Learning Semantic Representations for Unsupervised Domain Adaptation

Shaoan Xie [1 2] **Zibin Zheng** [1 2] **Liang Chen** [1 2] **Chuan Chen** [1 2]

Abstract

It is important to transfer the knowledge from label-rich source domain to unlabeled target domain due to the expensive cost of manual labeling efforts. Prior domain adaptation methods address this problem through aligning the global distribution statistics between source domain and target domain, but a drawback of prior methods is that they ignore the semantic information contained in samples, e.g., features of backpacks in target domain might be mapped near features of cars in source domain. In this paper, we present moving semantic transfer network, which learn semantic representations for unlabeled target samples by aligning labeled source centroid and pseudo-labeled target centroid. Features in same class but different domains are expected to be mapped nearby, resulting in an improved target classification accuracy. Moving average centroid alignment is cautiously designed to compensate the insufficient categorical information within each mini batch. Experiments testify that our model yields state of the art results on standard datasets.

1. Introduction

Deep learning approaches have gained prominence in various machine learning problems and applications. However, the recent success of deep learning depends on massive labeled data. Manual large scale labeled data on the target domain are too expensive or impossible to collect in practice. Therefore, there is a strong motivation to build an effective classification model using available labeled data from other domains. But, this learning paradigms suffers from the domain shift problem, which is an huge obstacle for adapting predictive models to the target domain (Pan & Yang, 2010).

Learning a discriminative predictor in the presence of the shift between source domain and target domain is known as domain adaptation (Pan & Yang, 2010). In recent years, deep learning has shown its potential to produce transferable features for domain adaptation. Fruitful line of works have been done in deep domain adaptation (Motiian et al., 2017b; Tzeng et al., 2014; Long et al., 2015). These methods aim at matching the marginal distributions across domains while (Zhang et al., 2013; Gong et al., 2016) considers the conditional distribution shift problem. Recently adversarial adaptation methods (Ganin & Lempitsky, 2015; Tzeng et al., 2017; Motiian et al., 2017a; Bousmalis et al., 2016) have shown promising results in domain adaptation. Adversarial adaptation methods is analogous to generative adversarial networks (GAN) (Goodfellow et al., 2014). A domain classifier is trained to tell whether the sample comes from source domain or target domain. The feature extractor is trained to minimize the classification loss and maximize the domain confusion loss. Domain-invariant yet discriminative features are seemingly obtainable through the principled lens of adversarial training.

Prior adversarial adaptation methods suffer a main limitation: as the discriminator only enforces the alignment of global domain statistics, crucial semantic information for each category might be lost. Even with perfect confusion alignment, there is no guarantee that samples from different domains but with the same class label will map nearby in the feature space, e.g, features of backpacks in the target domain may be mapped near features of cars in the source domain. This lack of semantic alignment is an important source of performance reduction (Motiian et al., 2017a; Hoffman et al., 2017; Luo et al., 2017). Recently, semantic transfer for supervised domain adaptation has received wide attention (Motiian et al., 2017a; Luo et al., 2017). To date, semantic alignment has not been addressed in unsupervised domain adaptation due to the lack of target label information.

In this paper, we propose a novel moving semantic transfer network (MSTN) for unsupervised domain adaptation, where our feature extractor learns to align the distributions semantically without any labeled target samples. We large-

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China [2] National Engineering Research Center of Digital Life, Sun Yat-sen University, Guangzhou, China. Correspondence to: Zibin Zheng <zhzibin@mail.sysu.edu.cn>, Liang Chen <chenliang6@mail.sysu.edu.cn>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

ly extend the ability of prior adversarial adaptation methods by our proposed semantic representation learning module. We firstly assign pseudo labels to target samples to fix the problem of lacking target label information. Since there are obviously false labels in pseudo labels, we wish to use correctly-pseudo-labeled samples to reduce the bias caused by falsely-pseudo-labeled samples. So we propose to align the centroid for each class in source and target domains instead of treating the pseudo-labeled samples as true directly. In particular, as we use mini batch SGD in practice, categorical information is usually insufficient and even one false label could lead to extremely biased estimation of the true centroid, *moving average centroid* is designed for safer semantic representation learning. Experiments have proven that MSTN yields state of the art results on standard datasets. Furthermore, we also find that MSTN stabilizes the adversarial learning for unsupervised domain adaptation.

2. Related Work

Recently, adversarial learning has been widely adopted in domain adaptation (Ganin & Lempitsky, 2015; Tzeng et al., 2015; Hoffman et al., 2017; Motiian et al., 2017a; Tzeng et al., 2017; Saito et al., 2017b; Long et al., 2017a; Luo et al., 2017; Sankaranarayanan et al.). Most of adversarial adaptation methods are based on generative adversarial networks (GAN) (Goodfellow et al., 2014). A discriminator is trained to tell whether the sampled feature comes from the source domain or target domain while the feature extractor is trained to fool the discriminator. However, prior unsupervised adversarial domain adaptation methods only enforce embedding alignment in domain-level instead of class-level transfer. Lacking the semantic alignment hurts the performance of domain adaptation significantly (Motiian et al., 2017a;b).

Semantic transfer is much easier in supervised domain adaptation as labeled target samples are available. In recent years, few-shot adversarial learning (Tzeng et al., 2015; Motiian et al., 2017a; Luo et al., 2017) have been explored in domain adaptation. Few-shot domain adaptation considers the task where very few labeled target data are available in training. (Tzeng et al., 2015) computes the average output probability with source training samples for each category, then for each labeled target sample, they optimize the model to match the distributions over classes to the average probability. FADA (Motiian et al., 2017a) pairs the labeled target sample and labeled source sample and the discriminator is trained to tell whether the pair comes from same domain and same class. (Luo et al., 2017) proposes cross category similarity for semantic transfer.

