

Multi-source Distilling Domain Adaptation

Sicheng Zhao^{1*#}, Guangzhi Wang^{2#}, Shanghang Zhang^{1#}, Yang Gu², Yaxian Li²³,
Zhichao Song², Pengfei Xu², Runbo Hu², Hua Chai², Kurt Keutzer¹

¹University of California, Berkeley, USA ²Didi Chuxing, China ³Renmin University of China, China
schzhao@gmail.com, gzwang98@gmail.com, shzhang.pku@gmail.com, liyaxian@ruc.edu.cn
{guyangdavid, songzhichao, xupengfeipf, hurunbo, chaihua}@didiglobal.com, keutzer@berkeley.edu

Abstract

Deep neural networks suffer from performance decay when there is domain shift between the labeled source domain and unlabeled target domain, which motivates the research on domain adaptation (DA). Conventional DA methods usually assume that the labeled data is sampled from a single source distribution. However, in practice, labeled data may be collected from multiple sources, while naive application of the single-source DA algorithms may lead to suboptimal solutions. In this paper, we propose a novel multi-source distilling domain adaptation (MDDA) network, which not only considers the different distances among multiple sources and the target, but also investigates the different similarities of the source samples to the target ones. Specifically, the proposed MDDA includes four stages: (1) pre-train the source classifiers separately using the training data from each source; (2) adversarially map the target into the feature space of each source respectively by minimizing the empirical Wasserstein distance between source and target; (3) select the source training samples that are closer to the target to fine-tune the source classifiers; and (4) classify each encoded target feature by corresponding source classifier, and aggregate different predictions using respective domain weight, which corresponds to the discrepancy between each source and target. Extensive experiments are conducted on public DA benchmarks, and the results demonstrate that the proposed MDDA significantly outperforms the state-of-the-art approaches. Our source code is released at: <https://github.com/daoyuan98/MDDA>.

Introduction

One key element of the significant success of deep learning algorithms is the availability of large-scale labeled data (He et al. 2016). However, in many practical applications, only limited or even no training data is provided. On the one hand, it is prohibitively labor-intensive and expensive to obtain abundant labeled data. On the other hand, visual data possess variance in nature, which fundamentally limits the scalability and applicability of supervised learning models for handling new scenarios with few labeled examples (Ni, Zhang, and Xie 2019). In such cases, conven-

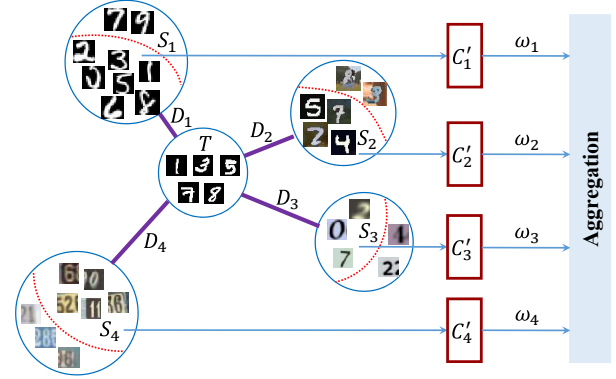


Figure 1: Illustration of MDDA which explores the relationships among different sources and the target. We employ a discriminator D to measure the similarity ω between each source and target in an adversarial manner. The samples that are closer to the target are selected to distill the source classifier C' . The prediction of different distilled source classifiers are aggregated based on the domain similarity to obtain the final prediction of the target samples.

tional deep learning approaches suffer from performance decay. Directly transferring the learned models trained on labeled source domains to unlabeled target domains may result in unsatisfying performance, because of the presence of domain shift (Torralla and Efros 2011), which calls for domain adaptation (DA) methods (Bousmalis et al. 2016; Zhao et al. 2018b; Hoffman et al. 2018). Unsupervised DA (UDA) addresses such problems by establishing knowledge transfer from a labeled source domain to an unlabeled target domain, and exploring domain-invariant structures and representations to bridge the domain gap (Netzer et al. 2011). Both theoretical results (Ben-David et al. 2010; Gopalan, Li, and Chellappa 2014; Tzeng et al. 2017) and algorithms for domain adaptation (Pan and Yang 2010; Long et al. 2015; Hoffman et al. 2018; Zhao et al. 2019b) have been proposed recently.

Though these methods make progress on DA, most of them focus on the single-source setting (Sun et al. 2011; Ganin et al. 2016) and fail to consider a more practical

*Corresponding Author. # Equal Contribution.

