→D, A→D, D→A and W→A.

ImageCLEF-DA is a benchmark for ImageCLEF 2014 domain adaptation challenges. It is organized by selecting 12 common classes shared by three domains: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). Each domain contains 600 images and 50 images for each class. We evaluate all methods on 6 tasks: I→P, P→I, I→C, C→I, C→P and P→C.

Visda-2017 (Peng et al., 2017) is a large simulation-to-real dataset with two domains. The source domain is termed as Synthetic which contains images obtained by rendering 3D models of the same object classes as in the real data from different angles and under different lighting conditions. The target domain is termed as Real which comprises natural images. We evaluate our methods on the task: Synthetic→Real.

Digits. We investigate three digits datasets: MNIST, USPS and Street View House Numbers (SVHN). Each dataset contains digit images of 10 classes (0-9). We adopt the experimental settings of CyCADA (Hoffman et al., 2018) with three tasks: MNIST to USPS (M→U), USPS to MNIST (U→M) and SVHN to MNIST (S→M).

4.2 COMPARISON METHODS

We compare our proposed DAT method with a number of state-of-the-art methods. Deep adaptation network (DAN) (Long et al., 2015), domain adversarial neural network (DANN) (Ganin et al., 2016), joint adaptation network (JAN) (Long et al., 2017), generate-to-adapt (GTA) (Sankaranarayanan et al., 2018), multi-adversarial domain adaptation (MADA) (Pei et al., 2018), conditional domain adaptation network (CDAN) (Long et al., 2018b), adversarial discriminative domain adaptation (ADDA) (Tzeng et al., 2017) and cycle-consistent adversarial domain adaptation (CyCADA) (Hoffman et al., 2018).

4.3 IMPLEMENTATION DETAILS

The standard evaluation protocols (Long et al., 2018b; Hoffman et al., 2018) for unsupervised domain adaptation are followed in our experiments. We use all labeled source samples and unlabeled target samples and compare the average classification accuracy based on three random experiments. No data augmentation is used in any of the experiments to allow a fair comparison. For Office-31 and ImageCLEF-DA datasets, we use ResNet-50 (He et al., 2016) pre-trained on ImageNet (Krizhevsky et al., 2012) as the backbone. For Visda-2017 dataset, we use ResNet-101 (He et al., 2016) pre-trained on ImageNet (Krizhevsky et al., 2012) as the backbone. For digits datasets, we adopt a modified version of Lenet architecture as the base network, and train models from scratch. For each backbone network, we use all its layers up to the second last one as the class-related feature extractor F_s. The domain-related feature extractor F_d adopts the same architecture as F_s. The classifier C uses a single fully-connected layer whose input dimension should be the sum of the output dimensions of F_s and F_d. For binary domain discriminator D_d, we use the same architecture as DANN

Table 2: Accuracy (%) on ImageCLEF-DA.

Method	I→P	P→I	I→C	C→I	C→P	P→C	Avg
ResNet-50 (He et al., 2016)	74.8±0.3	83.9±0.1	91.5±0.3	78.0±0.2	65.5±0.3	91.2±0.3	80.7
DAN (Long et al., 2015)	74.5±0.4	82.2±0.2	92.8±0.2	86.3±0.4	69.2±0.4	89.8±0.4	82.5
DANN (Ganin & Lempitsky, 2015)	75.0±0.6	86.0±0.3	96.2±0.4	87.0±0.5	74.3±0.5	91.5±0.6	85.0
JAN (Long et al., 2017)	76.8±0.4	88.0±0.2	94.7±0.2	89.5±0.3	74.2±0.3	91.7±0.3	85.8
MADA (Pei et al., 2018)	75.0±0.3	87.9±0.2	96.0±0.3	88.8±0.3	75.2±0.2	92.2±0.3	85.8
CDAN (Long et al., 2018b)	76.7±0.3	90.6±0.3	97.0±0.4	90.5±0.4	**74.5**±0.3	93.5±0.4	87.1
DAT (Proposed)	**77.0**±0.2	**91.1**±0.2	**97.3**±0.3	**90.8**±0.3	**74.5**±0.5	**93.7**±0.3	**87.4**

Table 3: Accuracy (%) on Digits and VisDA-2017.

Method	M→U	U→M	S→M	Avg	Method	Synthetic→Real
No Adaptation (Hoffman et al., 2018)	82.2	69.6	67.1	73.0	ResNet-101 (He et al., 2016)	52.4
DANN (Ganin & Lempitsky, 2015)	90.4	94.7	84.2	89.8	DANN (Ganin & Lempitsky, 2015)	57.4
ADDA (Tzeng et al., 2017)	89.4	90.1	86.3	88.6	DAN (Long et al., 2015)	61.1
CyCADA (Hoffman et al., 2018)	**95.6**	96.5	90.4	94.2	JAN (Long et al., 2017)	65.7
CDAN (Long et al., 2018b)	93.9	96.9	88.5	93.1	CDAN (Long et al., 2018b)	73.7
DAT (Proposed)	93.7	**97.4**	**93.1**	**94.7**	**DAT (Proposed)**	**74.4**

(Ganin & Lempitsky, 2015). The architecture of the multinomial task discriminator D_t is similar to that of D_d, the only difference is that for the last layer, a task-specific fully-connected layer is used for D_t while a sigmoid layer is used for D_d.

We implement all experiments using **PyTorch**. We adopt mini-batch SGD with momentum of 0.9 and the learning rate annealing strategy as (Ganin et al., 2016): the learning rate is adjusted by $\eta_p = \frac{\eta_0}{(1+\theta p)^\beta}$, where p denotes the process of training epochs that is normalized to be in $[0, 1]$, and we set $\eta_0 = 0.01$, $\theta = 10$, $\beta = 0.75$, which are optimized to promote convergence and low errors on the source domain. λ_d is progressively changed from 0 to 1 by multiplying to $\frac{1-exp(-\delta p)}{1+exp(-\delta p)}$, where $\delta = 10$. For all experiments, we select λ_c in the range $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$ via tuning on the unlabeled target data.

4.4 Results

The results on Office-31 are reported in Table 1. Our proposed DAT method can outperform other comparison methods on three out of six tasks: D→W, W→D and A→D. Moreover, our method can achieve the best average classification accuracy on this dataset. For the benchmark ImageCLEF-DA, the results are shown in Table 2. The proposed DAT exceeds comparison methods in all tasks, but with small performance gains. This is reasonable because the three domains in ImageCLEF-DA has identical image size and are balanced in each class. Promising results are also obtained on VisDA-2017 and Digits datasets, as presented in Table 3. For Digits datasets, our method can yield better results on two tasks: U→M and S→M, we can also achieve best average performance compared with other methods. For VisDA-2017 dataset, the DAT method produces the best average classification accuracy among all comparison methods, and outperform the baseline of ResNet-101 model pre-trained on ImageNet with a great margin. For more experiments and the analysis of class-invariant feature, please see the Appendix.

5 Conclusion

In this work, we propose a novel dual adversarial training (DAT) framework for unsupervised domain adaptation. The DAT method introduces class-invariant features to adversarial domain adaptation for optimizing the adaptability to improve the system performance. This approach provides an alternative for the mainstream domain adaptation methods which focus on minimizing the divergence between the source and target domains. We demonstrate that the class-invariant features learned from the source domain can be beneficial for the classification on the target domain. The proposed method can yield state-of-the-art results on four unsupervised domain adaptation benchmarks, which illustrates the promise of our method.

REFERENCES

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in neural information processing systems*, pp. 343–351, 2016.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

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

Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 222–230, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pp. 513–520, 2007.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Shuang Li, Shi-ji Song, and Cheng Wu. Layer-wise domain correction for unsupervised domain adaptation. *Frontiers of Information Technology & Electronic Engineering*, 19(1):91–103, 2018.

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

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017.

Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3071–3085, 2018a.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018b.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pp. 213–226. Springer, 2010.

Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Yuan Wu and Yuhong Guo. Dual adversarial co-learning for multi-domain text classification. *arXiv preprint arXiv:1909.08203*, 2019.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. *arXiv preprint arXiv:2007.03141*, 2020.

Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2272–2281, 2017.

A EXPERIMENTS

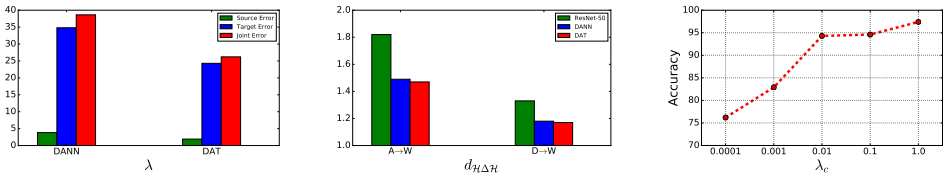
We also conduct experiments on the domain adaptation benchmark: Office-Home.

Office-Home (Venkateswara et al., 2017) is a more complicated dataset than Office-31, which consists of around 15500 images from 65 classes in office and home settings. There exists 4 domains in this dataset: Artistic Images (Ar), Clip Art (Cl), Product Images (Pr) and Real-World Images (Rw).

From Table 4, it can be noted that our proposed DAT method can yield the second best results. The improvement induced by DAT on Office-Home dataset is less than that on the other four benchmarks. An interpretation is that since Office-Home is with more classes compared with other benchmarks and difficult in each domain with much lower in-domain classification accuracy (Venkateswara et al., 2017), it's possible that the learned class-invariant features are contaminated by the task-related information in the presence of large number of classes. In summary, the proposed DAT approach works reasonably well on five domain adaptation benchmarks, highlighting the power of class-invariant feature in classification on both source and target domains.

Table 4: Accuracy (%) on Office-Home.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
ResNet-50 (He et al., 2016)	34.9	50.0	58.0	34.7	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN (Long et al., 2015)	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin & Lempitsky, 2015)	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN (Long et al., 2017)	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN (Long et al., 2018b)	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
DAT (Proposed)	47.8	65.9	73.4	48.3	62.7	64.2	48.7	46.9	74.5	68.2	53.4	80.7	61.2



(a) Adaptability (b) A-distance (c) λ_c

Figure 3: Analysis of adaptability, domain divergence and parameter sensitivity.

B Analysis

B.1 Adaptability

In this study, we investigate how the class-invariant feature influences the adaptability λ, which is measured by the expected error of the ideal joint hypothesis on both source and target domains. In order to compute λ, we train an multi-layer perceptron (MLP) classifier over the feature representations learned by DANN (Ganin & Lempitsky, 2015) and DAT on VisDA-2017. The MLP classifier is trained on all labeled data from both source and target domains. It should be noted that the target labels are only used for this analysis. When training the MLP classifier, the feature extractors in DANN and DAT should be fixed. The error of the ideal joint hypothesis on the source domain, the target domain, and their sum λ are shown in Figure 3(a). As expected, DAT has a lower λ than DANN, which suggests that the introduction of class-invariant feature in adversarial domain adaptation can effectively reduce the adaptability.

B.2 Distribution Discrepancy

As shown in the domain adaptation theory (Ben-David et al., 2010), the domain discrepancy $d_{\mathcal{H}\Delta\mathcal{H}}$ and adaptability λ are two important factors that bound the generalization error on the target domain. The A-distance (Ben-David et al., 2010) is a measure of domain discrepancy, defined as $d_A = 2(1 - 2\epsilon)$, where ϵ is the error rate of the domain discriminator trained to differ source features from target features. In this study, we investigate the A-distance of ResNet-50 (He et al., 2016), DANN (Ganin & Lempitsky, 2015) and DAT on two tasks of Office-31 dataset: A→W and D→W. The results are shown in Figure 3(b). It can be noted that the adversarial domain adaptation can effectively reduce the domain divergence. The A-distance with features of DAT is similar to that of DANN, this combined with the experimental results, reveals that our DAT approach focuses on optimizing the adaptability to improve the system performance.

B.3 Parameter Sensitivity Analysis

In this section, we discuss the sensitivity of our approach to the values of the hyperparameters λ_c. We evaluate the influence of λ_c on the Digits dataset, especially, the U→M task. λ_c is explored in the range $\{0.0001, 0.001, 0.01, 0.1, 1.0\}$. The results are shown in Figure 3(c). From Figure 3(c), we can see that the selection of λ_c has an influence for the classification accuracy. With the increase of λ_c, the accuracy increases dramatically and reaches its best at $\lambda_c = 1$. This illustrates that a properly selected λ_c can effectively boost the classification accuracy of the model.