In this paper, we consider a more challenging task: unsupervised semantic transfer where there is no labeled target samples. (Ghifary et al., 2016) proposes to add a decoder after the feature extractor to enforce the feature extractor preserving semantic information. (Bousmalis et al., 2016) propose to decouple the representation into the shared representation and private representation. It encourages the shared and private representation to be orthogonal while both the representations should be able to be decoded back to images. (Hoffman et al., 2017) adapts representations at both the pixel-level and feature-level. It encourages the feature extractor to preserve semantic information by using the cycle consistency constraints. (Saito et al., 2017b) uses the dropout to obtain two different views of input and if the prediction results are different, these target samples are regarded as near decision boundary. They use the boundary information to achieve low-density separation of aligned points. (Saito et al., 2017c) proposes to use two classifiers as discriminators to detect target samples that are far from the support of the source. These two classifiers are trained adversarial to view input differently. (Pinheiro, 2017) classifies the input samples by computing the distances between prototype representations of each category.

Previous unsupervised adaptation methods do not necessarily align distributions semantically across domains as they can not ensure features in same class but different domains are mapped nearby owing to the huge gap for semantic alignment: no labeled information for target samples. It means that explicit matching the distributions for each category is impossible. To fill this gap, we assign pseudo labels to target samples. Contrary to prior domain adaptation methods that assign pseudo labels (Chen et al., 2011; Saito et al., 2017a), we doubt the pseudo labels and propose to align the centroid to reduce the shift brought by false labels instead of direct matching distributions using pseudo labels.

3. Method

In this section, we provide details of the proposed model for domain adaptation. In unsupervised domain adaptation, we are given by n_s labeled samples $\left\{(x_S^{(i)}, y_S^{(i)})\right\}_{i=1}^{n_s}$ from the source domain $\mathcal{D}_S$, where $x_S^{(i)} \in \mathcal{X}_S$ and $y_S^{(i)} \in \mathcal{Y}_S$. Additionally, we are also given with n_t unlabeled target samples $\left\{(x_T^{(i)})\right\}_{i=1}^{n_t}$ from the target domain $\mathcal{D}_T$, where $x_T^{(i)} \in \mathcal{X}_T$. $\mathcal{X}_S$ and $\mathcal{X}_T$ are assumed to be different but related (referred as *covariate shift* in literature (Shimodaira, 2000)). Target task is assumed to be same with source task. Our ultimate goal is to develop a deep neural network $f : \mathcal{X}_T \to \mathcal{Y}_T$ that is able to predict labels for the samples from target domain.

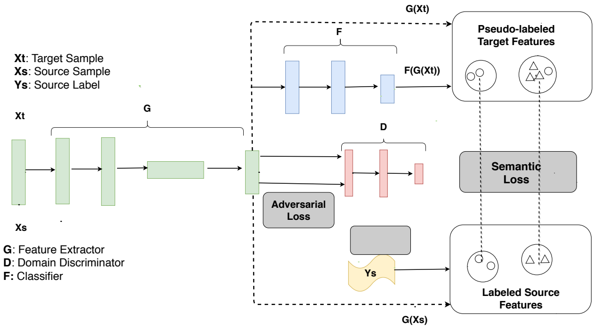


Figure 1. Besides the standard source classification loss, we also employ the domain adversarial loss to align distributions for two domains. In particular, to learn semantic representations, we maintain global centroids C_S^k and C_T^k for each class k in two domains at feature level, i.e., $\mathrm{G(X)}$. In each step, source centroids will be updated with the labeled features $\mathrm{(G(Xs), Ys)}$ while target centroids will be updated with pseudo-labeled features $\mathrm{(G(Xt), F{\circ}G(Xt))}$. Our model learns to semantically align the embedding by explicitly restricting the distance between centroids in same class but different domains.

3.1. The model

For unsupervised domain adaptation, in the presence of covariate shift, a visual classifier $f = \mathrm{F} \circ \mathrm{G}$ is trained by minimizing the source classification error and the discrepancy between source domain and target domain:

$$\mathcal{L} = \underbrace{\mathbb{E}_{(x,y)\sim D_S}[J(f(x),y)]}_{\mathcal{L}_C(\mathcal{X}_S,\mathcal{Y}_S)} + \lambda \underbrace{d(\mathcal{X}_S,\mathcal{X}_T)}_{\mathcal{L}_{DC}(\mathcal{X}_S,\mathcal{X}_T)} \tag{1}$$

where $J(.,.)$ is typically the cross entropy loss, λ is the balance parameter, $d(.,.)$ represents the divergence between two domains. Typically maximum mean discrepancy (MMD) (Long et al., 2015; Tzeng et al., 2014) or domain adversarial similarity loss (Bousmalis et al., 2016; Ganin & Lempitsky, 2015) are used to measure the divergence. We opt to use the domain adversarial similarity loss in our model. In other words, we employ an additional domain classifier **D** to tell whether the features from feature extractor **G** arise from source or target domain while **G** is trained to fool **D**. This two-player game is expected to reach an equilibrium where features from **G** are domain-invariant. Formally,

$$\begin{aligned} d(\mathcal{X}_S,\mathcal{X}_T) = &\mathbb{E}_{x\sim D_S}[\log(1-\mathrm{D}\circ\mathrm{G}(\mathrm{x}))] \\ &\mathbb{E}_{x\sim D_T}[\log(\mathrm{D}\circ\mathrm{G}(\mathrm{x}))] \end{aligned} \tag{2}$$

However, domain-invariance does not mean discriminability. Features of target backpacks can be mapped near features of source cars while satisfying the condition of domain-invariant. Separately, it has been shown that supervised domain adaptation (SDA) method improves upon unsupervised domain adaptation (UDA) by making the alignment semantic since SDA can ensure features of same class in different domains are mapped nearby (Motiian et al., 2017b). Motivated by this key observation, we endeavor to learn **semantic representations** for UDA.