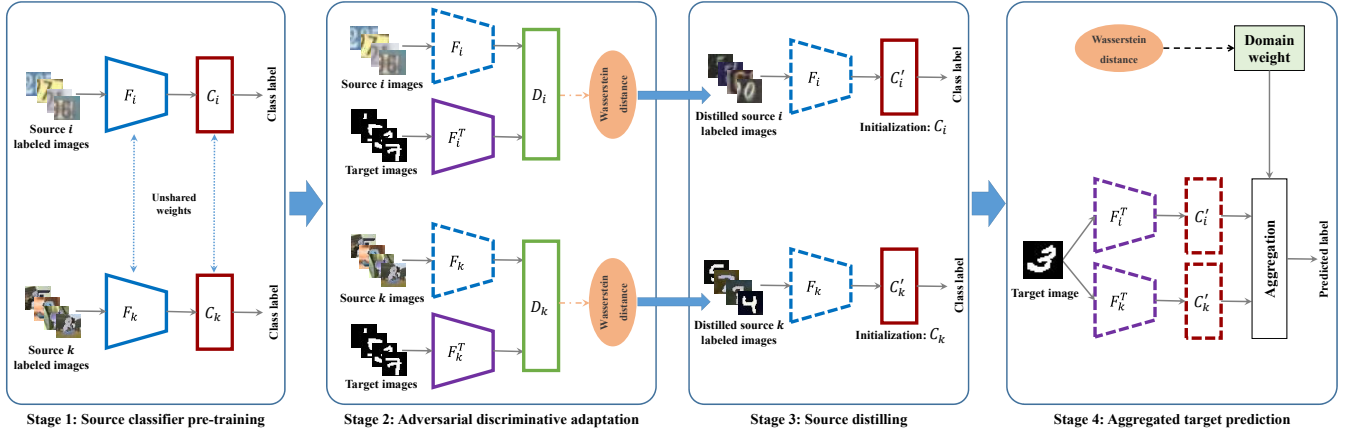


Figure 2: The framework of the proposed multi-source distilling domain adaptation (MDDA) network. Dashed rectangles and trapezoids indicate fixed network parameters. F , C , and D are short for feature extractor, classifier, and domain discriminator, respectively. For simplicity, we just consider the i th and k th source domains. The Proposed MDDA consists of four stages, as shown from left to right: source classifier pre-training, adversarial discriminative adaptation, source distilling, and aggregated target prediction. Best viewed in color.

scenario in which there are multiple labeled source domains with different distributions. Naive application of the single-source DA algorithms may lead to suboptimal solution (Shen et al. 2017), which calls for effective multi-source domain adaptation (MDA) techniques. Recently, some deep MDA approaches have been proposed (Zhao et al. 2018a; Xu et al. 2018; Li et al. 2018; Peng et al. 2019; Zhao et al. 2019a), but most of them suffer from the following limitations. (1) They sacrifice the discriminative property of the extracted features for the desired task learner in order to learn domain invariant features. (2) They treat the multiple sources equally and fail to consider the different discrepancy among sources and target, as illustrated in Figure 1. Such treatment may lead to suboptimal performance when some sources are very different from the target (Zhao et al. 2018a). (3) They treat different samples from each source equally, without distilling the source data based on the fact that different samples from the same source domain may have different similarities from the target. (4) The adversarial learning based methods suffer from vanishing gradient problem when the domain classifier network can perfectly distinguish target representations from the source ones.

In this paper, we propose a novel multi-source distilling domain adaptation (MDDA) network to address the above challenges by thoroughly exploring the relationships among different sources and the target. As shown in Figure 2, MDDA can be divided into four stages. (1) We first pre-train the source classifiers separately using the training data from each source. (2) We fix the feature extractor of each source and adversarially map the target into the feature space of each source respectively by minimizing the empirical Wasserstein distance between the source and target (Arjovsky, Chintala, and Bottou 2017), which provides more stable gradients even when the target and source distributions are non-overlap. (3) We select the source training samples that are closer to the target to fine-tune the source classifiers. (4) We build the target predictor by ag-

gregating the source predictions based on the source domain weights, which corresponds to the discrepancy between each source and target. We propose a mechanism to automatically choose a weighting strategy over source domains to emphasize more relevant sources and suppress the irrelevant ones, and aggregate multiple source classifiers based on these weights. With the above four stages, the proposed MDDA can extract features that are both discriminative for the learning task and indiscriminate with respect to the shift among the multiple source and target domains.

The main contributions of this paper are summarized as follows:

- We propose MDDA to explore the relationships among different sources and target, and achieve more accurate inference on the target by finetuning and aggregating the source classifiers based on these relationships.
- Compared to (Xu et al. 2018), which symmetrically maps the multiple sources and target into the same space, MDDA learns more discriminative target representations and avoids the oscillation from the simultaneous changing of the multi-source and target distributions by using separate feature extractors that asymmetrically map the target to the feature space of the source in an adversarial manner. Wasserstein distance is used in the adversarial training to achieve more stable gradients even when the target and source distributions are non-overlap.
- We propose the source distilling mechanism to select the source training samples that are closer to the target and fine-tune the source classifiers with these samples.
- We propose a novel mechanism to automatically choose a weighting strategy over source domains to emphasize more relevant sources and suppress the irrelevant ones, and aggregate the multiple source classifiers based on these weights to build more accurate target predictor.
- We extensively evaluate MDDA on the public benchmarks, achieving the state-of-the-art performance and verifying the efficacy of MDDA.

Related Work

Single-source UDA The emphasis of recent single-source UDA (SUDA) methods has shifted to deep learning architectures in an end-to-end fashion. Most deep SUDA methods employ a conjoined architecture with two streams to respectively represent the models for the source domain and the target domain (Zhuo et al. 2017). Generally, these methods are trained jointly with a traditional task loss based on the labeled source data and another loss to tackle the domain shift problem, such as **discrepancy loss, adversarial loss, reconstruction loss, etc.** Discrepancy-based methods explicitly measure the discrepancy between the source and target domains of the two network streams, such as the multiple kernel variant of maximum mean discrepancies (Long et al. 2015), correlation alignment (CORAL) (Sun, Feng, and Saenko 2016; 2017; Zhuo et al. 2017), and contrastive domain discrepancy (Kang et al. 2019). Adversarial generative models combine the domain discriminative model with a generative component to generate fake source or target data generally based on GAN (Goodfellow et al. 2014) and its variants, such as CoGAN (Liu and Tuzel 2016), SimGAN (Shrivastava et al. 2017), CycleGAN (Zhu et al. 2017; Zhao et al. 2019b), and CyCADA (Hoffman et al. 2018). Adversarial discriminative models usually employ an adversarial objective with respect to a domain discriminator to encourage domain confusion (Ganin et al. 2016; Tzeng et al. 2017; Chen et al. 2017; Shen et al. 2017; Tsai et al. 2018; Huang, Huang, and Krahenbuhl 2018). Most of these methods suffer from low accuracy when directly applied to the MDA problem.

Multi-source DA MDA assumes training data are collected from multiple sources (Sun, Shi, and Wu 2015; Zhao et al. 2019a). There are some theoretical analysis (Ben-David et al. 2010; Hoffman, Mohri, and Zhang 2018) to support existing MDA algorithms. The early MDA methods mainly focus on shallow models, including two categories (Sun, Shi, and Wu 2015): feature representation approaches (Sun et al. 2011; Duan, Xu, and Chang 2012; Chattopadhyay et al. 2012; Duan, Xu, and Tsang 2012) and combination of pre-learned classifiers (Xu and Sun 2012; Sun and Shi 2013). Some novel shallow MDA methods aim to deal with special cases, such as incomplete MDA (Ding, Shao, and Fu 2018) and target shift (Redko et al. 2019).

Recently, some representative deep learning based MDA methods are proposed, such as **multisource domain adversarial network (MDAN)** (Zhao et al. 2018a), **deep cocktail network (DCTN)** (Xu et al. 2018), and **moment matching network (MMN)** (Peng et al. 2019). All these MDA methods employ a shared feature extractor network to symmetrically map the multiple sources and target into the same space. For each source-target pair in MDAN and DCTN, a discriminator is trained to distinguish the source and target features. MDAN concatenates all extracted source features and labels into one domain to train a single task classifier, while DCTN trains a classifier for each source domain and combines the predictions of different classifiers for a target image using perplexity scores as weights. MMN transfers the learned knowledge from multiple sources to the target by dynamically aligning moments of their feature distributions. The fi-

nal prediction of a target image is averaged uniformly based on the classifiers from different source domains. Different from these works, we employ an unshared feature extractor to obtain the feature representation for each source, match the target feature to each source feature space asymmetrically, distill the pre-trained classifiers with selected representative samples, and combine the predictions of different classifiers using a novel weighting strategy.

Problem Definition

Suppose we have M source domains S_1, S_2, \dots, S_M and one target domain T . In unsupervised domain adaptation (UDA) scenario, S_1, S_2, \dots, S_M are labeled and T is fully unlabeled. For the i th source domain S_i , the observed images and corresponding labels drawn from the source distribution $p_i(\mathbf{x}, y)$ are $\mathbf{X}_i = \{\mathbf{x}_i^j\}_{j=1}^{N_i}$ and $Y_i = \{y_i^j\}_{j=1}^{N_i}$, where N_i is the number of source images. The target images drawn from the target distribution $p_T(\mathbf{x}, y)$ are $\mathbf{X}_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$ **without label observation**, where N_T is the number of target images. Unless otherwise specified, we assume (1) homogeneity, i.e. $\mathbf{x}_i^j \in \mathbb{R}^d, \mathbf{x}_T^j \in \mathbb{R}^d$, which indicates that the data from different domains are observed in the same feature space but exhibit different distributions; (2) closed set, i.e. $y_i^j \in \mathcal{Y}, y_T^j \in \mathcal{Y}$, where \mathcal{Y} is the class label space, indicating that all the domains share their categories. Our goal is to learn an adaptation model that can correctly predict a sample from the target domain based on $\{(\mathbf{X}_i, Y_i)\}_{i=1}^M$ and $\{\mathbf{X}_T\}$. Please note that our method can be easily extended to tackle **heterogeneous DA** (Li et al. 2014; Hubert Tsai, Yeh, and Frank Wang 2016) by changing the network structure of the target feature extractor, open set DA (Panareda Busto and Gall 2017) by adding an “unknown” class, or category shift DA (Xu et al. 2018) by reweighing the predictions of only those domains that contain the specified category. We will investigate such study in our future work.

Multi-source Distilling Domain Adaptation

In this section, we introduce the proposed multi-source distilling domain adaptation (MDDA) network. MDDA is a novel approach to overcome the limitations of existing methods for multiple source domain adaptation by thoroughly exploring the relationships among different sources and the target. It achieves more accurate inference on the target by finetuning and aggregating the source classifiers based these relationships. As shown in Figure 2, MDDA can be divided into four stages. We first pre-train the source classifiers separately with the training data from each source. Then, we fix the feature extractor of each source and map the target into the feature space of each source adversarially by minimizing the estimated Wasserstein distance between the source and target. MDDA learns more discriminative target representations and avoids the oscillation from the simultaneous changing of the multi-source and target distributions by using separate feature extractors that asymmetrically map the target to the feature space of the source in an adversarial manner. In the third stage, the source samples closer to

the target are selected to fine-tune the source classifiers. Finally, we build the target predictor by aggregating the source predictions based on the discrepancy between each source and target. We propose a novel mechanism to automatically choose a weighting strategy over source domains to emphasize more relevant sources and suppress the irrelevant ones. With the above four stages, MDDA extracts features that are both discriminative for the learning task and indiscriminate with respect to the shift among the multiple source and target domains. We will explain each stage in the following subsections.

