Multi-Adversarial Domain Adaptation*

Zhongyi Pei[†], Zhangjie Cao[†], Mingsheng Long, and Jianmin Wang
KLiss, MOE; NEL-BDS; TNList; School of Software, Tsinghua University, China
{peizhyi,caozhangjie14}@gmail.com {mingsheng,jimwang}@tsinghua.edu.cn

Abstract

Recent advances in deep domain adaptation reveal that adversarial learning can be embedded into deep networks to learn transferable features that reduce distribution discrepancy between the source and target domains. Existing domain adversarial adaptation methods based on single domain discriminator only align the source and target data distributions without exploiting the complex multimode structures. In this paper, we present a multi-adversarial domain adaptation (MADA) approach, which captures multimode structures to enable fine-grained alignment of different data distributions based on multiple domain discriminators. The adaptation can be achieved by stochastic gradient descent with the gradients computed by back-propagation in linear-time. Empirical evidence demonstrates that the proposed model outperforms state of the art methods on standard domain adaptation datasets.

Introduction

Deep networks, when trained on large-scale labeled datasets, can learn transferable representations which are generically useful across diverse tasks and application domains (Donahue et al. 2014; Yosinski et al. 2014). However, due to a phenomenon known as dataset bias or domain shift (Torralba and Efros 2011), predictive models trained with these deep representations on one large dataset do not generalize well to novel datasets and tasks. The typical solution is to further fine-tune these networks on task-specific datasets, however, it is often prohibitively expensive to collect enough labeled data to properly fine-tune the high-capacity deep networks. Hence, there is strong motivation to establishing effective algorithms to reduce the labeling consumption by leveraging readily-available labeled data from a different but related source domain. This promising transfer learning paradigm, however, suffers from the shift in data distributions across different domains, which poses a major obstacle in adapting classification models to target tasks (Pan and Yang 2010).

Existing transfer learning methods assume shared label space and different feature distributions across the source and target domains. These methods bridge different domains by learning domain-invariant feature representations without using target labels, and the classifier learned from source domain can be directly applied to target domain. Recent studies have revealed that deep neural networks can learn more transferable features for domain adaptation (Donahue et al. 2014; Yosinski et al. 2014), by disentangling explanatory factors of variations behind domains. The latest advances have been achieved by embedding domain adaptation modules in the pipeline of deep feature learning to extract domain-invariant representations (Tzeng et al. 2014; Long et al. 2015; Ganin and Lempitsky 2015; Tzeng et al. 2015; Long et al. 2016; Bousmalis et al. 2016; Long et al. 2017).

Recently, adversarial learning has been successfully embedded into deep networks to learn transferable features to reduce distribution discrepancy between the source and target domains. Domain adversarial adaptation methods (Ganin and Lempitsky 2015; Tzeng et al. 2015) are among the top-performing deep architectures. These methods mainly align the whole source and target distributions, without considering the complex multimode structures underlying the data distributions. As a result, not only all data from the source and target domains will be confused, but also the discriminative structures could be mixed up, leading to false alignment of the corresponding discriminative structures of different distributions, with intuitive example shown in Figure 1. Hence, matching the whole source and target domains as previous methods without exploiting the discriminative structures may not work well for diverse domain adaptation scenarios.

There are two technical challenges to enabling domain adaptation: (1) enhancing *positive* transfer by maximally matching the multimode structures underlying data distributions across domains, and (2) alleviating *negative* transfer by preventing false alignment of modes in different distributions across domains. Motivated by these challenges, we present a multi-adversarial domain adaptation (MADA) approach, which captures multimode structures to enable fine-grained alignment of different data distributions based on multiple domain discriminators. A key improvement over previous methods is the capability to simultaneously promote positive transfer of relevant data and alleviate negative transfer of irrelevant data. The adaptation can be achieved by stochastic gradient descent with the gradients computed by back-propagation in linear-time. Empirical evidence demonstrates that the proposed MADA approach outperforms state of the art methods on standard domain adaptation benchmarks.

*Corresponding author: Mingsheng Long

[†]Equal contribution

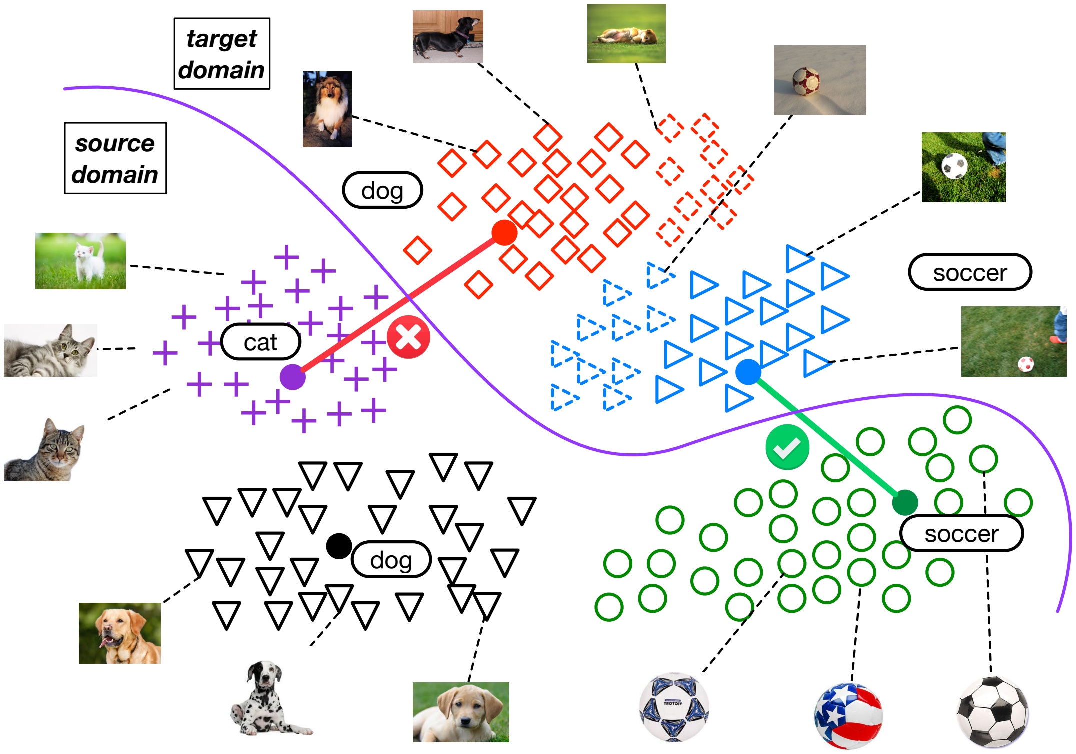


Figure 1: The difficulty of domain adaptation: discriminative structures may be mixed up or falsely aligned across domains. As an intuitive example, in this figure, the source class *cat* is falsely aligned with target class *dog*, making final classification wrong.

Related Work

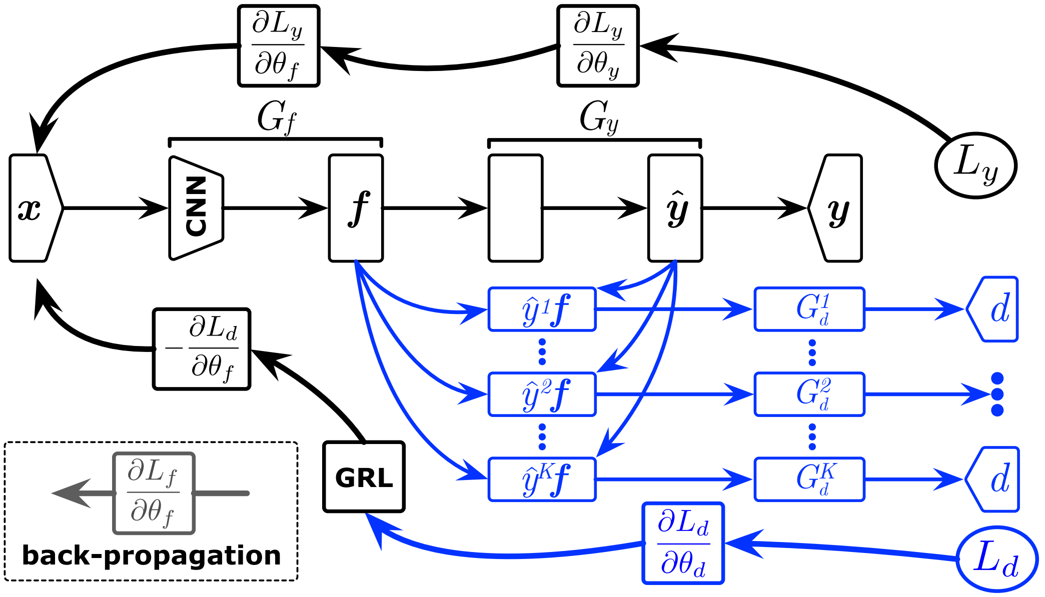
Transfer learning (Pan and Yang 2010) bridges different domains or tasks to mitigate the burden of manual labeling for machine learning (Pan et al. 2011; Duan, Tsang, and Xu 2012; Zhang et al. 2013; Wang and Schneider 2014), computer vision (Saenko et al. 2010; Gong et al. 2012; Hoffman et al. 2014) and natural language processing (Collobert et al. 2011). The main technical difficulty of transfer learning is to formally reduce the distribution discrepancy across different domains. Deep networks can learn abstract representations that disentangle different explanatory factors of variations behind data (Bengio, Courville, and Vincent 2013) and manifest invariant factors underlying different populations that transfer well from original tasks to similar novel tasks (Yosinski et al. 2014). Thus deep networks have been explored for transfer learning (Glorot, Bordes, and Bengio 2011; Oquab et al. 2013; Hoffman et al. 2014), multimodal and multi-task learning (Collobert et al. 2011; Ngiam et al. 2011), where significant performance gains have been witnessed against prior shallow transfer learning methods.

However, recent advances show that deep networks can learn abstract feature representations that can only reduce, but not remove, the cross-domain discrepancy (Glorot, Bordes, and Bengio 2011; Tzeng et al. 2014), resulting in unbounded risk for target tasks (Mansour, Mohri, and Rostamizadeh 2009; Ben-David et al. 2010). Some recent work bridges deep learning and domain adaptation (Tzeng et al. 2014; Long et al. 2015; Ganin and Lempitsky 2015; Tzeng et al. 2015; Long et al. 2016; Bousmalis et al. 2016; Long et al. 2017), which extends deep convolutional networks (CNNs) to domain adaptation by adding adaptation layers through which the mean embeddings of distributions are matched (Tzeng et al. 2014; Long et al. 2015; 2016), or by adding a subnetwork as domain discriminator while the deep features are learned to confuse the discriminator in a domain-adversarial training paradigm (Ganin and Lempitsky 2015; Tzeng et al. 2015). While performance was significantly improved, these state of the art methods may be restricted by the fact that the discriminative structures as well as complex multimode structures are not exploited for fine-grained alignment of different distributions.

Adversarial learning has been explored for generative modeling in Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). Recently, several difficulties of GANs have been addressed, e.g. ease training (Arjovsky, Chintala, and Bottou 2017; Arjovsky and Bottou 2017), avoid mode collapse (Mirza and Osindero 2014; Che et al. 2017; Metz et al. 2017). In particular, Generative Multi-Adversarial Network (GMAN) (Durugkar, Gemp, and Mahadevan 2017) extends GANs to multiple discriminators including formidable adversary and forgiving teacher, which significantly eases model training and enhances distribution matching.

Multi-Adversarial Domain Adaptation

In unsupervised domain adaptation, we are given a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ of n_s labeled examples and



$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
G_f
G_y
L_y
x
CNN
f
$\hat{y}$
y
$\hat{y}^1 f$
$\hat{y}^2 f$
$\hat{y}^K f$
G_d^1
G_d^2
G_d^K
d
d
$-\frac{\partial L_d}{\partial \theta_f}$
GRL
$\frac{\partial L_f}{\partial \theta_f}$
back-propagation
$\frac{\partial L_d}{\partial \theta_d}$
L_d

Figure 2: The architecture of the proposed Multi-Adversarial Domain Adaptation (MADA) approach, where $\mathbf{f}$ is the extracted deep features, $\hat{\mathbf{y}}$ is the predicted data label, and $\hat{\mathbf{d}}$ is the predicted domain label; G_f is the feature extractor, G_y and L_y are the label predictor and its loss, G_d^k and L_d^k are the domain discriminator and its loss; GRL stands for Gradient Reversal Layer. The blue part shows the multiple adversarial networks (each for a class, K in total) crafted in this paper. *Best viewed in color.*

a target domain $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ of n_t unlabeled examples. The source domain and target domain are sampled from joint distributions $P(\mathbf{X}^s, \mathbf{Y}^s)$ and $Q(\mathbf{X}^t, \mathbf{Y}^t)$ respectively, and note that $P \neq Q$. The goal of this paper is to design a deep neural network that enables learning of transfer features $\mathbf{f} = G_f(\mathbf{x})$ and adaptive classifier $y = G_y(\mathbf{f})$ to reduce the shifts in the joint distributions across domains, such that the target risk $\Pr_{(\mathbf{x},y)\sim q}[G_y(G_f(\mathbf{x})) \neq \mathbf{y}]$ minimized by jointly minimizing source risk and distribution discrepancy by multi-adversarial domain adaptation.

There are two technical challenges to enabling domain adaptation: **(1)** enhancing *positive* transfer by maximally matching the multimode structures underlying data distributions P and Q across domains, and **(2)** alleviating *negative* transfer by preventing false alignment of different distribution modes across domains. These two challenges motivate the multi-adversarial domain adaptation approach.

Domain Adversarial Network

Domain adversarial networks have been successfully applied to transfer learning (Ganin and Lempitsky 2015; Tzeng et al. 2015) by extracting transferable features that can reduce the distribution shift between the source domain and the target domain. The adversarial learning procedure is a two-player game, where the first player is the domain discriminator G_d trained to distinguish the source domain from the target domain, and the second player is the feature extractor G_f fine-tuned simultaneously to confuse the domain discriminator.

To extract domain-invariant features $\mathbf{f}$, the parameters θ_f of feature extractor G_f are learned by maximizing the loss of domain discriminator G_d, while the parameters θ_d of domain discriminator G_d are learned by minimizing the loss of the domain discriminator. In addition, the loss of label predictor G_y is also minimized. The objective of domain adversarial network (Ganin and Lempitsky 2015) is the functional:

$$\begin{aligned} C_0(\theta_f, \theta_y, \theta_d) = &\frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_s} L_y(G_y(G_f(\mathbf{x}_i)), y_i) \\ &- \frac{\lambda}{n} \sum_{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} L_d(G_d(G_f(\mathbf{x}_i)), d_i), \end{aligned} \tag{1}$$

where $n = n_s + n_t$ and λ is a trade-off parameter between the two objectives that shape the features during learning. After training convergence, the parameters $\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_d$ will deliver a saddle point of the functional (1):

$$\begin{aligned} (\hat{\theta}_f, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_y} C_0(\theta_f, \theta_y, \theta_d), \\ (\hat{\theta}_d) &= \arg\max_{\theta_d} C_0(\theta_f, \theta_y, \theta_d). \end{aligned} \tag{2}$$

Domain adversarial networks (Ganin and Lempitsky 2015; Tzeng et al. 2015) are the top-performing architectures for standard domain adaptation when the distributions of the source domain and target domain can be aligned successfully.

Multi-Adversarial Domain Adaptation

In practical domain adaptation problems, however, the data distributions of the source domain and target domain usually embody complex multimode structures, reflecting either the class boundaries in supervised learning or the cluster boundaries in unsupervised learning. Thus, previous domain adversarial adaptation methods that only match the data distributions without exploiting the multimode structures may be prone to either under transfer or negative transfer. Under transfer may happen when different modes of the distributions cannot be maximally matched. Negative transfer may happen when the corresponding modes of the distributions across domains are falsely aligned. To promote positive transfer and combat negative transfer, we should find a technology to reveal the multimode structures underlying distributions on which multi-adversarial domain adaptation can be performed.

To match the source and target domains upon the multimode structures underlying data distributions, we notice that the source domain labeled information provides strong signals to reveal the multimode structures. Therefore, we split the domain discriminator G_d in Equation (1) into K class-wise domain discriminators $G_d^k, k = 1, \ldots, K$, each is responsible for matching the source and target domain data associated with class k, as shown in Figure 2. Since target domain data are fully unlabeled, it is not easy to decide which domain discriminator G_d^k is responsible for each target data point. Fortunately, we observe that the output of the label predictor $\hat{\mathbf{y}}_i = G_y(\mathbf{x}_i)$ to each data point $\mathbf{x}_i$ is a probability distribution over the label space of K classes. This distribution well characterizes the probability of assigning $\mathbf{x}_i$ to each of the K classes. Thus, it is a natural idea to use $\hat{\mathbf{y}}_i$ as the probability to indicate how much each data point $\mathbf{x}_i$ should be attended to the K domain discriminators $G_d^k, k = 1, \ldots, K$. The attention of each point $\mathbf{x}_i$ to a domain discriminator G_d^k can be modeled by weighting its features $G_f(\mathbf{x}_i)$ with probability $\hat{y}_i^k$. Applying this to all K domain discriminators $G_d^k, k = 1, \ldots, K$ yields

$$L_d = \frac{1}{n} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}_s \cup \mathcal{D}_t} L_d^k \left(G_d^k \left(\hat{y}_i^k G_f \left(\mathbf{x}_i \right) \right), d_i \right), \tag{3}$$

where G_d^k is the k-th domain discriminator while L_d^k is its cross-entropy loss, and d_i is the domain label of point $\mathbf{x}_i$. We note that the above strategy shares similar ideas with the attention mechanism.

Compared with the previous single-discriminator domain adversarial network in Equation (1), the proposed multi-adversarial domain adaptation network enables fine-grained adaptation where each data point $\mathbf{x}_i$ is matched only by those relevant domain discriminators according to its probability $\hat{\mathbf{y}}_i$. This fine-grained adaptation may introduce three benefits. **(1)** It avoids the hard assignment of each point to only one domain discriminator, which tends to be inaccurate for target domain data. **(2)** It circumvents negative transfer since each point is only aligned to the most relevant classes, while the irrelevant classes are filtered out by the probability and will not be included in the corresponding domain discriminators, hence avoiding false alignment of the discriminative structures in different distributions. **(3)** The multiple domain discriminators are trained with probability-weighted data points $\hat{y}_i^k G_f(\mathbf{x}_i)$, which naturally learn multiple domain discriminators with different parameters θ_d^k; these domain discriminators with different parameters promote *positive transfer* for each instance.

Integrating all things together, the objective of the Multi-Adversarial Domain Adaptation (MADA) is

$$\begin{aligned} C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^K\right) &= \frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_s} L_y \left(G_y \left(G_f \left(\mathbf{x}_i \right) \right), y_i \right) \\ &- \frac{\lambda}{n} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{D}} L_d^k \left(G_d^k \left(\hat{y}_i^k G_f \left(\mathbf{x}_i \right) \right), d_i \right), \end{aligned} \tag{4}$$

where $n = n_s + n_t$, $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_t$ and λ is a hyper-parameter that trade-offs the two objectives in the unified optimization problem. The optimization problem is to find the parameters $\hat{\theta}_f$, $\hat{\theta}_y$ and $\hat{\theta}_d^k (k = 1, 2, ..., K)$ that jointly satisfy

$$\begin{aligned} (\hat{\theta}_f, \hat{\theta}_y) &= \arg \min_{\theta_f, \theta_y} C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^K\right), \\ (\hat{\theta}_d^1, ..., \hat{\theta}_d^K) &= \arg \max_{\theta_d^1, ..., \theta_d^K} C\left(\theta_f, \theta_y, \theta_d^k|_{k=1}^K\right). \end{aligned} \tag{5}$$

The multi-adversarial domain adaptation (MADA) model simultaneously enhances *positive* transfer by maximally matching the multimode structures underlying data distributions across domains, and circumvents *negative* transfer by avoiding false alignment of the distribution modes across domains.

Experiments

We evaluate the proposed multi-adversarial domain adaptation (MADA) model with state of the art transfer learning and deep learning methods. The codes, datasets and configurations will be available online at `github.com/thuml`.

Setup

Office-31 (Saenko et al. 2010) is a standard benchmark for visual domain adaptation, comprising 4,652 images and 31 categories collected from three distinct domains: *Amazon* (**A**), which contains images downloaded from `amazon.com`, *Webcam* (**W**) and *DSLR* (**D**), which contain images respectively taken by web camera and digital SLR camera with different environments. We evaluate all methods across three transfer tasks **A** → **W**, **D** → **W** and **W** → **D**, which are widely used by previous deep transfer learning methods (Tzeng et al. 2014; Ganin and Lempitsky 2015), and another three transfer tasks **A** → **D**, **D** → **A** and **W** → **A** as used in (Long et al. 2015; Tzeng et al. 2015; Long et al. 2016).

ImageCLEF-DA[1] is a benchmark dataset for ImageCLEF 2014 domain adaptation challenge, which is organized by selecting the 12 common categories shared by the following three public datasets, each is considered as a domain: *Caltech-256* (**C**), *ImageNet ILSVRC 2012* (**I**), and *Pascal VOC 2012* (**P**). The 12 common categories are aeroplane, bike, bird, boat, bottle, bus, car, dog, horse, monitor, motorbike, and people. There are 50 images in each category and 600 images in each domain. We use all domain combinations and build 6 transfer tasks: **I** → **P**, **P** → **I**, **I** → **C**, **C** → **I**, **C** → **P**, and **P** → **C**. Different from the *Office-31* dataset where different domains

[1] `http://imageclef.org/2014/adaptation`

Table 1: Accuracy (%) on *Office-31* for unsupervised domain adaptation (AlexNet and ResNet)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
AlexNet (Krizhevsky, Sutskever, and Hinton 2012)	60.6±0.4	95.4±0.2	99.0±0.1	64.2±0.3	45.5±0.5	48.3±0.5	68.8
TCA (Pan et al. 2011)	59.0±0.0	90.2±0.0	88.2±0.0	57.8±0.0	51.6±0.0	47.9±0.0	65.8
GFK (Gong et al. 2012)	58.4±0.0	93.6±0.0	91.0±0.0	58.6±0.0	52.4±0.0	46.1±0.0	66.7
DDC (Tzeng et al. 2014)	61.0±0.5	95.0±0.3	98.5±0.3	64.9±0.4	47.2±0.5	49.4±0.4	69.3
DAN (Long et al. 2015)	68.5±0.3	96.0±0.1	99.0±0.1	66.8±0.2	50.0±0.4	49.8±0.3	71.7
RTN (Long et al. 2016)	73.3±0.2	96.8±0.2	99.6±0.1	71.0±0.2	50.5±0.3	51.0±0.1	73.7
RevGrad (Ganin and Lempitsky 2015)	73.0±0.5	96.4±0.3	99.2±0.3	72.3±0.3	52.4±0.4	50.4±0.5	74.1
MADA	**78.5**±0.2	**99.8**±0.1	**100.0**±.0	**74.1**±0.1	**56.0**±0.2	**54.5**±0.3	**77.1**
ResNet (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
TCA (Pan et al. 2011)	74.7±0.0	96.7±0.0	99.6±0.0	76.1±0.0	63.7±0.0	62.9±0.0	79.3
GFK (Gong et al. 2012)	74.8±0.0	95.0±0.0	98.2±0.0	76.5±0.0	65.4±0.0	63.0±0.0	78.8
DDC (Tzeng et al. 2014)	75.8±0.2	95.0±0.2	98.2±0.1	77.5±0.3	67.4±0.4	64.0±0.5	79.7
DAN (Long et al. 2015)	83.8±0.4	96.8±0.2	99.5±0.1	78.4±0.2	66.7±0.3	62.7±0.2	81.3
RTN (Long et al. 2016)	84.5±0.2	96.8±0.1	99.4±0.1	77.5±0.3	66.2±0.2	64.8±0.3	81.6
RevGrad (Ganin and 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
MADA	**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**

are of different sizes, the three domains in this dataset are of equal size, making it a good alternative dataset.

We compare the proposed multi-adversarial domain adaptation (**MADA)** with both shallow and deep transfer learning methods: Transfer Component Analysis (**TCA**) (Pan et al. 2011), Geodesic Flow Kernel (**GFK**) (Gong et al. 2012), Deep Domain Confusion (**DDC**) (Tzeng et al. 2014), Deep Adaptation Network (**DAN**) (Long et al. 2015), Residual Transfer Network (**RTN**) (Long et al. 2016), and Reverse Gradient (**RevGrad**) (Ganin and Lempitsky 2015). TCA learns a shared feature space by Kernel PCA with linear-MMD penalty. GFK interpolates across an infinite number of intermediate subspaces to bridge the source and target subspaces. For these shallow transfer methods, we adopt SVM as the base classifier. DDC maximizes domain confusion by adding to deep networks a single adaptation layer that is regularized by linear-kernel MMD. DAN learns transferable features by embedding deep features of multiple domain-specific layers to reproducing kernel Hilbert spaces (RKHSs) and matching different distributions optimally using multi-kernel MMD. RTN jointly learns transferable features and adapts different source and target classifiers via deep residual learning (He et al. 2016). RevGrad enables domain adversarial learning (Goodfellow et al. 2014) by adapting a single layer of deep networks, which matches the source and target domains by making them indistinguishable for a domain discriminator.

We follow standard evaluation protocols for unsupervised domain adaptation (Long et al. 2015; Ganin and Lempitsky 2015). For both *Office-31* and *ImageCLEF-DA* datasets, we use all labeled source examples and all unlabeled target examples. We compare the average classification accuracy of each method on three random experiments, and report the standard error of the classification accuracies by different experiments of the same transfer task. For all baseline methods, we either follow their original model selection procedures, or conduct *transfer cross-validation* (Zhong et al. 2010) if their model selection strategies are not specified. We also adopt transfer cross-validation (Zhong et al. 2010) to select parameter λ for the MADA models. Fortunately, our models perform very stably under different parameter values, thus we fix $\lambda = 1$ throughout all experiments. For MMD-based methods (TCA, DDC, DAN, and RTN), we use Gaussian kernel with bandwidth set to the median pairwise squared distances on the training data, i.e. median trick (Gretton et al. 2012; Long et al. 2015). We examine the influence of deep representations for domain adaptation by exploring **AlexNet** (Krizhevsky, Sutskever, and Hinton 2012) and **ResNet** (He et al. 2016) as base architectures for learning deep representations. For shallow methods, we follow DeCAF (Donahue et al. 2014) and use as deep representations the activations of the $fc7$ (AlexNet) and $pool5$ (ResNet) layers.

We implement all deep methods based on the **Caffe** (Jia et al. 2014) framework, and fine-tune from AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and ResNet (He et al. 2016) models pre-trained on the ImageNet dataset (Russakovsky et al. 2014). We fine-tune all convolutional and pooling layers and train the classifier layer via back propagation. Since the classifier is trained from scratch, we set its learning rate to be 10 times that of the lower layers. We employ the mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and the learning rate strategy implemented in RevGrad (Ganin and Lempitsky 2015): the learning rate is not selected by a grid search due to high computational cost—it is adjusted during SGD using these formulas: $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p is the training progress linearly changing from 0 to 1, $\eta_0 = 0.01, \alpha = 10$ and $\beta = 0.75$, which is optimized to promote convergence and low error on source domain. To suppress noisy activations at the early stages of training, instead of fixing parameter λ, we gradually change it by multiplying $\frac{2}{1+\exp(-\delta p)} - 1$, where $\delta = 10$ (Ganin and Lempitsky 2015). This progressive training strategy significantly stabilizes parameter sensitivity of the proposed approach.

Results

The classification accuracy results on the *Office-31* dataset for unsupervised domain adaptation based on AlexNet and ResNet are shown in Table 1. For fair comparison, the re-

Table 2: Accuracy (%) on *ImageCLEF-DA* for unsupervised domain adaptation (AlexNet and ResNet)

Method	I → P	P → I	I → C	C → I	C → P	P → C	Avg
AlexNet (Krizhevsky, Sutskever, and Hinton 2012)	66.2±0.2	70.0±0.2	84.3±0.2	71.3±0.4	59.3±0.5	84.5±0.3	73.9
DAN (Long et al. 2015)	67.3±0.2	80.5±0.3	87.7±0.3	76.0±0.3	61.6±0.3	88.4±0.2	76.9
RTN (Long et al. 2016)	67.4±0.3	82.3±0.3	89.5±0.4	78.0±0.2	63.0±0.2	90.1±0.1	78.4
RevGrad (Ganin and Lempitsky 2015)	66.5±0.5	81.8±0.4	89.0±0.5	79.8±0.5	63.5±0.4	88.7±0.4	78.2
MADA	**68.3**±0.3	**83.0**±0.1	**91.0**±0.2	**80.7**±0.2	**63.8**±0.2	**92.2**±0.3	**79.8**
ResNet (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)	75.0±0.4	86.2±0.2	93.3±0.2	84.1±0.4	69.8±0.4	91.3±0.4	83.3
RTN (Long et al. 2016)	**75.6**±0.3	86.8±0.1	95.3±0.1	86.9±0.3	72.7±0.3	92.2±0.4	84.9
RevGrad (Ganin and 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
MADA	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**

Table 3: Accuracy (%) on *Office-31* for domain adaptation from 31 classes to 25 classes (AlexNet)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
AlexNet (Krizhevsky, Sutskever, and Hinton 2012)	58.2±0.4	95.9±0.2	99.0±0.1	60.4±0.3	49.8±0.5	47.3±0.5	68.4
RevGrad (Ganin and Lempitsky 2015)	65.1±0.5	91.7±0.3	97.1±0.3	60.6±0.3	42.1±0.4	42.9±0.5	66.6
MADA	**70.8**±0.2	**96.6**±0.1	**99.5**±.0	**69.6**±0.1	**51.4**±0.2	**54.2**±0.3	**73.7**

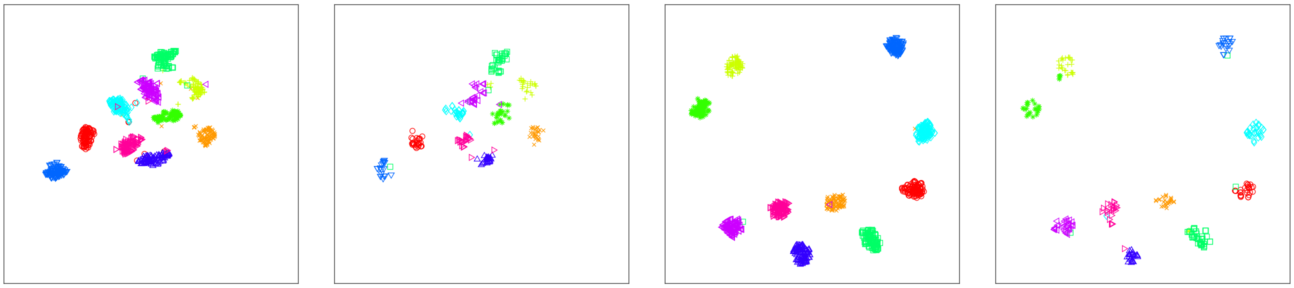
sults of DAN (Long et al. 2015), RTN (Long et al. 2016), and RevGrad (Ganin and Lempitsky 2015) are directly reported from their original papers. MADA outperforms all comparison methods on most transfer tasks. It is noteworthy that MADA promotes the classification accuracies substantially on hard transfer tasks, e.g. **A** → **W**, **A** → **D**, **D** → **A**, and **W** → **A**, where the source and target domains are substantially different, and produce comparable classification accuracies on easy transfer tasks, **D** → **W** and **W** → **D**, where the source and target domains are similar (Saenko et al. 2010). The three domains in the *ImageCLEF-DA* dataset are balanced in each category. As reported in Table 2, the MADA approach outperforms the comparison methods on most transfer tasks. The encouraging results highlight the importance of multi-adversarial domain adaptation in deep neural networks, and suggest that MADA is able to learn more transferable representations for effective domain adaptation.

The experimental results reveal several insightful observations. **(1)** Standard deep learning methods (AlexNet and ResNet) either outperform or underperform traditional shallow transfer learning methods (TCA and GFK) using deep features as input. This confirms the current practice that deep networks, even the extremely deep ones (ResNet), can learn abstract feature representations that only reduce but not remove the cross-domain discrepancy (Yosinski et al. 2014). **(2)** Deep transfer learning methods substantially outperform both standard deep learning methods and traditional shallow transfer learning methods with deep features as input. This validates that explicitly reducing the cross-domain discrepancy by embedding domain-adaptation modules into deep networks (DDC, DAN, RTN, and RevGrad) can learn more transferable features. **(3)** MADA substantially outperforms previous methods based on either multilayer adaptation (DAN), semi-supervised adaptation (RTN), and domain adversarial training (RevGrad). Although both MADA and RevGrad (Ganin and Lempitsky 2015) perform domain adversarial adaptation, the improvement from RevGrad to MADA is crucial for domain adaptation: RevGrad matches data distributions across domains without exploiting the complex multimode structures; MADA enables domain adaptation by making the source and target domains indistinguishable multiple domain discriminators, each responsible for matching the source and target data associated with the same class, which can essentially reduce the shift in the data distributions of complex multimode structures.

Negative transfer is an important technical bottleneck for successful domain adaptation. Negative transfer is more likely to happen when the source domain is substantially larger than the target domain, in which there exist many source data points that are irrelevant to the target domain. To evaluate the robustness against negative transfer, we randomly remove 6 classes from all transfer tasks constructed from the *Office-31* dataset. For example, we perform domain adaptation on transfer task **A 31** → **W 25**, where the source domain **A** has 31 classes but the target domain **W** has only 25 classes. In this more general and challenging scenario, we observe from Table 3 that the top-performing adversarial adaptation method, RevGrad, significantly underperforms standard AlexNet on most transfer tasks. This is an evidence of the negative transfer difficulty. The proposed MADA approach significantly exceeds the performance of both AlexNet and RevGrad, and successfully avoids the negative transfer trap. These positive results imply that the multi-adversarial adaptation can alleviate negative transfer.

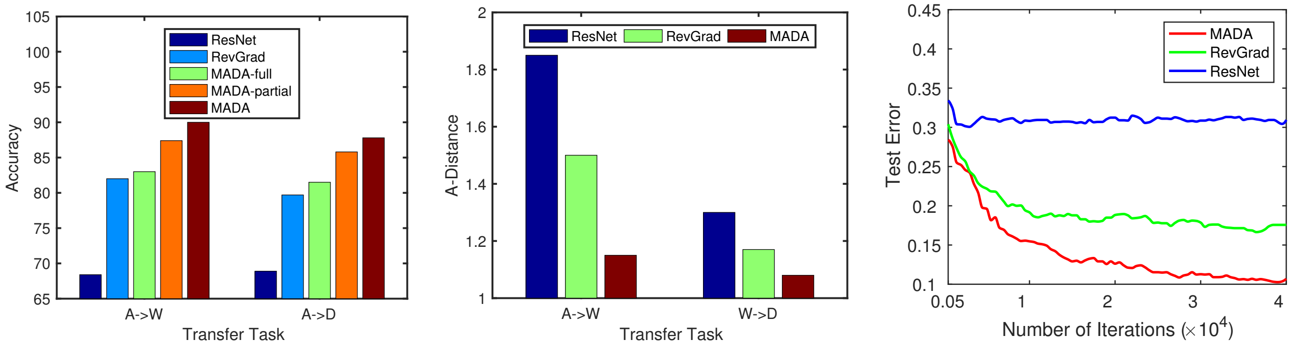
Analysis

Feature Visualization: We go deeper into the feature transferability by visualizing in Figures 3(a)–3(d) the network activations of task **A** → **W** (10 classes) learned by RevGrad (the bottleneck layer fcb) and MADA (the bottleneck layer fcb) respectively using t-SNE embeddings (Donahue et al. 2014). The visualization results reveal several interesting observations. **(1)** Under RevGrad features, the source and target domains are made indistinguishable; however, different cate-



(a) RevGrad: source=**A** (b) RevGrad: target=**W** (c) MADA: source=**A** (d) MADA: target=**W**

Figure 3: The t-SNE visualization of deep features extracted by RevGrad (a)(b) and MADA (c)(d).



(a) Sharing Strategies (b) $\mathcal{A}$-distance (c) Convergence

Figure 4: Empirical analysis: (a) Sharing strategies, (b) $\mathcal{A}$-distance, and (c) Convergence performance.

gories are not well discriminated clearly. The reason is that domain adversarial learning is performed only at the feature layer fcb, while the discriminative information is not taken into account by the domain adversary. **(2)** Under MADA features, not only the source and target domains are made more indistinguishable but also different categories are made more discriminated, which leads to the best adaptation accuracy. This superior results benefit from the integration of discriminative information into multiple domain discriminators, which enables matching of complex multimode structures of the source and target data distributions.

Sharing Strategies: Besides the proposed multi-adversarial strategy, one may consider other sharing strategies for multiple domain discriminators. For example, one can consider sharing all network parameters in the multiple domain discriminators, which is similar to previous domain adversarial adaptation methods with single domain discriminator; or consider sharing only a fraction of the network parameters for more flexibility. To examine different sharing strategies, we compare different variants of MADA: **MADA-full**, which shares all parameters of the multiple domain discriminator networks; **MADA-partial**, which shares only the lowest layers of the multiple discriminator networks. The accuracy results of tasks **A** $\rightarrow$ **W** and **A** $\rightarrow$ **D** in Figure 4(a) reveal that the transfer performance decreases when we share more parameters of multiple discriminators. This confirms our motivation that multiple domain discriminators are necessary to establish fine-grained distribution alignment.

Distribution Discrepancy: The domain adaptation theory (Ben-David et al. 2010; Mansour, Mohri, and Rostamizadeh 2009) suggests $\mathcal{A}$-distance as a measure of cross-domain discrepancy, which, together with the source risk, will bound the target risk. The proxy $\mathcal{A}$-distance is defined as $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier (e.g. kernel SVM) trained on the binary task of discriminating source and target. Figure 4(b) shows $d_{\mathcal{A}}$ on tasks **A** $\rightarrow$ **W**, **W** $\rightarrow$ **D** with features of ResNet, RevGrad, and MADA. We observe that $d_{\mathcal{A}}$ using MADA features is much smaller than $d_{\mathcal{A}}$ using ResNet and RevGrad features, which suggests that MADA features can reduce the cross-domain gap more effectively. As domains **W** and **D** are similar, $d_{\mathcal{A}}$ of task **W** $\rightarrow$ **D** is smaller than that of **A** $\rightarrow$ **W**, which well explains better accuracy of **W** $\rightarrow$ **D**.

Convergence Performance: Since MADA involves alternating optimization procedures, we testify the convergence performance with ResNet and RevGrad. Figure 4(c) demonstrates the test errors of different methods on task **A** $\rightarrow$ **W**, which suggests that MADA has similarly stable convergence performance as RevGrad while significantly outperforming RevGrad in the whole process of convergence. Also, the computational complexity of MADA is similar to RevGrad since the multiple domain discriminators only occupy a small fraction of the overall computational complexity.

Conclusion

This paper presented a novel multi-adversarial domain adaptation approach to enable effective deep transfer learning. Unlike previous domain adversarial adaptation methods that only match the feature distributions across domains without exploiting the complex multimode structures, the proposed approach further exploits the discriminative structures to enable fine-grained distribution alignment in a multi-adversarial adaptation framework, which can simultaneously promote positive transfer and circumvent negative transfer. Experiments show state of the art results of the proposed approach.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2016YFB1000701), National Natural Science Foundation of China (61772299, 61325008, 61502265, 61672313) and Tsinghua National Laboratory (TNList) Key Project.

References

Arjovsky, M., and Bottou, L. 2017. Towards principled methods for training generative adversarial networks. In *ICLR*.

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

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(1-2):151–175.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35(8):1798–1828.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *NIPS*, 343–351.

Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2017. Mode regularized generative adversarial networks. *ICLR*.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)* 12:2493–2537.

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.; Tsang, I. W.; and Xu, D. 2012. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34(3):465–479.

Durugkar, I.; Gemp, I.; and Mahadevan, S. 2017. Generative multi-adversarial networks. *ICLR*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

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*.

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.; Rasch, M.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)* 13:723–773.

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

Hoffman, J.; Guadarrama, S.; Tzeng, E.; Hu, R.; Donahue, J.; Girshick, R.; Darrell, T.; and Saenko, K. 2014. LSDA: Large scale detection through adaptation. In *NIPS*.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

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

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2016. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 136–144.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *ICML*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation: Learning bounds and algorithms. In *COLT*.

Metz, L.; Poole, B.; Pfau, D.; and Sohl-Dickstein, J. 2017. Unrolled generative adversarial networks. *ICLR*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2013. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 22(10):1345–1359.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks (TNN)* 22(2):199–210.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2014. ImageNet Large Scale Visual Recognition Challenge.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*.

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

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*.

Wang, X., and Schneider, J. 2014. Flexible transfer learning under support and model shift. In *NIPS*.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *NIPS*.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *ICML*.

Zhong, E.; Fan, W.; Yang, Q.; Verscheure, O.; and Ren, J. 2010. Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML/PKDD*, 547–562. Springer.