Before we go further, we will stop to see how SDA achieves semantic transfer. For SDA, one could easily align the embeddings semantically by adding following objective,

$$\mathcal{L}_{SM}^{SDA}(\mathcal{X}_S,\mathcal{X}_T,\mathcal{Y}_S,\mathcal{Y}_T) = \sum_{k=1}^{K} d(\mathcal{X}_S^k,\mathcal{X}_T^k), \tag{3}$$

where K is the number of classes. It means that one can match the distributions for each class directly in SDA.

Unfortunately, for UDA, we do not have label information from target domain. To circumvent the impossibility of distribution matching at class-level, we resort to pseudo labels (Lee, 2013). We firstly assign pseudo labels to target samples with the training classifier f and we obtain a pseudo-labeled target domain. But obviously there must be some false labels and they may harm the performance of adaptation heavily. A natural question then arises as how to suppress the noisy signals conveyed in those false pseudo-labeled samples?

We approach the question by **centroid alignment**. Centroid has long been favored for its simplicity and effectiveness to represent a set of samples (Luo et al., 2017; Snell et al., 2017). When computing the centroid for each class, pseudo-labeled (correct or wrong) samples are being used together and the detrimental influences brought by false pseudo labels are expected be neutralized by correct pseudo labels. Inspired by this, we propose following semantic transfer objective for unsupervised domain adaptation:

$$\mathcal{L}_{SM}^{UDA}(\mathcal{X}_S,\mathcal{Y}_S,\mathcal{X}_T) = \underbrace{\sum_{k=1}^{K} \Phi(C_S^k, C_T^k)}_{\mathcal{L}_{SM}(\mathcal{X}_S,\mathcal{Y}_S,\mathcal{X}_T)}, \tag{4}$$

where C_S^k and C_T^k are centroid for each class in feature space, $\Phi(.,.)$ is any appropriate distance measure function. We use the squared Euclidean distance $\Phi(x,x') = ||x-x'||^2$ in our experiments. In total, we obtain 2K centroids. Through explicitly restricting the distance between centroids with same class label but different domains, we can ensure that features in the same class will be mapped nearby. More importantly, false signals in pseudo-labeled target domain are suppressed through centroid alignment.

More formally, our totally objective can be written as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{X}_S,\mathcal{Y}_S,\mathcal{X}_T) = &\mathcal{L}_C(\mathcal{X}_S,\mathcal{Y}_S) + \lambda\mathcal{L}_{DC}(\mathcal{X}_S,\mathcal{X}_T) \\ &+ \gamma\mathcal{L}_{SM}(\mathcal{X}_S,\mathcal{Y}_S,\mathcal{X}_T), \end{aligned} \tag{5}$$

where λ and γ are parameters that balance the classification loss, domain confusion loss and semantic loss. As we can see, our model is simple and the semantic transfer objective can be computed in linear time.

3.2. Moving Semantic Transfer Network

Algorithm 1 Moving semantic transfer loss computation in iteration t in our model. K is the number of classes.

```
Input: Labeled set S, Unlabeled set T, $N$ is the batch size, Training classifier $f$, Global centroids for two domains: $\{C^k_S\}^K_{k=1}$ and $\{C^k_T\}^K_{k=1}$
$S_t = $ RandomSample($S, N$)
$T_t = $ RandomSample($T, N$)
$\widehat{\boldsymbol{T_t}}$ =Labeling(G,$f$,$T_t$)
$\mathcal{L}_{SM} = 0$
for $k = 1$ to $K$ do
    $C^k_{S_{(t)}} \leftarrow \frac{1}{|S^k_t|} \sum_{(x_i,y_i)\in S^k_t} G(x_i)$ (From Scratch)
    $C^k_{T_{(t)}} \leftarrow \frac{1}{|\widehat{\boldsymbol{T^k_t}}|} \sum_{(x_i,y_i)\in \widehat{\boldsymbol{T^k_t}}} G(x_i)$ (From Scratch)
    $\boldsymbol{C^k_S \leftarrow \theta C^k_S + (1-\theta) C^k_{S_{(t)}}}$ (Moving Average)
    $\boldsymbol{C^k_T \leftarrow \theta C^k_T + (1-\theta) C^k_{T_{(t)}}}$ (Moving Average)
    $\mathcal{L}_{SM} \leftarrow \mathcal{L}_{SM} + \Phi(C^k_S, C^k_T)$
end for
return $\mathcal{L}_{SM}$
```

The proposed model achieves semantic transfer in very simple form but it suffers two limitations in practice: **(1)** As we always uses mini batch SGD for optimization in practice, categorical information in each batch is usually insufficient. For instance, it is possible that some classes are missing in the current batch of target data since the batch is randomly selected. **(2)** If the batch size is small, even one false pseudo label will lead to the huge deviation between the pseudo-labeled centroid and true centroid. For example, when there is one pseudo-labeled *car* sample in a target batch but the true label is *backpack*. Then it will wrongly guide the alignment between source *car* features and target *backpack* features.

Instead of aligning those newly obtained centroids in each iteration directly, we propose to align exponential moving average centroids to address the two aforementioned problems. As shown in algorithm 1, we maintain global centroids for each class. In each iteration, source centroids are updated by the labeled source samples while target centroids are updated by pseudo-labeled target samples. Then we can align those moving average centroids following equation (4).

Moving average centroid alignment works in an intuitive way: When *backpack* are missing in current source batch, we can align the target *backpack* centroid with the global source *backpack* centroid updated in last iteration. Under the reasonable assumption that centroids change by a limited step in each iteration, we can still ensure features of *backpacks* in two domains are mapped nearby. Meanwhile, when there is one pseudo-labeled *car* sample in a target mini batch but the true label is *backpack*, moving average centroids can avoid the aforementioned misalignment as it also considers the pseudo-labeled *backpacks* in the past mini batches.

Our method attempts to align the centroids in same class but different domains to achieve semantic transfer for unsupervised domain adaptation. We use pseudo labels from F to guide the semantic alignment for G. As the learning proceeds, G will learn semantic representations for target samples, resulting in an improved accuracy of F. This cycle will gradually enhance the accuracy for target domain. In addition, we suppress the noisy semantic information by assigning a small weight to γ in early training phase .

3.3. Analysis

In this section, we show the relationship between our method and the theory of domain adaptation (Ben-David et al., 2010). The theory bounds the expected error on the target samples $\varepsilon_\mathcal{T}(h)$ by three terms as follows.

Theorem 1. (Ben-David et al., 2010) Let $\mathcal{H}$ be the hypothesis class. Given two domains $\mathcal{S}$ and $\mathcal{T}$, we have

$$\forall h \in \mathcal{H}, \varepsilon_\mathcal{T}(h) \leq \varepsilon_\mathcal{S}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T}) + C, \tag{6}$$

where $\varepsilon_\mathcal{S}(h)$ is the expected error on the source samples which can be minimized easily with source label information, $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T})$ defines a discrepancy distance between two distributions $\mathcal{S}$ and $\mathcal{T}$ w.r.t. a hypothesis set $\mathcal{H}$. C is the shared expected loss and is expected to be negligibly small, thus usually disregarded by previous methods (Ganin & Lempitsky, 2015; Long et al., 2015). But it is very important and we cannot expect to learn a good target classifier by minimizing the source error if C is large (Ben-David et al., 2010).

It is defined as $C = \min_{h\in\mathcal{H}} \varepsilon_\mathcal{S}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(h, f_\mathcal{T})$ where $f_\mathcal{S}$ and $f_\mathcal{T}$ are labeling functions for source and target domain respectively. We show that our method is trying to optimize the upper bound for C. Recall the triangle inequality for classification error (Ben-David et al., 2010; Crammer et al., 2008), which implies that for any labeling functions f_1, f_2 and f_3, we have $\varepsilon(f_1, f_2) \leq \varepsilon(f_1, f_3) + \varepsilon(f_2, f_3)$. Then

$$\begin{aligned} C &= \min_{h\in\mathcal{H}} \varepsilon_\mathcal{S}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(h, f_\mathcal{T}) \\ &\leq \min_{h\in\mathcal{H}} \varepsilon_\mathcal{S}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(f_\mathcal{S}, f_\mathcal{T}) \\ &\leq \min_{h\in\mathcal{H}} \varepsilon_\mathcal{S}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(h, f_\mathcal{S}) + \varepsilon_\mathcal{T}(f_\mathcal{S}, f_{\widehat{\mathcal{T}}}) + \varepsilon_\mathcal{T}(f_\mathcal{T}, f_{\widehat{\mathcal{T}}}) \end{aligned} \tag{7}$$

The first and second term denotes the disagreement between h and the source labeling function $f_{\mathcal{S}}$. These two terms should be small as we can easily find such a h in our hypothesis space to approximate the $f_{\mathcal{S}}$ since we have source labels. Therefore, we seek to minimize the last two terms. Obviously the last term denotes the false pseudo rate in our method which would be minimized as learning proceeds. Now our focus should be the third term $\varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\widehat{\mathcal{T}}})$. This term denotes the disagreement between the source labeling function and pseudo target labeling function on target samples. $\varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\widehat{\mathcal{T}}}) = \mathbb{E}_{x \sim \mathcal{T}}[l(f_{\mathcal{S}}(x), f_{\widehat{\mathcal{T}}}(x))]$, where $l(.,.)$ is typically the 0-1 loss function.

Our method aligns the centroid for class k in source domain $\mathcal{S}^k$ and pseudo-labeled target domain $\widehat{\mathcal{T}^k}$. We can decompose the hypothesis h into the feature extractor G and classifier F. Then we have $\mathbb{E}_{x \sim \mathcal{S}^k} G(x) = \mathbb{E}_{x \sim \widehat{\mathcal{T}^k}} G(x)$. For $\varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\widehat{\mathcal{T}}})$, it could be rewritten as

$$\mathbb{E}_{x \sim \mathcal{T}}[l(F_{\mathcal{S}} \circ G(x), F_{\widehat{\mathcal{T}}} \circ G(x))] \tag{8}$$

Now the relationship is clear: for source samples in class k, the source labeling function should return k. We wish to have target features in class k to be similar with source features in class k, so the source labeling function would also predict those target samples as k, which is consistent with the prediction results made by pseudo target labeling function. Consequently, $\varepsilon_{\mathcal{T}}(f_{\mathcal{S}}, f_{\widehat{\mathcal{T}}})$ is expected to be small.

In summary, the premise for the success of domain adaptation methods is that the shared expected loss C should be small. Our method attempts to minimize this item through aligning the centroid between source domain and pseudo-labeled target domain.

4. Experiments

4.1. Setup

We evaluate the semantic transfer network with state of art transfer learning methods. Codes are available at `https://github.com/Mid-Push/Moving-Semantic-Transfer-Network`.

Office-31 (Saenko et al., 2010) is a standard dataset used for domain adaptation. It contains three distinct domains: Amazon (A) with 2817 images, Webcam (W) with 795 images and DSLR (D) with 498 images. Each domain contains 31 categories. We examine our methods by employing the frequently used network structures: **AlexNet** (Krizhevsky et al., 2012). For fair comparison, we report results of methods that are also based on AlexNet.

ImageCLEF-DA is a benchmark dataset for ImageCLEF 2014 domain adaptation challenges. Three domains including Caltech-256 (C), ImageNet ILSVRC 2012 (I) and Pascal VOC 2012 (P) share 12 categories. Each domain contains 600 images and 50 images for each category. Images in ImageCLEF-DA are of equal size. This dataset has been used by JAN (Long et al., 2017b). Same, we also examine our method in AlexNet (Krizhevsky et al., 2012).

MNIST-USPS-SVHN. We explore three digits datasets of varying difficulty: MNIST (LeCun et al., 1998), USPS and SVNH (Netzer et al., 2011). Different from Office-31, MNIST consists grey digits images of size 28x28, USPS contains 16x16 grey digits and SVHN composes color 32x32 digits images which might contain more than one digit in each image. MNIST-USPS-SVHN makes a good complement to previous datasets for diverse domain adaptation scenarios. We conduct experiments in a resolution-going-down way, SVHN→ MNIST and MNIST →USPS.

Baseline Methods For Office-31 and ImageCLEF-DA datasets, we compare with state-of-art transfer learning methods: Deep Domain Confusion (**DDC**) (Tzeng et al., 2014), Deep Reconstruction Classification Network (**DRCN**) (Ghifary et al., 2016), Gradient Reversal (**RevGrad**) (Ganin & Lempitsky, 2015), Residual Transfer Network (**RTN**) (Long et al., 2016), Joint Adaptation Network (**JAN**) (Long et al., 2017b), Automatic Domain Alignment Layer (**AutoDIAL**) (Carlucci et al., 2017). We cite the results of AlexNet, DDC, RevGrad, RTN, JAN from (Long et al., 2017b). For DRCN and AutoDIAL, we cite the results in their papers. For ImageCLEF-DA, we compare with AlexNet, RTN, RevGrad and JAN. Results are cited from (Long et al., 2017a). To further validate our method, we also conduct experiments on MNIST-USPS-SVHN. We compare with Domain of Confusion (**DOC**) (Tzeng et al., 2014), **RevGrad** (Ganin & Lempitsky, 2015), Asymmetric Tri-Training (**AsmTri**) (Saito et al., 2017a), Couple GAN (**CoGAN**) (Liu & Tuzel, 2016), Label Efficient Learning (**LEL**) (Luo et al., 2017) and Adversarial Discriminative Domain Adaptation (**ADDA**) (Tzeng et al., 2017). Results of source only, DOC, RevGrad, CoGAN and ADDA are cited from (Tzeng et al., 2017). For the rest, we cite the result in their papers respectively.

We follow standard evaluation protocols for unsupervised domain adaptation as (Long et al., 2015; Ganin & Lempitsky, 2015; Long et al., 2017b). We use all labeled source examples and all unlabeled target examples. We repeat each transfer task three times and report the mean accuracy as well as the standard error.

4.2. Implementation Detail

CNN architecture. In our experiments on Office and ImageCLEF-DA, we employed the AlexNet architecture. Following RTN (Long et al., 2016) and RevGrad (Ganin & Lempitsky, 2015), a bottleneck layer fcb with 256 units is added after the $fc7$ layer for safer transfer representation learning. We use fcb as inputs to the discriminator

Table 1. Classification accuracies (%) on office-31 datasets.(AlexNet)

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
AlexNet (Krizhevsky et al., 2012)	61.6±0.5	95.4±0.3	99.0±0.2	63.8±0.5	51.1±0.6	49.8±0.4	70.1
DDC (Tzeng et al., 2014)	61.8±0.4	95.0±0.5	98.5±0.4	64.4±0.3	52.1±0.6	52.2±0.4	70.6
DRCN (Ghifary et al., 2016)	68.7±0.3	96.4±0.3	99.0±0.2	66.8±0.5	56.0±0.5	54.9±0.5	73.6
RevGrad (Ganin & Lempitsky, 2015)	73.0±0.5	96.4±0.3	99.2±0.3	72.3±0.3	53.4±0.4	51.2±0.5	74.3
RTN (Long et al., 2016)	73.3±0.3	96.8±0.2	99.6±0.1	71.0±0.2	50.5±0.3	51.0±0.1	73.7
JAN (Long et al., 2017b)	74.9±0.3	96.6±0.2	99.5±0.2	71.8±0.2	58.3±0.3	55.0±0.4	76.0
AutoDIAL (Carlucci et al., 2017)	75.5	96.6	99.5	73.6	58.1	59.4	77.1
MSTN (centroid from scratch,ours)	80.3±0.7	96.8±0.1	**100**±0.1	73.8±0.1	60.7±0.1	59.9±0.3	78.6
MSTN (ours)	**80.5**±0.4	**96.9**±0.1	99.9±0.1	**74.5**±0.4	**62.5**±0.4	**60.0**±0.6	**79.1**

Table 2. Classification accuracies (%) on ImageCLEF-DA datasets.(AlexNet)

Method	I → P	P → I	I → C	C → I	C → P	P → C	Avg
AlexNet (Krizhevsky et al., 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
RTN (Long et al., 2016)	**67.4**±0.3	81.3±0.3	89.5±0.4	78.0±0.2	62.0±0.2	89.1±0.1	77.9
RevGrad (Ganin & 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
JAN (Long et al., 2017b)	67.2±0.5	**82.8**±0.4	91.3±0.5	80.0±0.5	63.5±0.4	91.0±0.4	79.3
MSTN (ours)	67.3±0.3	**82.8**±0.2	**91.5**±0.1	**81.7**±0.3	**65.3**±0.2	**91.2**±0.2	**80.0**

as well as the centroid computation. Image random flipping and cropping are adopted following JAN (Long et al., 2017b). For a fair comparison with other methods, we also finetune the $conv1, conv2, conv3, conv4, conv5, fc6, fc7$ layers with pretrained AlexNet. For discriminator, we use same architecture with RevGrad, x→1024→1024→1, dropout is used.

For digit classification datasets, we use same architecture with ADDA (Tzeng et al., 2017): two convolution layers followed by max pool layers and two fully connected layers are placed behind. Digit images are also cast to 28x28x1 in all experiments for fair comparison. For discriminator, we also use same architecture with ADDA, x→500→500→1. Batch Normalization is inserted in convolutional layers.

Hyper-parameters tuning. A good unsupervised domain adaptation method should provide ways to tune hyper-parameters in an unsupervised way. Therefore, no labeled target samples are referred for tuning hyper-paramters. We essentially tune the three hyper-parameters: weight balance parameter λ, γ and moving average coefficient θ. For θ, we first apply reverse validation (Ganin & Lempitsky, 2015) on the experiments MNIST→USPS. Then we use the optimal value for θ in all experiments. We set θ=0.7 in all our experiments. For the weight balance parameter, we set $\lambda = \frac{2}{1+exp(-\gamma.p)} - 1$, where γ is set to 10 and p is training progress changing from 0 to 1. It is optimized by (Ganin & Lempitsky, 2015) to suppress noisy signal from the discriminator at the early stages of training. Considering that our pseudo-labeled semantic loss would be inaccurate in early training phase, we also set $\gamma = \lambda$ to suppress the noisy information brought by false labels. Stochastic gradient descent with 0.9 momentum is used. The learning rate is annealed by $\mu_p = \frac{\mu_0}{(1+\alpha.p)^\beta}$, where μ_0=0.01, α=10 and β=0.75 (Ganin & Lempitsky, 2015). We set the learning rate for finetuned layers to be 0.1 times of that from scratch. We set the batch size to 128 for each domain. Domain adversarial loss is scaled by 0.1 following (Ganin & Lempitsky, 2015).

4.3. Results

We now discuss the experiment settings and results.

Office-31 We follow the fully transductive evaluation protocol in (Ganin & Lempitsky, 2015). Results of office-31 are shown in Table 1. The proposed model outperforms all comparison methods on all transfer tasks. It is noteworthy that MSTN results in improved accuracies on four hard transfer task: **A→W**, **A→D**, **D→A** and **W→A**. On these four difficult tasks, our method promote classification accuracies substantially. The encouraging improvement on hard transfer tasks proves the importance of semantic alignment and suggests that our method is able to learn semantic representations effectively despite of its simplicity.

The results reveal several interesting observations. **(1)** Deep transfer learning methods outperform standard deep learning methods. It validates that the idea that domain shift in two distributions can not be removed by deep networks (Yosinski et al., 2014). **(2)** DRCN (Ghifary et al., 2016) trains an extra decoder to enforce the extracted features contain semantic information and thus outperformed standard deep learning methods by about 5%. This improvement also indicates the importance to learn seman-

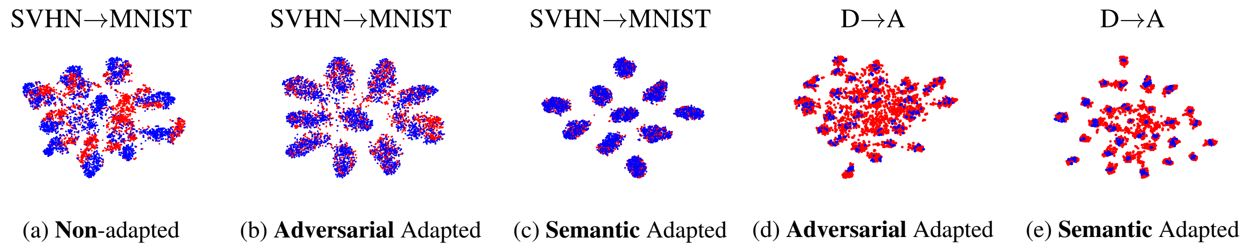


Figure 2. SVHN→MNIST and D→A. We confirmed the effects our method through a visualization of the learned representations using t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008). Blue points are source samples and red are target samples. **(a)** are trained without any adaptation. **(b)(d)** are trained with previous adversarial domain adaptation methods. **(c)(e)** Adaptation using our proposed method. As we can see, compared to non-adapted method, adversarial adaptation methods successfully fuse the source features and target features. But semantic information are ignored and ambiguous features are generated near class boundary, which is catastrophic for classification task. Our model attempts to fuse features in the same class while separate features in different classes.

tic representations. **(3)** Separately, distribution matching methods RevGrad, RTN and JAN, also bring significant improvement over source only. Our method combines the advantages of DRCN and distribution matching methods in a very simple form. In particular, in contrast to using a decoder to extract semantic information, our method also ensures that the features in same classes but different domains are similar, which has not been addressed by any existing methods. For completeness, we also conduct a visualization over transfer task **D**→ **A** for comparison between our learned representation and prior adversarial adaptation method RevGrad (Ganin & Lempitsky, 2015). See Fig (2d) and (2e). Representations learned by our model are better behaved compared to RevGrad and representations in different classes are dispersed instead of mixing up.

To dive deeper into our method, we present the results of one variants of MSTN: MSTN with centroid from scratch. We try to align the centroids directly computed in each iteration instead of using moving average. The results are interesting, for the simple transfer task D→W, W→D, this variant are comparable or outperforms the moving average. This phenomenon is plausible since the prediction accuracy for target domains is already very high and introducing the past semantic information might introduce noisy information too. But take a look at the hard transfer task D→A and A→D, the improvement carried by the moving average centroid is obvious. This curious result provides us two training instructions: **(1)** for easy transfer tasks or large batch size, one could just align the centroids directly to learn semantic representation in each iteration. **(2)** for hard transfer tasks or small batch size, one could effectively pass the semantic information by aligning the moving average centroids. Note that our method does not introduce any extra network architecture but only few memory that are used to keep these global centroids.

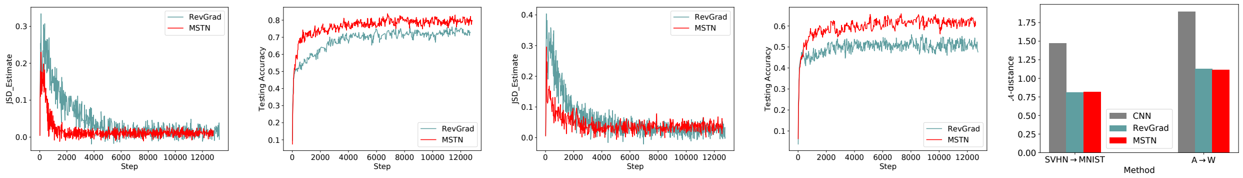
ImageCLEF-DA For ImageCLEF-DA, results are shown in Table 2. Images are balanced in ImageCLEF-DA, so our model could be more focused on transfer learning by avoiding the class imbalance problem. But the domain size is limited to 600, which might not be sufficient for training the network. Our model outperforms existing methods in most transfer tasks, but with less improvement compared to Office-31. This result also validates hypothesis in (Long et al., 2017b) that the domain size may cause shift.

Table 3. Classification accuracies (%) on digit recognitions tasks

Source	SVHN	MNIST
Target	MNIST	USPS
Source Only	60.1±1.1	75.2±1.6
DOC (Tzeng et al., 2014)	68.1±0.3	79.1±0.5
RevGrad (Ganin & Lempitsky, 2015)	73.9	77.1±1.8
AsmTri (Saito et al., 2017a)	86.0	-
coGAN (Liu & Tuzel, 2016)	-	91.2±0.8
ADDA (Tzeng et al., 2017)	76.0±1.8	89.4±0.2
LEL (Luo et al., 2017)	81.0±0.3	-
MSTN (ours)	**91.7**±1.5	**92.9**±1.1

MNIST-USPS-SVHN We follow the protocols in (Tzeng et al., 2017): For adaptation between SVHN and MNIST, we use the training set of SVHN and test set of MNIST for evaluation. For adaptation between MNIST and USPS, we randomly sample 2000 images from MNIST and 1800 from USPS. For SVHN→MNIST, the transfer gap is huge since images in SVHN might contain multiple digits. Thus, to avoid ending up in a local minimum, we do not use the learning rate annealing as suggested by (Ganin & Lempitsky, 2015).

Results of MNIST-USPS-SVHN are shown in Table 3. It shows that our model outperforms all comparison methods. For MNIST → USPS, our method obtains a desirable performance. On the difficult transfer task SVHN → MNIST, Our model outperforms existing methods by about 6.6%. In Fig. 2, the representations in **SVHN→MNIST**



(a) JSD A→W (b) Accuracy A→W (c) JSD D→A (d) Accuracy D→A (e) $\mathcal{A}$-distance

Figure 3. Standard CNN in **grey**, Revgrad (Ganin & Lempitsky, 2015) in **green**, our model MSTN in **red**. **(a)(c)**: Comparison of Jensen-Shannon divegence (JSD) estimate during training for RevGrad and our proposed method MSTN. Our model stabilizes and accelerates the adversarial learning process. **(b)(d)**: Comparison of testing accuracies of different models. **(e)**: Comparison of $\mathcal{A}$-distance of different models.

are visualized. Fig (2a) shows the representations without any adapt. As we can see, the distributions are separated between domains. This highlights the importance for transfer learning. Fig (2b) shows the result for RevGrad (Ganin & Lempitsky, 2015), a typical adversarial domain adaptation method. Features are successfully fused but it also exhibits a serious problem: features generated are near class boundary. Features of digit 1 in target domain could be easily mapped to the intermediate space between class 1 and class 2, which is obviously a damage to classification tasks. In contrast, Fig (2c) shows the representations that learned by our method. Features in the same class are mapped closer. In particular, features with different classes are dispersed, making the features more discriminative. The well-behaved learned features suggests that our model successfully pass the semantic information to the feature generator and our model is capable to learn semantic representations without any label information for target domain.

$\mathcal{A}$-distance. Based on the theory in (Ben-David et al., 2010), $\mathcal{A}$-distance is usually used to measure domain discrepancy. The empirical $\mathcal{A}$-distance is simple to compute: $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier trained with the binary classification task of discriminating the source and target. Results are shown in Fig (3e). We compared our method with domain adaptation methods RevGrad(Ganin & Lempitsky, 2015). We use a kernel SVM as the classifier. We compare our model to the standard CNN and RevGrad. From this graph, we can see that with the adversarial adaptation module embedded, our model reduces the A distances compared to CNN. But when compared to RevGrad, the results are close. This finding tells us that our semantic representation module is not focusing on reducing the global distribution discrepancy. The superior performance lead by our method shows that only reducing the global distribution discrepancy for domain adaptation is far from enough.

Convergence As our model involves the adversarial adaptation module, we testify their performance on convergence from two different aspects. The first is the testing accuracy as shown in Fig (3b)(3d). Our model has similar convergence speed as RevGrad.

Since the adversarial module in our model and RevGrad works analogous to GAN (Goodfellow et al., 2014), we will check our model from GAN's perspective. We adopt the *min-max game* in GAN. It has been proved that when the discriminator is optimal, the generator involved in the min-max game in a GAN is reducing the **Jenson-Shannon Divergence** (JSD). For the discriminator in adversarial adaptation, it is trained to maximize $\mathcal{L}_D = \mathbb{E}_{x \sim D_S}[log1 - D(x)] + \mathbb{E}_{x \sim D_T}[logD(x)]$, which is a lower bound of $2JS(D_S, D_T)$-2log2. Therefore, following (Arjovsky & Bottou, 2017), we plot the quantity of $\frac{1}{2}\mathcal{L}_D + log2$, which is the lower bound of the JS distance. Results are shown in Fig (3a)(3c). We can make following observations: **(1)** different from the vanishing generator gradient problem in traditional GANs, the manifolds where features generated by adversarial adaptation methods lies seems to be perfectly aligned. So the gradients for the feature extractor will not vanish but towards reducing the JS distance. This justifies the feasibility for adversarial domain adaptation methods. **(2)** Compared to RevGrad, our model is more stable and accelerate the minimization process for JSD. It indicates that our method stabilize the notorious unstable adversarial training through semantic alignment.

5. Conclusion

In this paper, we propose a novel method which aims at learning semantic representations for unsupervised domain adaptation. Unlike previous domain adaptation methods that solely match distribution at domain-level, we proposes to match distribution at class-level and align features semantically without any target labels. We use centroid alignment to guide the feature extractor to preserve class information for target samples in aligning domains and moving average centroid is cautiously designed to tackle the problem where a mini-batch may be insufficient for covering all class distribution in each training step. Experiments on three different domain adaptation scenarios testify the efficacy of our proposed approach.

Acknowledgements

The work described in this paper was supported by the National Key R&D Program of China2018YFB1003800), the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme 2016, the National Natural Science Foundation of China (No.61722214), and and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams(No.2016ZT06D211). We thank Peifeng Wang, Weili Chen, Fanghua Ye, Honglin Zheng and Jie Hu for discussions over the paper.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.

Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., and Bulò, S. R. Autodial: Automatic domain alignment layers. In *International Conference on Computer Vision*, 2017.

Chen, M., Weinberger, K. Q., and Blitzer, J. Co-training for domain adaptation. In *Advances in neural information processing systems*, pp. 2456–2464, 2011.

Crammer, K., Kearns, M., and Wortman, J. Learning from multiple sources. *Journal of Machine Learning Research*, 9(Aug):1757–1774, 2008.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189, 2015.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., and Li, W. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision*, pp. 597–613. Springer, 2016.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, pp. 2, 2013.

Liu, M.-Y. and Tuzel, O. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pp. 469–477, 2016.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105, 2015.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pp. 136–144, 2016.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Domain adaptation with randomized multilinear adversarial networks. *arXiv preprint arXiv:1705.10667*, 2017a.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 2208–2217, 2017b.

Luo, Z., Zou, Y., Hoffman, J., and Fei-Fei, L. F. Label efficient learning of transferable representations acrosss domains and tasks. In *Advances in Neural Information Processing Systems*, pp. 164–176, 2017.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov): 2579–2605, 2008.

Motiian, S., Jones, Q., Iranmanesh, S. M., and Doretto, G. Few-shot adversarial domain adaptation. *arXiv preprint arXiv:1711.02536*, 2017a.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017b.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 5, 2011.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Pinheiro, P. O. Unsupervised domain adaptation with similarity learning. *arXiv preprint arXiv:1711.08995*, 2017.

Saenko, K., Kulis, B., Fritz, M., and Darrell, T. Adapting visual category models to new domains. *Computer Vision–ECCV 2010*, pp. 213–226, 2010.

Saito, K., Ushiku, Y., and Harada, T. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017a.

Saito, K., Ushiku, Y., Harada, T., and Saenko, K. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017b.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017c.

Sankaranarayanan, S., