The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Discriminative Adversarial Domain Adaptation

Hui Tang, Kui Jia*
South China University of Technology
eehuitang@mail.scut.edu.cn, kuijia@scut.edu.cn

Abstract

Given labeled instances on a source domain and unlabeled ones on a target domain, unsupervised domain adaptation aims to learn a task classifier that can well classify target instances. Recent advances rely on domain-adversarial training of deep networks to learn domain-invariant features. However, due to an issue of mode collapse induced by the separate design of task and domain classifiers, these methods are limited in aligning the joint distributions of feature and category across domains. To overcome it, we propose a novel adversarial learning method termed Discriminative Adversarial Domain Adaptation (DADA). Based on an integrated category and domain classifier, DADA has a novel adversarial objective that encourages a mutually inhibitory relation between category and domain predictions for any input instance. We show that under practical conditions, it defines a minimax game that can promote the joint distribution alignment. Except for the traditional closed set domain adaptation, we also extend DADA for extremely challenging problem settings of partial and open set domain adaptation. Experiments show the efficacy of our proposed methods and we achieve the new state of the art for all the three settings on benchmark datasets.

Introduction

Many machine learning tasks are advanced by large-scale learning of deep models, with image classification (Russakovsky et al. 2015) as one of the prominent examples. A key factor to achieve such advancements is the availability of massive labeled data on the domains of the tasks of interest. For many other tasks, however, training instances on the corresponding domains are either difficult to collect, or their labeling costs prohibitively. To address the scarcity of labeled data for these *target* tasks/domains, a general strategy is to leverage the massively available labeled data on related *source* ones via domain adaptation (Pan and Yang 2010). Even though the source and target tasks share the same label space (i.e. closed set domain adaptation), domain adaptation still suffers from the shift in data distributions. The main objective of domain adaptation is thus to learn domain-invariant features, so that task classifiers learned from the source data can be readily applied to the target domain. In this work, we focus on the unsupervised setting where training instances on the target domain are completely unlabeled.

Recent domain adaptation methods are largely built on modern deep architectures. They rely on great model capacities of these networks to learn hierarchical features that are empirically shown to be more transferable across domains (Yosinski et al. 2014; Zhang, Tang, and Jia 2018). Among them, those based on domain-adversarial training (Ganin et al. 2016; Wang et al. 2019) achieve the current state of the art. Based on the seminal work of DANN (Ganin et al. 2016), they typically augment a classification network with an additional domain classifier. The domain classifier takes features from the feature extractor of the classification network as inputs, which is trained to differentiate between instances from the two domains. By playing a minimax game (Goodfellow et al. 2014), adversarial training aims to learn domain-invariant features.

Such domain-adversarial networks can largely reduce the domain discrepancy. However, the separate design of task and domain classifiers has the following shortcomings. Firstly, feature distributions can only be aligned to a certain level, since model capacity of the feature extractor could be large enough to compensate for the less aligned feature distributions. More importantly, given practical difficulties of aligning the source and target distributions with high granularity to the category level (especially for complex distributions with multi-mode structures), the task classifier obtained by minimizing the empirical source risk cannot well generalize to the target data due to an issue of mode collapse (Kurmi and Namboodiri 2019; Tran et al. 2019), i.e., the joint distributions of feature and category are not well aligned across the source and target domains.

Recent methods (Kurmi and Namboodiri 2019; Tran et al. 2019) take the first step to address the above shortcomings by jointly parameterizing the task and domain classifiers into an integrated one. To further push this line, based on such a classifier, we propose a novel adversarial learning method termed *Discriminative Adversarial Domain Adaptation (DADA)*, which encourages a *mutually inhibitory* relation between its domain prediction and category prediction for any input instance, as illustrated in Figure 1. This dis-

*Corresponding author.

criminative interaction between category and domain predictions underlies the ability of DADA to reduce domain discrepancy at both the feature and category levels. Intuitively, the adversarial training of DADA mainly conducts competition between the domain neuron (output) and the true category neuron (output). Different from the work (Tran et al. 2019) whose mechanism to align the joint distributions is rather implicit, DADA enables explicit alignment between the joint distributions, thus improving the classification of target data. Except for closed set domain adaptation, we also extend DADA for partial domain adaptation (Cao et al. 2018b), i.e. the target label space is subsumed by the source one, and open set domain adaptation (Saito et al. 2018c), i.e. the source label space is subsumed by the target one. Our main contributions can be summarized as follows.

- We propose in this work a novel adversarial learning method, termed DADA, for closed set domain adaptation. Based on an integrated category and domain classifier, DADA has a novel adversarial objective that encourages a *mutually inhibitory* relation between category and domain predictions for any input instance, which can promote the joint distribution alignment across domains.
- For more realistic partial domain adaptation, we extend DADA by a reliable category-level weighting mechanism, termed DADA-P, which can significantly reduce the negative influence of outlier source instances.
- For more challenging open set domain adaptation, we extend DADA by balancing the joint distribution alignment in the shared label space with the classification of outlier target instances, termed DADA-O.
- Experiments show the efficacy of our proposed methods and we achieve the new state of the art for all the three adaptation settings on benchmark datasets.

Related Works

Closed Set Domain Adaptation After the seminal work of DANN (Ganin et al. 2016), ADDA (Tzeng et al. 2017) proposes an untied weight sharing strategy to align the target feature distribution to a fixed source one. SimNet (Pinheiro 2018) replaces the standard FC-based cross-entropy classifier by a similarity-based one. MADA (Pei et al. 2018) and CDAN (Long et al. 2018b) integrate the discriminative category information into domain-adversarial training. VADA (Shu et al. 2018) reduces the cluster assumption violation to constrain domain-adversarial training. Some methods (Wang et al. 2019; Wen et al. 2019) focus on transferable regions to learn domain-invariant features and task classifier. TAT (Liu et al. 2019) enhances the discriminability of features to guarantee the adaptability. Some methods (Saito et al. 2018b; 2018a; Lee et al. 2019) utilize category predictions from two task classifiers to measure the domain discrepancy. The most related works (Kurmi and Namboodiri 2019; Tran et al. 2019) to us propose joint parameterization of the task and domain classifiers, which implicitly align the joint distributions. Differently, our proposed DADA makes the joint distribution alignment more explicit, thus promoting classification on the target domain.

Partial Domain Adaptation The work (Zhang et al. 2018) weights each source instance by its importance to the target domain based on one domain classifier, and then trains another domain classifier on target and weighted source instances. The works (Cao et al. 2018a; 2018b) reduce the contribution of outlier source instances to the task or domain classifiers by utilizing category predictions. Differently, DADA-P weights the proposed source discriminative adversarial loss by a reliable category confidence.

Open Set Domain Adaptation Previous research (Jain, Scheirer, and Boult 2014) proposes to reject an instance as the unknown category by threshold filtering. The work (Saito et al. 2018c) proposes to utilize adversarial training for both domain adaptation and unknown outlier detection. Differently, DADA-O balances the joint distribution alignment in the shared label space with the outlier rejection.

Method

Given $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ of labeled instances sampled from the source domain $\mathcal{D}_s$, and $\{\mathbf{x}_j^t\}_{j=1}^{n_t}$ of unlabeled instances sampled from the target domain $\mathcal{D}_t$, the objective of unsupervised domain adaptation is to learn a feature extractor $G(\cdot)$ and a task classifier $C(\cdot)$ such that the expected target risk $\mathbb{E}_{(\mathbf{x}^t, y^t)\sim\mathcal{D}_t}[\mathcal{L}_{cls}(C(G(\mathbf{x}^t)), y^t)]$ is low for a certain classification loss function $\mathcal{L}_{cls}(\cdot)$. The domains $\mathcal{D}_s$ and $\mathcal{D}_t$ are assumed to have different distributions. To achieve a low target risk, a typical strategy is to learn $G(\cdot)$ and $C(\cdot)$ by minimizing the sum of the source risk and some notion of *distance* between the source and target domain distributions, inspired by domain adaptation theories (Ben-David et al. 2007; 2010). This strategy is based on a simple rational that the source risk would become a good indicator of the target risk when the distance between the two distributions is getting closer. While most of existing methods use distance measures based on the marginal distributions, it is arguably better to use those based on the joint distributions.

The above strategy is generally implemented by domain-adversarial learning (Ganin et al. 2016; Wang et al. 2019), where separate task classifier $C(\cdot)$ and domain classifier $D(\cdot)$ are typically stacked on top of the feature extractor $G(\cdot)$. As discussed before, this type of design has the following shortcomings: (1) model capacity of $G(\cdot)$ could be large enough to make $D(G(\mathbf{x}^s))$ and $D(G(\mathbf{x}^t))$ hardly differentiable for any instance, even though the marginal feature distributions are not well aligned; (2) more importantly, it is difficult to align the source and target distributions with high granularity to the category level (especially for complex distributions with multi-mode structures), and thus $C(\cdot)$ obtained by minimizing the empirical source risk cannot perfectly generalize to the target data due to an issue of mode collapse, i.e. the joint distributions are not well aligned.

To alleviate the above shortcomings, inspired by semi-supervised learning methods based on GANs (Salimans et al. 2016; Dai et al. 2017), the recent work (Tran et al. 2019) proposes joint parameterization of $C(\cdot)$ and $D(\cdot)$ into an integrated one $F(\cdot)$. Suppose the classification task of interest has K categories, $F(\cdot)$ is formed simply by augmenting the last FC layer of $C(\cdot)$ with one additional neuron.

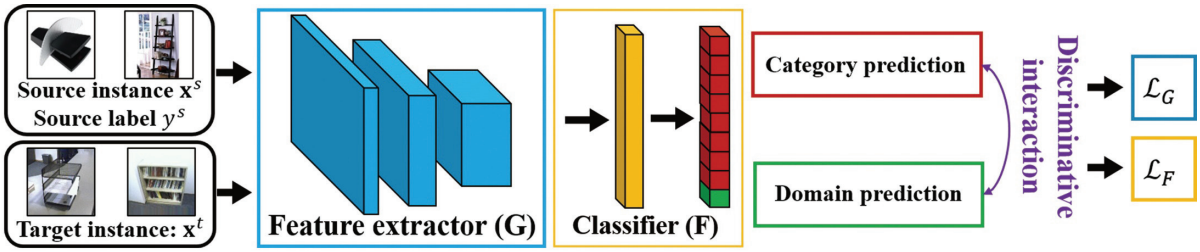

Figure 1: (Best viewed in color.) Discriminative Adversarial Domain Adaptation (DADA), which includes a feature extractor $G(\cdot)$ and an integrated category and domain classifier $F(\cdot)$. The blue and orange colors denote $G(\cdot)$ and $F(\cdot)$, and the losses applied to them, respectively. Note that DADA explicitly establishes a discriminative interaction between category and domain predictions. Please refer to the main text for how the adversarial training objective of DADA is defined.

Denote $\mathbf{p}(\mathbf{x}) \in [0,1]^{K+1}$ as the output vector of class probabilities of $F(G(\mathbf{x}))$ for an instance $\mathbf{x}$, and $p_k(\mathbf{x})$, $k \in \{1,\ldots,K+1\}$, as its k^{th} element. The k^{th} element of the conditional probability vector $\bar{\mathbf{p}}(\mathbf{x})$ is written as follows

$$\bar{p}_k(\mathbf{x}) = \begin{cases} \dfrac{p_k(\mathbf{x})}{1-p_{K+1}(\mathbf{x})}, & k = 1,2,...,K \\ 0 \quad , & k = K+1 \end{cases}. \tag{1}$$

For ease of subsequent notations, we also write $p^s_k = p_k(\mathbf{x}^s)$ and $p^t_k = p_k(\mathbf{x}^t)$. Then, such a network is trained by the classification-aware adversarial learning objective

$$\begin{aligned} &\min_F -\frac{1}{n_s}\sum_{i=1}^{n_s}\log p_{y^s_i}(\mathbf{x}^s_i) - \frac{1}{n_t}\sum_{j=1}^{n_t}\log p_{K+1}(\mathbf{x}^t_j) \\ &\max_G \frac{1}{n_s}\sum_{i=1}^{n_s}\log \bar{p}_{y^s_i}(\mathbf{x}^s_i) + \lambda\frac{1}{n_t}\sum_{j=1}^{n_t}\log(1-p_{K+1}(\mathbf{x}^t_j)), \end{aligned} \tag{2}$$

where λ balances category classification and domain adversarial losses. The mechanism of this objective to align the joint distributions across domains is rather implicit.

To make it more explicit, based on the integrated classifier $F(\cdot)$, we propose a novel adversarial learning method termed *Discriminative Adversarial Domain Adaptation (DADA)*, which explicitly enables a discriminative interplay of predictions among the domain and K categories for any input instance, as illustrated in Figure 1. This discriminative interaction underlies the ability of DADA to promote the joint distribution alignment, as explained shortly.

Discriminative Adversarial Learning

To establish a direct interaction between category and domain predictions, we propose a novel source discriminative adversarial loss that is tailored to the design of the integrated classifier $F(\cdot)$. The proposed loss is inspired by the principle of binary cross-entropy loss. It is written as

$$\begin{aligned} \mathcal{L}^s(G,F) = -\frac{1}{n_s}\sum_{i=1}^{n_s}[&(1-p_{K+1}(\mathbf{x}^s_i))\log p_{y^s_i}(\mathbf{x}^s_i) \\ &+ p_{K+1}(\mathbf{x}^s_i)\log\left(1-p_{y^s_i}(\mathbf{x}^s_i)\right)]. \end{aligned} \tag{3}$$

Intuitively, the proposed loss (3) establishes a mutually inhibitory relation between $p_{y^s}(\mathbf{x}^s)$ of the prediction on the true category of $\mathbf{x}^s$, and $p_{K+1}(\mathbf{x}^s)$ of the prediction on the domain of $\mathbf{x}^s$. We first discuss how the proposed loss (3) works during adversarial training, and we show that under practical conditions, minimizing (3) over the classifier $F(\cdot)$ has the effects of discriminating among task categories while distinguishing the source domain from the target one, and maximizing (3) over the feature extractor $G(\cdot)$ can discriminatively align the source domain to the target one.

Discussion We first write the gradient formulas of $\mathcal{L}^s$ on any source instance $\mathbf{x}^s$ w.r.t. $p^s_{y^s}$ and p^s_{K+1} as

$$\nabla_{p^s_{y^s}} = \frac{\partial\mathcal{L}^s}{\partial p^s_{y^s}} = \frac{p^s_{y^s}p^s_{K+1} - (1-p^s_{y^s})(1-p^s_{K+1})}{p^s_{y^s}(1-p^s_{y^s})},$$

$$\nabla_{p^s_{K+1}} = \frac{\partial\mathcal{L}^s}{\partial p^s_{K+1}} = \log\frac{p^s_{y^s}}{1-p^s_{y^s}}.$$

Since both $p^s_{y^s}$ and p^s_{K+1} are among the $K+1$ output probabilities of the classifier $F(G(\mathbf{x}^s))$, we always have $p^s_{y^s} \le 1-p^s_{K+1}$ and $p^s_{K+1} \le 1-p^s_{y^s}$, suggesting $\nabla_{p^s_{y^s}} \le 0$. When the loss (3) is minimized over $F(\cdot)$ via stochastic gradient descent (SGD), we have the update $p^s_{y^s} \leftarrow p^s_{y^s} - \eta\nabla_{p^s_{y^s}}$ where η is the learning rate, and since $\nabla_{p^s_{y^s}} \le 0$, $p^s_{y^s}$ increases; when it is maximized over $G(\cdot)$ via stochastic gradient ascent (SGA), we have the update $p^s_{y^s} \leftarrow p^s_{y^s} + \eta\nabla_{p^s_{y^s}}$, and since $\nabla_{p^s_{y^s}} \le 0$, $p^s_{y^s}$ decreases. Then, we discuss the change of p^s_{K+1} in two cases: (1) in case of $p^s_{y^s} > 0.5$ that guarantees $\nabla_{p^s_{K+1}} > 0$, when minimizing the loss (3) over $F(\cdot)$ by SGD update $p^s_{K+1} \leftarrow p^s_{K+1} - \eta\nabla_{p^s_{K+1}}$, we have decreased p^s_{K+1}, and when maximizing it over $G(\cdot)$ by SGA update $p^s_{K+1} \leftarrow p^s_{K+1} + \eta\nabla_{p^s_{K+1}}$, we have increased p^s_{K+1}; (2) in case of $p^s_{y^s} < 0.5$ that guarantees $\nabla_{p^s_{K+1}} < 0$, when minimizing the loss (3) over $F(\cdot)$ by SGD update, we have increased p^s_{K+1}, and when maximizing it over $G(\cdot)$ by SGA update, we have decreased p^s_{K+1}, as shown in Figure 2.

For discriminative adversarial domain adaptation, we expect that (1) when minimizing the proposed loss (3) over $F(\cdot)$, task categories of the source domain is discriminative and the source domain is distinctive from the target one, which can be achieved when $p^s_{y^s}$ increases and p^s_{K+1} decreases; (2) when maximizing it over $G(\cdot)$, the source domain is aligned to the target one while retains discriminability, which can be achieved when $p^s_{y^s}$ decreases and p^s_{K+1}

Cases	$min_F \ \mathcal{L}^s$		$max_G \ \mathcal{L}^s$	
	$p^s_{y^s}$	p^s_{K+1}	$p^s_{y^s}$	p^s_{K+1}
$p^s_{y^s} > 0.5$	↑	↓	↓	↑
$p^s_{y^s} < 0.5$	↑	↑	↓	↓

Figure 2: Changes of $p^s_{y^s}$ and p^s_{K+1} when minimizing and maximizing the loss (3) in the two cases.

increases in the case of $p^s_{y^s} > 0.5$. To meet the expectations, the condition of $p^s_{y^s} > 0.5$ for all source instances should be always satisfied. This is practically achieved by pre-training DADA on the labeled source data using a K-way cross-entropy loss, and maintaining in the adversarial training of DADA the same supervision signal. We present in the supplemental material empirical evidence on benchmark datasets that shows the efficacy of our used scheme.

To achieve the joint distribution alignment, the explicit interplay between category and domain predictions for any target instance should also be created. Motivated by recent works (Pei et al. 2018; Long et al. 2018b) which alleviate the issue of mode collapse by aligning each instance to several most related categories, we propose a target discriminative adversarial loss based on the design of the integrated classifier $F(\cdot)$, by using the conditional category probabilities to weight the domain predictions. It is written as

$$\begin{aligned}\mathcal{L}^t_F(G,F) &= -\frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K}\bar{p}_k(\mathbf{x}^t_j)\log\hat{p}^k_{K+1}(\mathbf{x}^t_j)\\ \mathcal{L}^t_G(G,F) &= \frac{1}{n_t}\sum_{j=1}^{n_t}\sum_{k=1}^{K}\bar{p}_k(\mathbf{x}^t_j)\log(1-\hat{p}^k_{K+1}(\mathbf{x}^t_j)),\end{aligned} \tag{4}$$

where the k'^{th} element of the domain prediction vector $\hat{\mathbf{p}}^k$ for the k^{th} category is written as follows

$$\hat{p}^k_{k'}(\mathbf{x}) = \begin{cases}\dfrac{p_{k'}(\mathbf{x})}{p_k(\mathbf{x})+p_{K+1}(\mathbf{x})}, & k'=k,K+1\\ 0 \quad , & \text{otherwise}\end{cases}. \tag{5}$$

An intuitive explanation for our proposed (4) is provided in the supplemental material.

Established knowledge from cluster analysis (Nalewajski 2012) indicates that we can estimate clusters with a low probability of error only if the conditional entropy is small. To this end, we adopt the entropy minimization principle (Grandvalet and Bengio 2005), which is written as

$$\mathcal{L}^t_{em}(G,F) = \frac{1}{n_t}\sum_{j=1}^{n_t}\mathcal{H}(\bar{\mathbf{p}}(\mathbf{x}^t_j)), \tag{6}$$

where $\mathcal{H}(\cdot)$ computes the entropy of a probability vector. Combining (3), (4), and (6) gives the following minimax problem of our proposed DADA

$$\begin{aligned}\min_F \mathcal{L}_F &= \lambda(\mathcal{L}^s+\mathcal{L}^t_F)-\mathcal{L}^t_{em}\\ \max_G \mathcal{L}_G &= \lambda(\mathcal{L}^s+\mathcal{L}^t_G)-\mathcal{L}^t_{em},\end{aligned} \tag{7}$$

where λ is a hyper-parameter that trade-offs the adversarial domain adaptation objective with the entropy minimization one in the unified optimization problem. Note that in the minimization problem of (7), $\mathcal{L}^t_{em}$ serves as a regularizer for learning $F(\cdot)$ to avoid the trivial solution (i.e. all instances are assigned to the same category), and in the maximization problem of (7), it helps learn more target-discriminative features, which can alleviate the negative effect of adversarial feature adaptation on the adaptability (Liu et al. 2019).

By optimizing (7), the joint distribution alignment can be enhanced. This ability comes from the better use of discriminative information from both the source and target domains. Concretely, DADA constrains the domain classifier so that it clearly/explicitly knows the classification boundary, thus reducing false alignment between different categories. By deceiving such a strong domain classifier, DADA can learn a feature extractor that better aligns the two domains. *We also theoretically prove in the supplemental material that DADA can better bound the expected target error.*

Extension for Partial Domain Adaptation

Partial domain adaptation is a more realistic setting, where the target label space is subsumed by the source one. The false alignment between the outlier source categories and the target domain is unavoidable. To address it, existing methods (Cao et al. 2018a; Zhang et al. 2018; Cao et al. 2018b) utilize the category or domain predictions, to decrease the contribution of source outliers to the training of task or domain classifiers. Inspired by these ideas, we extend DADA for partial domain adaptation by using a reliable category-level weighting mechanism, which is termed DADA-P.

Concretely, we average the conditional probability vectors $\bar{\mathbf{p}}(\mathbf{x}^t) \in [0,1]^K$ over all target data and then normalize the averaged vector $\bar{\mathbf{c}} \in [0,1]^K$ by dividing its largest element. The category weight vector $\mathbf{c} \in [0,1]^K$ with c_k as its k^{th} element is derived by a convex combination of the normalized vector and an all-ones vector $\mathbf{1}$, as follows

$$\begin{aligned}\bar{\mathbf{c}} &= \frac{1}{n_t}\sum_{j=1}^{n_t}\bar{\mathbf{p}}(\mathbf{x}^t_j)\\ \mathbf{c} &= \lambda\frac{\bar{\mathbf{c}}}{\max(\bar{\mathbf{c}})}+(1-\lambda)\mathbf{1},\end{aligned} \tag{8}$$

where $\lambda \in [0,1]$ is to suppress the detection noise of outlier source categories in the early stage of training. Then, we apply the category weight vector $\mathbf{c}$ to the proposed discriminative adversarial loss for any source instance, leading to

$$\begin{aligned}\mathcal{L}^s(G,F) = -\frac{1}{n_s}\sum_{i=1}^{n_s} c_{y^s_i}[&(1-p_{K+1}(\mathbf{x}^s_i))\log p_{y^s_i}(\mathbf{x}^s_i)\\ &+p_{K+1}(\mathbf{x}^s_i)\log\left(1-p_{y^s_i}(\mathbf{x}^s_i)\right)].\end{aligned} \tag{9}$$

Since predicted probabilities on the outlier source categories are more likely to increase when minimizing $-\mathcal{L}^t_{em}$ over $F(\cdot)$, which incurs negative transfer. To avoid it, we minimize $\mathcal{L}^t_{em}$ over $F(\cdot)$ and the objective of DADA-P is

$$\begin{aligned}\min_F \mathcal{L}_F &= \lambda(\mathcal{L}^s+\mathcal{L}^t_F)+\mathcal{L}^t_{em}\\ \max_G \mathcal{L}_G &= \lambda(\mathcal{L}^s+\mathcal{L}^t_G)-\mathcal{L}^t_{em}.\end{aligned} \tag{10}$$

By optimizing it, DADA-P can simultaneously alleviate negative transfer and promote the joint distribution alignment across domains in the shared label space.

Extension for Open Set Domain Adaptation

Open set domain adaptation is a very challenging setting, where the source label space is subsumed by the target one. We denominate the shared category and all unshared categories between the two domains as the "known category" and "unknown category" respectively. The goal of open set domain adaptation is to correctly classify any target instance as the known or unknown category. The false alignment between the known and unknown categories is inevitable. To this end, the work (Saito et al. 2018c) proposes to make a pseudo decision boundary for the unknown category, which enables the feature extractor to reject some target instances as outliers. Inspired by this work, we extend DADA for open set domain adaptation by training the classifier to classify all target instances as the unknown category with a small probability q, which is termed DADA-O. Assuming the predicted probability on the unknown category as the K^{th} element of $\mathbf{p}(\mathbf{x}^t)$, i.e., $p_K(\mathbf{x}^t)$, the modified target adversarial loss when minimized over the integrated classifier $F(\cdot)$ is

$$\mathcal{L}_F^t(G,F) = \\ -\frac{1}{n_t}\sum_{j=1}^{n_t} q\log p_K(\mathbf{x}_j^t) - (1-q)\log p_{K+1}(\mathbf{x}_j^t), \tag{11}$$

where $0 < q < 0.5$. When maximized over the feature extractor $G(\cdot)$, we still use the discriminative loss $\mathcal{L}_G^t$ in (4). Replacing $\mathcal{L}_F^t$ in (7) with (11) gives the overall adversarial objective of DADA-O, which can achieve a balance between domain adaptation and outlier rejection.

We utilize all target instances to obtain the concept of "unknown", which is very helpful for the classification of unknown target instances as the unknown category but can cause the misclassification of known target instances as the unknown category. This issue can be alleviated by selecting an appropriate q. If q is too small, the unknown target instances cannot be correctly classified; if q is too large, the known target instances can be misclassified. By choosing an appropriate q, the feature extractor can separate the unknown target instances from the known ones while aligning the joint distributions in the shared label space.

Experiments

Datasets and Implementation Details

Office-31 (Saenko et al. 2010) is a popular benchmark domain adaptation dataset consisting of $4,110$ images of 31 categories collected from three domains: Amazon (**A**), Webcam (**W**), and DSLR (**D**). We evaluate on six settings.

Syn2Real (Peng et al. 2018) is the largest benchmark. Syn2Real-C has over $280K$ images of 12 shared categories in the combined training, validation, and testing domains. The $152,397$ images on the training domain are synthetic ones by rendering 3D models. The validation and test domains comprise real images, and the validation one has $55,388$ images. We use the training domain as the source domain and validation one as the target domain. For partial domain adaptation, we choose images of the first 6 categories (in alphabetical order) in the validation domain as the target domain and form the setting: **Synthetic 12** $\rightarrow$ **Real 6**. For open set domain adaptation, we evaluate on Syn2Real-O, which includes two domains. The training/synthetic domain uses synthetic images from the 12 categories of Syn2Real-C as "known". The validation/real domain uses images of the 12 categories from the validation domain of Syn2Real-C as "known", and 50k images from 69 other categories as "unknown". We use the training and validation domains of Syn2Real-O as the source and target domains respectively.

Implementation Details We follow standard evaluation protocols for unsupervised domain adaptation (Ganin et al. 2016; Wang et al. 2019): we use all labeled source and all unlabeled target instances as the training data. For all tasks of Office-31 and **Synthetic 12** $\rightarrow$ **Real 6**, based on ResNet-50 (He et al. 2016), we report the classification result on the target domain of mean($\pm$standard deviation) over three random trials. For other tasks of Syn2Real, we evaluate the accuracy of each category based on ResNet-101 and ResNet-152 (for closed and open set domain adaptation respectively). For each base network, we use all its layers up to the second last one as the feature extractor $G(\cdot)$, and set the neuron number of its last FC layer as $K+1$ to have the integrated classifier $F(\cdot)$. Exceptionally, we follow the work (Peng et al. 2018) and replace the last FC layer of ResNet-152 with three FC layers of 512 neurons. All base networks are pre-trained on ImageNet (Russakovsky et al. 2015). We firstly pre-train them on the labeled source data, and then fine-tune them on both the labeled source data and unlabeled target data via adversarial training, where we maintain the same supervision signal as the pre-training.

We follow DANN (Ganin et al. 2016) to use the SGD training schedule: the learning rate is adjusted by $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p denotes the process of training iterations that is normalized to be in $[0,1]$, and we set $\eta_0 = 0.0001$, $\alpha = 10$, and $\beta = 0.75$; the hyper-parameter λ is initialized at 0 and is gradually increased to 1 by $\lambda_p = \frac{2}{1+\exp(-\gamma p)} - 1$, where we set $\gamma = 10$. We empirically set $q = 0.1$. We implement all our methods by **PyTorch**. The code will be available at https://github.com/huitangtang/DADA-AAAI2020.

Analysis

Ablation Study We conduct ablation studies on Office-31 to investigate the effects of key components of our proposed DADA based on ResNet-50. Our ablation studies start with the very baseline termed "No Adaptation" that simply finetunes a ResNet-50 on the source data. To validate the mutually inhibitory relation enabled by DADA, we use DANN (Ganin et al. 2016) and DANN-CA (Tran et al. 2019) respectively as the second and third baselines. To investigate how the entropy minimization principle helps learn more target-discriminative features, we remove the entropy minimization loss (6) from our main minimax problem (7), denoted as "DADA (w/o em)". To know effects of the proposed source and target discriminative adversarial losses (3) and (4), we

Table 1: Ablation studies using Office-31 based on ResNet-50. Please refer to the main text for how they are defined.

Methods	A → W	D → W	W → D	A → D	D → A	W → A	Avg
No Adaptation	79.9±0.3	96.8±0.4	99.5±0.1	84.1±0.4	64.5±0.3	66.4±0.4	81.9
DANN	81.2±0.3	98.0±0.2	99.8±0.0	83.3±0.3	66.8±0.3	66.1±0.3	82.5
DANN-CA	85.4±0.4	98.2±0.2	99.8±0.0	87.1±0.4	68.5±0.2	67.6±0.3	84.4
DADA (w/o em + w/o td)	91.0±0.2	98.7±0.1	**100.0**±0.0	90.8±0.2	70.9±0.3	70.2±0.3	86.9
DADA (w/o em)	91.8±0.1	99.0±0.1	**100.0**±0.0	92.5±0.3	72.8±0.2	72.3±0.3	88.1
DADA	**92.3**±0.1	**99.2**±0.1	**100.0**±0.0	**93.9**±0.2	**74.4**±0.1	**74.2**±0.1	**89.0**

Table 2: Results for closed set domain adaptation on Office-31 based on ResNet-50. Note that SimNet is implemented by an **unknown** framework; MADA and DANN-CA are implemented by **Caffe**; all the other methods are implemented by **PyTorch**.

Methods	A → W	D → W	W → D	A → D	D → A	W → A	Avg
No Adaptation (He et al. 2016)	79.9±0.3	96.8±0.4	99.5±0.1	84.1±0.4	64.5±0.3	66.4±0.4	81.9
DAN (Long et al. 2018a)	81.3±0.3	97.2±0.0	99.8±0.0	83.1±0.2	66.3±0.0	66.3±0.1	82.3
DANN (Ganin et al. 2016)	81.2±0.3	98.0±0.2	99.8±0.0	83.3±0.3	66.8±0.3	66.1±0.3	82.5
ADDA (Tzeng et al. 2017)	86.2±0.5	96.2±0.3	98.4±0.3	77.8±0.3	69.5±0.4	68.9±0.5	82.9
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
VADA (Shu et al. 2018)	86.5±0.5	98.2±0.4	99.7±0.2	86.7±0.4	70.1±0.4	70.5±0.4	85.4
DANN-CA (Tran et al. 2019)	91.35	98.24	99.48	89.94	69.63	68.76	86.2
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
MCD (Saito et al. 2018b)	88.6±0.2	98.5±0.1	**100.0**±0.0	92.2±0.2	69.5±0.1	69.7±0.3	86.5
CDAN+E (Long et al. 2018b)	**94.1**±0.1	98.6±0.1	**100.0**±0.0	92.9±0.2	71.0±0.3	69.3±0.3	87.7
TADA (Wang et al. 2019)	94.3±0.3	98.7±0.1	99.8±0.2	91.6±0.3	72.9±0.2	73.0±0.3	88.4
SymNets (Zhang et al. 2019)	90.8±0.1	98.8±0.3	**100.0**±0.0	**93.9**±0.5	**74.6**±0.6	72.5±0.5	88.4
TAT (Liu et al. 2019)	92.5±0.3	**99.3**±0.1	**100.0**±0.0	93.2±0.2	73.1±0.3	72.1±0.3	88.4
DADA	92.3±0.1	99.2±0.1	**100.0**±0.0	**93.9**±0.2	74.4±0.1	**74.2**±0.1	**89.0**

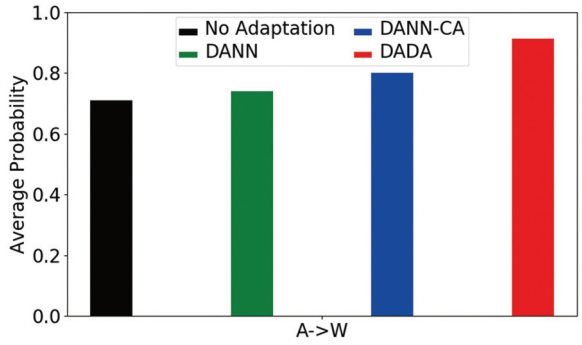

Figure 3: Average probability on the true category over all target instances by task classifiers of different methods.

remove both (6) and (4) from (7), denoted as "DADA (w/o em + w/o td)".

Results in Table 1 show that although DANN improves over "No Adaptation", its result is much worse than DANN-CA, verifying the efficacy of the design of the integrated classifier $F(\cdot)$. "DADA (w/o em + w/o td)" improves over DANN-CA and "DADA (w/o em)" improves over "DADA (w/o em + w/o td)", showing the efficacy of our proposed discriminative adversarial learning. DADA significantly outperforms DANN and DANN-CA, confirming the efficacy of the proposed mutually inhibitory relation between the category and domain predictions in aligning the joint distributions of feature and category across domains. Table 1 also confirms that entropy minimization is helpful to learn more target-discriminative features.

Quantitative Comparison To compare the efficacy of different methods in reducing domain discrepancy at the category level, we visualize the average probability on the true category over all target instances by task classifiers of No Adaptation, DANN, DANN-CA, and DADA on **A** → **W** in Figure 3. Note that here we use labels of the target data for the quantization of category-level domain discrepancy. Figure 3 shows that our proposed DADA gives the predicted probability on the true category of any target instance a better chance to approach 1, meaning that target instances are more likely to be correctly classified by DADA, i.e., a better category-level domain alignment.

Results

Closed Set Domain Adaptation We compare in Tables 2 and 3 our proposed method with existing ones on Office-31 and Syn2Real-C based on ResNet-50 and ResNet-101 respectively. Whenever available, results of existing methods are quoted from their respective papers or the recent works (Pei et al. 2018; Long et al. 2018b; Liu et al. 2019; Saito et al. 2018b). Our proposed DADA outperforms existing methods, testifying the efficacy of DADA in aligning the joint distributions of feature and category across domains.

Partial Domain Adaptation We compare in Table 5 our proposed method to existing ones on Syn2Real-C based on ResNet-50. Results of existing methods are quoted from the work (Cao et al. 2018b). Our proposed DADA-P substantially outperforms all comparative methods by $+15.53\%$, showing the effectiveness of DADA-P on reducing the neg-

Table 3: Results for closed set domain adaptation on Syn2Real-C based on ResNet-101. Note that all compared methods are based on **PyTorch** implementation.

Methods	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	mean
No Adaptation (He et al. 2016)	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DANN (Ganin et al. 2016)	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
DAN (Long et al. 2018a)	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
MCD (Saito et al. 2018b)	87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	**76.9**	88.6	40.3	83.0	25.8	71.9
GPDA (Kim et al. 2019)	83.0	74.3	80.4	**66.0**	87.6	75.3	83.8	73.1	**90.1**	57.3	80.2	37.9	73.3
ADR (Saito et al. 2018a)	87.8	**79.5**	**83.7**	65.3	**92.3**	61.8	**88.9**	73.2	87.8	60.0	85.5	32.3	74.8
DADA	**92.9**	74.2	82.5	65.0	90.9	**93.8**	87.2	74.2	89.9	**71.5**	**86.5**	**48.7**	**79.8**

Table 4: Results for open set domain adaptation on Syn2Real-O based on ResNet-152. *Known* indicates the mean classification result over the known categories whereas *Mean* also includes the unknown category. The table below shows the results when the Known-to-Unknown Ratio in the target domain is set to $1:10$. All compared methods are based on **PyTorch** implementation.

Known-to-Unknown Ratio $= 1:1$															
Methods	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	unk	Known	Mean
No Adaptation (He et al. 2016)	49	20	29	47	62	27	79	3	37	19	70	1	62	36	38
DAN (Long et al. 2018a)	51	40	42	56	68	24	75	2	39	30	71	2	75	41	44
DANN (Ganin et al. 2016)	59	41	16	54	77	18	88	4	44	32	68	4	61	42	43
AODA (Saito et al. 2018c)	85	71	65	53	**83**	10	79	36	**73**	56	79	**32**	**87**	60	62
DADA-O	**88**	**76**	**76**	**64**	79	**46**	**91**	**62**	52	**63**	**86**	8	55	**66**	**65**
Known-to-Unknown Ratio $= 1:10$															
AODA (Saito et al. 2018c)	**80**	**63**	59	63	**83**	12	89	5	**61**	14	79	0	**69**	51	52
DADA-O	77	**63**	**75**	**71**	38	**33**	**92**	**58**	47	**50**	**89**	**1**	50	**58**	**57**

Table 5: Results for partial domain adaptation on Syn2Real-C based on ResNet-50. Note that all compared methods are based on **PyTorch** implementation.

Methods	Synthetic 12→Real 6
No Adaptation (He et al. 2016)	45.26
DAN (Long et al. 2018a)	47.60
DANN (Ganin et al. 2016)	51.01
RTN (Long et al. 2016)	50.04
PADA (Cao et al. 2018b)	53.53
DADA-P	**69.06**

ative influence of source outliers while promoting the joint distribution alignment in the shared label space.

Open Set Domain Adaptation We compare in Table 4 our proposed method with existing ones on Syn2Real-O based on ResNet-152. Results of existing methods are quoted from the recent work (Peng et al. 2018). Our proposed DADA-O outperforms all comparative methods in both evaluation metrics of Known and Mean, showing the efficacy of DADA-O in both aligning joint distributions of the known instances and identifying the unknown target instances. It is noteworthy that DADA-O improves over the state-of-the-art method AODA by a large margin when the known-to-unknown ratio in the target domain is much smaller than 1, i.e. the false alignment between the known source and unknown target instances will be much more serious. This observation confirms the efficacy of DADA-O.

We provide more results and analysis for the three problem settings in the supplemental material.

Conclusion

We propose a novel adversarial learning method termed Discriminative Adversarial Domain Adaptation (DADA) to overcome the limitation in aligning the joint distributions of feature and category across domains, which is due to an issue of mode collapse induced by the separate design of task and domain classifiers. Based on an integrated task and domain classifier, DADA has a novel adversarial objective that encourages a mutually inhibitory relation between the category and domain predictions, which can promote the joint distribution alignment. Unlike previous methods, DADA explicitly enables a discriminative interaction between category and domain predictions. Except for closed set domain adaptation, we also extend DADA for more challenging problem settings of partial and open set domain adaptation. Experiments on benchmark datasets testify the efficacy of our proposed methods for all the three settings.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No.: 61771201), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), and the Guangdong R&D key project of China (Grant No.: 2019B010155001).

References

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In Schölkopf, B.; Platt, J. C.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems*. MIT Press. 137–144.

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* 79(1):151–175.

Cao, Z.; Long, M.; Wang, J.; and Jordan, M. I. 2018a. Partial transfer learning with selective adversarial networks. In *Computer Vision and Pattern Recognition*.

Cao, Z.; Ma, L.; Long, M.; and Wang, J. 2018b. Partial adversarial domain adaptation. In *European Conference on Computer Vision*.

Dai, Z.; Yang, Z.; Yang, F.; Cohen, W. W.; and Salakhutdinov, R. R. 2017. Good semi-supervised learning that requires a bad gan. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 6510–6520.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17(1):2096–2030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 2672–2680.

Grandvalet, Y., and Bengio, Y. 2005. Semi-supervised learning by entropy minimization. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. MIT Press. 529–536.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*.

Jain, L. P.; Scheirer, W. J.; and Boult, T. E. 2014. Multi-class open set recognition using probability of inclusion. In *European Conference on Computer Vision*.

Kim, M.; Sahu, P.; Gholami, B.; and Pavlovic, V. 2019. Unsupervised visual domain adaptation: A deep max-margin gaussian process approach. In *Computer Vision and Pattern Recognition*.

Kurmi, V. K., and Namboodiri, V. P. 2019. Looking back at labels: A class based domain adaptation technique. *ArXiv* abs/1904.01341.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*.

Liu, H.; Long, M.; Wang, J.; and Jordan, M. 2019. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*. Curran Associates Inc.

Long, M.; Cao, Y.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018a. Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1.

Long, M.; CAO, Z.; Wang, J.; and Jordan, M. I. 2018b. Conditional adversarial domain adaptation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc. 1640–1650.

Nalewajski, R. F. 2012. *Elements of Information Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg. 371–395.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22:1345–1359.

Pei, Z.; Cao, Z.; Long, M.; and Wang, J. 2018. Multi-adversarial domain adaptation. In *Association for the Advancement of Artificial Intelligence*.

Peng, X.; Usman, B.; Saito, K.; Kaushik, N.; Hoffman, J.; and Saenko, K. 2018. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *ArXiv* abs/1806.09755.

Pinheiro, P. O. 2018. Unsupervised domain adaptation with similarity learning. In *Computer Vision and Pattern Recognition*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *European Conference on Computer Vision*.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2018a. Adversarial dropout regularization. In *International Conference on Learning Representations*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018b. Maximum classifier discrepancy for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition*.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018c. Open set domain adaptation by backpropagation. In *European Conference on Computer Vision*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; and Chen, X. 2016. Improved techniques for training gans. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc. 2234–2242.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Computer Vision and Pattern Recognition*.

Shu, R.; Bui, H.; Narui, H.; and Ermon, S. 2018. A DIRT-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations*.

Tran, L.; Sohn, K.; Yu, X.; Liu, X.; and Chandraker, M. K. 2019. Gotta adapt 'em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In *Computer Vision and Pattern Recognition*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition*.

Wang, X.; Li, L.; Ye, W.; Long, M.; and Wang, J. 2019. Transferable attention for domain adaptation. In *Association for the Advancement of Artificial Intelligence*.

Wen, J.; Liu, R.; Zheng, N.; Zheng, Q.; Gong, Z.; and Yuan, J. 2019. Exploiting local feature patterns for unsupervised domain adaptation. In *Association for the Advancement of Artificial Intelligence*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 3320–3328.

Zhang, J.; Ding, Z.; Li, W.; and Ogunbona, P. 2018. Importance weighted adversarial nets for partial domain adaptation. In *Computer Vision and Pattern Recognition*.

Zhang, Y.; Tang, H.; Jia, K.; and Tan, M. 2019. Domain-symmetric networks for adversarial domain adaptation. In *Computer Vision and Pattern Recognition*.

Zhang, Y.; Tang, H.; and Jia, K. 2018. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *The European Conference on Computer Vision*.