Source Classifier Pre-training

To extract more task discriminative features and learn accurate classifiers, we pre-train a feature extractor F_i and classifier C_i for each labeled source domain S_i with unshared weights between different domains. Take the N -class classification task as an example, F_i and C_i are optimized by minimizing the following cross-entropy loss:

$$\mathcal{L}_{cls}(F_i, C_i) = -\mathbb{E}_{(\mathbf{x}_i, y_i) \sim p_i} \sum_{n=1}^N \mathbb{1}_{[n=y_i]} \log(\sigma(C_i(F_i(\mathbf{x}_i)))), \quad (1)$$

where σ is the softmax function, and $\mathbb{1}$ is an indicator function. Comparing with a shared feature extractor network to extract domain-invariant features among different source domains (Zhao et al. 2018a; Xu et al. 2018), the unshared feature extractor network can obtain the discriminative feature representations and accurate classifiers for each source domain. When aggregating the multiple predictions based on the source classifier and matched target features in the later stage, the final target prediction would be better boosted.

Adversarial Discriminative Adaptation

After the pre-training stage, we learn separate target encoder F_i^T to map the target feature into the same space of source S_i . A discriminator D_i is trained adversarially to maximize the Wasserstein distance of correctly classifying the encoded target features from F_i^T and the encoded source feature from pre-trained F_i , while F_i^T tries to maximize the probability of D_i making a mistake, *i.e.* minimizing the Wasserstein distance. Similar to GAN (Goodfellow et al. 2014), we model this as a two-player minimax game. Following (Arjovsky, Chintala, and Bottou 2017), **we suppose the discriminators $\{D_i\}$ are all 1-Lipschitz** and then we can optimize D_i by maximizing the Wasserstein distance

$$\mathcal{L}_{wd_D}(D_i) = \mathbb{E}_{\mathbf{x}_i \sim p_i} D_i(F_i(\mathbf{x}_i)) - \mathbb{E}_{\mathbf{x}_T \sim p_T} [D_i(F_i^T(\mathbf{x}_T))], \quad (2)$$

while F_i^T is obtained by minimizing

$$\mathcal{L}_{wd_F}(F_i^T) = -\mathbb{E}_{\mathbf{x}_T \sim p_T} D_i(F_i^T(\mathbf{x}_T)). \quad (3)$$

In this way, the target encoder F_i^T tries to confuse the discriminator D_i by minimizing the Wasserstein distance between the encoded target features as the source ones.

To enforce the Lipschitz constraint (Goodfellow et al. 2014), we add a gradient penalty for the parameters of each discriminator D_i as in (Gulrajani et al. 2017)

$$\mathcal{L}_{grad}(D_i) = (\|\nabla_{\hat{\mathbf{x}}} D_i(\hat{\mathbf{x}})\|_2 - 1)^2, \quad (4)$$

where $\hat{\mathbf{x}}$ is a feature set that contains not only the source and target features but also the random points along the straight line between source and target feature pairs (Gulrajani et al. 2017). D_i can then be optimized by

$$\max_{D_i} \mathcal{L}_{wd_D}(D_i) - \alpha \mathcal{L}_{grad}(D_i), \quad (5)$$

where α is a balancing coefficient, the value of which can be empirically set.

Source Distilling

We further dig into each source domain to select the source training samples that are closer to the target based on the estimated Wasserstein distance to fine-tune the source classifiers. Such source distilling mechanism utilizes more relevant training data and further improves the target performance on the aggregated source classifiers. We select the source samples based on the estimated Wasserstein distance, since it can represent the divergence between source data and target data. For each source sample x_i^j in the i th source domain, we calculate the Wasserstein distance between each source sample and target domain:

$$\tau_i^j = \|D_i(F_i(x_j)) - \frac{1}{N_T} \sum_{k=1}^{N_T} D_i(F_i^T(x_k))\|. \quad (6)$$

For each source sample x_i^j , τ_i^j reflects the its distance to the target domain. The smaller the τ_i^j value is, the closer it is to the target domain. Therefore, in each source domain X_i , we select $\frac{N_i}{2}$ of the source data $\hat{p}_i = \{\hat{x}_i^j, \hat{y}_i^j\}_{j=1}^{N_i}$ whose τ_i^j is larger than the left ones. With these selected source data, we finetune C_i by minimizing the following objective:

$$\mathcal{L}_{distill}(C_i) = -\mathbb{E}_{(\hat{\mathbf{x}}_i, \hat{y}_i) \sim \hat{p}_i} \sum_{n=1}^N \mathbb{1}_{[n=\hat{y}_i]} \log(\sigma(C_i(F_i(\hat{\mathbf{x}}_i)))), \quad (7)$$

Aggregated Target Prediction

In the testing stage, the goal is to accurately classify a given target image \mathbf{x}_T . Corresponding to each source domain, we extract the features $F_i^T(\mathbf{x}_T)$ of the target image based on the learned target encoder from stage 2, and obtain source-specific prediction $C'_i(F_i^T(\mathbf{x}_T))$ using the distilled source classifier. Next, we combine the different predictions from each source classifier to obtain the final prediction:

$$Result(\mathbf{x}_T) = \sum_{i=1}^N \omega_i C'_i(F_i^T(\mathbf{x}_T)). \quad (8)$$

The key problem here is how to select the weights ω_i for the predictions from different source classifiers. We design a novel weighting strategy based on the discrepancy between each source and target to emphasize more relevant sources and suppress the irrelevant ones. We assume after training in stage 2, the estimated Wasserstein distance $\mathcal{L}_{wd_{D_i}}$ between each source S_i and target T subordinates to a standard Gaussian Distribution $\mathcal{N}(0, 1)$. Therefore, the weight of each domain can be computed by the following equation

$$\omega_i = e^{\frac{-\mathcal{L}_{wd_{D_i}}^2}{2}}. \quad (9)$$

Table 1: Classification accuracy (%) on Digits-five dataset for multi-source unsupervised domain adaptation. The best method is emphasized in bold. Our method achieves 88.1% accuracy, significantly outperforming the state-of-the-art approaches.

Standards	Models	mm	mt	up	sv	sy	Avg
Source-only	Combined	63.7	92.3	87.2	66.3	84.8	78.9
	Single-best	59.2	97.2	84.7	77.7	85.2	80.8
Single-best DA	DAN (2015)	63.8	96.3	94.2	62.5	85.4	80.4
	CORAL (2016)	62.5	97.2	93.5	64.4	82.8	80.1
	DANN (2016)	71.3	97.6	92.3	63.5	85.3	82.0
	ADDA (2017)	71.6	97.9	92.8	75.5	86.5	84.9
Source-combined DA	DAN (2015)	67.9	97.5	93.5	67.8	86.9	82.7
	DANN (2016)	70.8	97.9	93.5	68.5	87.4	83.6
	ADDA (2017)	72.3	97.9	93.1	75.0	86.7	85.0
Multi-source DA	DCTN (2018)	70.5	96.2	92.8	77.6	86.8	84.8
	MDAN (2018a)	69.5	98.0	92.5	69.2	87.4	83.3
	MDDA (ours)	78.6	98.8	93.9	79.3	89.7	88.1

Experiments

We evaluate the proposed MDDA model on multi-source domain adaptation task in visual classification applications, including digit recognition and object classification.

Experimental Settings

Benchmarks Digits-five includes 5 digit image datasets sampled from different domains, including *handwritten mt* (MNIST) (LeCun et al. 1998), *combined mm* (MNIST-M) (Ganin and Lempitsky 2015), *street image sv* (SVHN) (Netzer et al. 2011), *synthetic sy* (Synthetic Digits) (Ganin and Lempitsky 2015), and *handwritten up* (USPS) (Hull 1994). Following (Xu et al. 2018; Peng et al. 2019), we sample 25,000 images for training and 9,000 for testing in **mt**, **mm**, **sv**, **sy**, and select the entire 9,298 images in **up** as a domain.

Office-31 (Saenko et al. 2010) contains 4,110 images within 31 categories, which are collected from office environment in 3 image domains: **A** (Amazon) downloaded from amazon.com, **W** (Webcam) and **D** (DSLR) taken by web camera and digital SLR camera, respectively.

Baselines To compare MDDA with the state-of-the-art approaches for MDA, we select the following methods as baselines. (1) **Source-only**, *i.e.* train on the source domains and test on the target domain directly. We can view this as a lower bound of DA. (2) **Single-source DA**, perform multi-source DA via single-source DA, including conventional models, *i.e.* TCA (Pan et al. 2011) and GFK (Gong et al. 2012), and deep methods, *i.e.* DDC (Tzeng et al. 2015), DRCN (Ghifary et al. 2016), RevGrad (Ganin and Lempitsky 2015), DAN (Long et al. 2015), RTN (Long et al. 2016), CORAL (Sun, Feng, and Saenko 2016), DANN (Ganin et al. 2016), and ADDA (Tzeng et al. 2017). (3) **Multi-source DA**, extend some single-source DA method to multi-source settings, including DCTN (Xu et al. 2018) and MDAN (Zhao et al. 2018a).

For the source-only and single-source DA standards, we employ two strategies: (1) source-combined, *i.e.* all source domains are combined into a traditional single source; (2) single-best, *i.e.* performing adaptation on each single source and selecting the best adaptation result in the target test set.

Table 2: Classification accuracy (%) on Office31 dataset for multi-source unsupervised domain adaptation. The best method is emphasized in bold. Our method achieves 84.2% accuracy, achieving the state-of-the-art performances.

Standards	Models	D	W	A	Avg
Source-only	Combined	97.1	92.0	51.6	80.2
	Single-best	99.0	95.3	50.2	81.5
Single-best DA	TCA (2011)	95.2	93.2	51.6	80.0
	GFK (2012)	95.0	95.6	52.4	81.0
	DDC (2015)	98.5	95.0	52.2	81.9
	DRCN (2016)	99.0	96.4	56.0	83.8
	RevGrad (2015)	99.2	96.4	53.4	83.0
	DAN (2015)	99.0	96.0	54.0	83.0
	RTN (2016)	99.6	96.8	51.0	82.5
Source-combined DA	ADDA (2017)	99.4	95.3	54.6	83.1
	RevGrad (2015)	98.8	96.2	54.6	83.2
	DAN (2015)	98.8	96.2	54.9	83.3
Multi-source DA	ADDA (2017)	99.2	96.0	55.9	83.7
	DCTN (2018)	99.6	96.9	54.9	83.8
	MDAN (2018a)	99.2	95.4	55.2	83.3
	MDDA (ours)	99.2	97.1	56.2	84.2

Implementation Details In Digits-five experiments, we use three convolutional layers and two fully connected layers as encoder and one fully connected layer as classifier. In Office-31 experiments, we use Alexnet as our backbone. The last layer is used as classifier and the other layers are used as encoder. Following (Gulrajani et al. 2017), we set α in Eq. (5) to 10.

Comparison with the State-of-the-art

The performance comparisons between MDDA and the state-of-the-art approaches as measured by classification accuracy are shown in Table 1 and Table 2 on Digits-five and Office-31 datasets, respectively. From the results, we have the following observations.

(1) The source-only method *i.e.* directly transferring the models trained on the source domains to the target domain performs the worst in most adaptation settings. Due to the presence of domain shift, the joint probability distributions of observed images and class labels greatly differ in the source and target domains. This results in the model’s low transferability from the source domains to the target domain. Further, even with more training samples, the Combined setting does not guarantee to perform better than the Single-best one. This is because domain shift also exists across different source domains, which may confuse the classifier. For example, if one source domain and the target is very similar, such as sv and sy, and the other source domains are quite different, simple combination would enlarge the domain shift between the Single-best and the target. This observation demonstrates the necessity of designing DA algorithms to address the domain shift problem.

(2) Almost all adaptation methods outperform the source-only methods, demonstrating the effectiveness of DA in image classification. Comparing the Single-best DA and Source-combined DA, it is clear that on average the Source-combined DA performs better, which is different from the source-only scenario. This is because after adaptation,

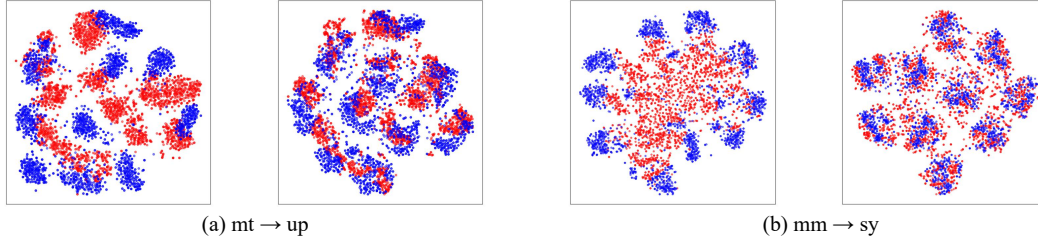


Figure 3: The t-SNE (Maaten and Hinton 2008) visualization of the digit-5 dataset features for (a) $mt \rightarrow up$ and (b) $mm \rightarrow sy$. In each pair, the features are extracted using the last layer of source domain encoder from the samples of source and target domain in the first image, and the target domain features are extracted using the last layer of adapted encoder in the second one.

Table 3: Ablation study of different weighting strategies in the proposed MDDA model on Digits-five dataset for multi-source unsupervised domain adaptation.

Weighting	mm	mt	up	sv	sy	Avg
Uniform	74.3	95.8	93.7	64.2	79.3	81.5
Ours	78.6	98.8	93.9	79.3	89.7	88.1

Table 4: Ablation study of different weighting strategies in the proposed MDDA model on Office31 dataset for multi-source unsupervised domain adaptation.

Weighting	D	W	A	Avg
Uniform	98.4	95.2	55.7	83.1
Ours	99.2	97.1	56.2	84.2

domain-invariant representations are learned for the samples of different domains. Therefore, the Source-combined DA works better with the help of more training data.

(3) Generally, multi-source DA performs better than other adaptation standards. This is more clear when comparing the methods that employ similar adaptation architectures, such as our MDDA vs. ADDA (Tzeng et al. 2017) and MDAN (Zhao et al. 2018a) vs. DANN (Ganin et al. 2016). Not only the domain shift between the sources and the target, but also the shift across the different source domains is bridged in multi-source DA, which boosts the adaptation by exploring the complementarity of different sources.

(4) The proposed MDDA model performs better than state-of-the-art multi-source methods in most cases. On one hand, the performance improvements of MDDA over the best Source-combined method are 3.1% and 0.5% on Digits-five and Office-31 datasets, respectively. On the other hand, the proposed MDDA method achieves 3.3%, 4.8% and 0.4%, 0.9% performance improvements as compared to DCTN (Xu et al. 2018) and MDAN (Zhao et al. 2018a) on Digits-five and Office-Home datasets, respectively. These results demonstrate that the proposed MDDA model can achieve superior performance relative to state-of-the-art approaches. The performance improvements benefit from the advantages of MDDA. First, the unshared weights enable to learn the best feature extractor and classifiers for each source domain, which would boost the performance when aggregation. Second, a novel weighting strategy based on

Table 5: Ablation study of whether distilling the source classifiers in the proposed MDDA model on Digits-five dataset for multi-source unsupervised domain adaptation.

	mm	mt	up	sv	sy	Avg
w/o	78.4	98.8	93.2	79.1	89.6	87.8
w	78.6	98.8	93.9	79.3	89.7	88.1

Table 6: Ablation study of whether distilling the source classifiers in the proposed MDDA model on Office31 dataset for multi-source unsupervised domain adaptation.

	D	W	A	Avg
w/o	99.2	96.0	55.8	83.7
w	99.2	97.1	56.2	84.2

the Wasserstein distance can better emphasize the domains that are more closer to the target. Finally, for each source domain, selective samples are distilled to fine-tune the source classifier, which also adapt better to the target features.

Interpretability and Ablation Study

Feature Visualization To show the adaptation ability of the proposed MDDA model, we visualize the features before and after adversarial adaptation with t-SNE embedding (Maaten and Hinton 2008) in tasks: $mt \rightarrow up$ and $mm \rightarrow sy$. As illustrated in Figure 3, we have two observations: (1) target features become more dense while using adversarial adaptation; (2) target domain fits source domain more tightly after the adversarial adaptation, which demonstrates that MDDA can align the distributions between the source and target domains.

Ablation Study The proposed MDDA model contains two major components: source distilling for fine-tuning the source classifiers and a novel weighting strategy for aggregating target prediction. We conduct ablation study to further verify their effectiveness by changing one component while fixing the other.

We compare the proposed weighting strategy with one straightforward baseline: uniform weight. The results on Digits-five and Office31 datasets are shown in Table 3 and Table 4, respectively. From the results, we can observe that the proposed weighting strategy outperforms the uniform weight. This is reasonable because the uniform weight does

Table 7: An example of detailed distilling result from each source to the target sy on Digits-five dataset.

Source	mt	mm	sv	up	Avg
w/o	52.0	70.8	89.4	38.6	62.7
w	54.5	71.0	89.5	40.8	64.0

not reveal the importance of different sources, which might have different similarities to the target. By considering the relative similarity of different sources to the target based on the Wasserstein distance, the proposed MDDA achieves 6.6% and 1.1% improvements on Digits-five and Office31 datasets, respectively. These observations demonstrate the effectiveness of the proposed weighting strategy.

Table 5 and Table 6 show the comparison between with and without fine-tuning the source classifiers by the distilled source samples on Digits-five and Office31 datasets, respectively. It is clear that without distilling, the adaptation performance drops in most cases. For example, we can achieve 0.3% and 0.5% average accuracy improvements by source distilling on Digits-five and Office31 datasets. This confirms the validity of distilling the sources, since the selected source samples are more similar to the target ones and the fine-tuned classifier can enhance the transferability.

To better demonstrate the effectiveness of source distilling, we give an example of Wasserstein Distance based ADDA method before and after distilling on the Digits-five dataset when sy is set as the target domain and the others as source domains. As shown in Table 7, we find that the performance gains of source distilling vary across different sources. For the sources with larger domain discrepancies to the target, *e.g.* mt to sy and up to sy, source distilling may yield higher improvement (2.5% and 2.1%, respectively), while the improvement is not that obvious for the sources with smaller discrepancy to the target, *e.g.* sv to sy (0.1%), mm to sy (0.2%). This is reasonable because when one source domain is far away from the target, the distilled samples can lead the classifier closer to target domain. If the source is already very similar to the target, the influence of distilled samples will be not that obvious.

Model Interpretability In order to show the interpretability of our model, we use the heat map generated by the Grad-Cam algorithm (Selvaraju et al. 2017) to visualize the attention before and after our proposed domain adaptation method. As illustrated in Figure 4, we observe that after the domain adaptation: the attentions generated by our model can better focus on the more “discriminative” regions, which indicates that our model can pay more attention to the discriminative regions of the objects for classification even though the background or view point are changed. Such observation verifies that our model learns the features that are more invariant to different domains, while they are discriminative for the desired learning task (*i.e.* image classification).

For example, the ring binder in the first row shows that before adaptation, the model focuses on a region in the background, instead of the central target object. However, after our domain adaptation, the model can correctly focus on the ring binder and thus is more discriminative for the classifi-

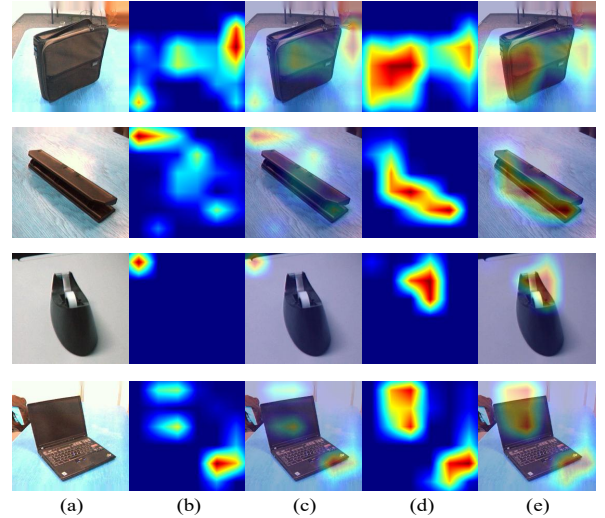


Figure 4: Comparison of the attention maps before and after adversarial training on Office-31 dataset. From left to right: (a) original image; (b) attention map before adversarial training; (c) image with attention map before adversarial training; (d) attention map after adversarial training; (e) image with attention map after adversarial training. Brighter regions indicate more attention. Comparison shows the attention shifts to more discriminative regions of the image after adversarial training. Best viewed in color.

cation. Similar observations can be found in the second and third rows. In the last row, we find that attention is enhanced on the discriminative regions of the object (the laptop) after our domain adaptation.

Conclusion

In this paper, we have proposed an effective multi-source domain adaptation approach MDDA. The separately pre-trained feature extractor and classifier for each source domain can sufficiently explore the discriminability of labeled source data. The adversarial discriminative-adaptation and source distilling aim to match the target feature distribution to the source ones and to fine-tune the pre-trained classifiers. A novel weighting strategy is designed to jointly combine the predictions from different source classifiers. The extensive experiments conducted on Digits-five and Office-31 benchmarks demonstrate that MDDA achieves 3.3% and 0.4% performance improvements as compared to the state-of-the-art multi-source domain adaptation approaches (*i.e.* DCTN) for digit and object classification. In future studies, we plan to extend the MDDA model to more challenging vision tasks, such as scene segmentation. We also aim to investigate methods that can combine generative and discriminative pipelines for multi-source domain adaptation.

Acknowledgments

This work is supported by Berkeley DeepDrive and the National Natural Science Foundation of China (No. 61701273).

References

- Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *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*.
- Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *NIPS*.
- Chattopadhyay, R.; Sun, Q.; Fan, W.; Davidson, I.; Panchanathan, S.; and Ye, J. 2012. Multisource domain adaptation and its application to early detection of fatigue. *ACM TKDD*.
- Chen, Y.-H.; Chen, W.-Y.; Chen, Y.-T.; Tsai, B.-C.; Frank Wang, Y.-C.; and Sun, M. 2017. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*.
- Ding, Z.; Shao, M.; and Fu, Y. 2018. Incomplete multisource transfer learning. *IEEE TNNLS*.
- Duan, L.; Xu, D.; and Chang, S.-F. 2012. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *CVPR*.
- Duan, L.; Xu, D.; and Tsang, I. W.-H. 2012. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE TNNLS*.
- GANIN, Y., and LEMPITSKY, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.
- GANIN, Y.; USTINOV, E.; AJAKAN, H.; GERMAIN, P.; LAROCHELLE, H.; LAVIOLETTE, F.; MARC-HAND, M.; and LEMPITSKY, V. 2016. Domain-adversarial training of neural networks. *JMLR*.
- Ghifary, M.; Kleijn, W. B.; Zhang, M.; Balduzzi, D.; and Li, W. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*.
- 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*.
- Gopalan, R.; Li, R.; and Chellappa, R. 2014. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE TPAMI*.
- Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *NIPS*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.
- Hoffman, J.; Mohri, M.; and Zhang, N. 2018. Algorithms and theory for multiple-source adaptation. In *NeurIPS*.
- Huang, H.; Huang, Q.; and Krahenbuhl, P. 2018. Domain transfer through deep activation matching. In *ECCV*.
- Hubert Tsai, Y.-H.; Yeh, Y.-R.; and Frank Wang, Y.-C. 2016. Learning cross-domain landmarks for heterogeneous domain adaptation. In *CVPR*.
- Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE TPAMI*.
- Kang, G.; Jiang, L.; Yang, Y.; and Hauptmann, A. G. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*.
- LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *PIEEE*.
- Li, W.; Duan, L.; Xu, D.; and Tsang, I. W. 2014. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE TPAMI*.
- Li, Y.; Murias, M.; Major, S.; Dawson, G.; and Carlson, D. E. 2018. Extracting relationships by multi-domain matching. In *NeurIPS*.
- Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*.
- Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 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*.
- Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR*.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshops*.
- Ni, J.; Zhang, S.; and Xie, H. 2019. Dual adversarial semantics-consistent network for generalized zero-shot learning. In *NeurIPS*.
- Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE TKDE*.
- Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011. Domain adaptation via transfer component analysis. *IEEE TNN*.
- Panareda Busto, P., and Gall, J. 2017. Open set domain adaptation. In *ICCV*.
- Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*.
- Redko, I.; Courty, N.; Flamary, R.; and Tuia, D. 2019. Optimal transport for multi-source domain adaptation under target shift. In *AISTATS*.
- Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *ECCV*.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.
- Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2017. Wasserstein distance guided representation learning for domain adaptation. *arXiv:1707.01217*.
- Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR*.
- Sun, S.-L., and Shi, H.-L. 2013. Bayesian multi-source domain adaptation. In *ICMLC*.
- Sun, Q.; Chattopadhyay, R.; Panchanathan, S.; and Ye, J. 2011. A two-stage weighting framework for multi-source domain adaptation. In *NIPS*.
- Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI*.
- Sun, B.; Feng, J.; and Saenko, K. 2017. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*.
- Sun, S.; Shi, H.; and Wu, Y. 2015. A survey of multi-source domain adaptation. *INF*.
- Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*.
- Tsai, Y.-H.; Hung, W.-C.; Schuler, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*.
- Tzeng, E.; Hoffman, J.; Darrell, T.; and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. In *ICCV*.
- Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.
- Xu, Z., and Sun, S. 2012. Multi-source transfer learning with multi-view adaboost. In *ICONIP*.
- Xu, R.; Chen, Z.; Zuo, W.; Yan, J.; and Lin, L. 2018. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *CVPR*.
- Zhao, H.; Zhang, S.; Wu, G.; Moura, J. M.; Costeira, J. P.; and Gordon, G. J. 2018a. Adversarial multiple source domain adaptation. In *NeurIPS*.
- Zhao, S.; Zhao, X.; Ding, G.; and Keutzer, K. 2018b. Emotiongan: unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM MM*.
- Zhao, S.; Li, B.; Yue, X.; Gu, Y.; Xu, P.; Hu, R.; Chai, H.; and Keutzer, K. 2019a. Multi-source domain adaptation for semantic segmentation. In *NeurIPS*.
- Zhao, S.; Lin, C.; Xu, P.; Zhao, S.; Guo, Y.; Krishna, R.; Ding, G.; and Keutzer, K. 2019b. Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions. In *AAAI*.
- Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
- Zhuo, J.; Wang, S.; Zhang, W.; and Huang, Q. 2017. Deep unsupervised convolutional domain adaptation. In *ACM MM*.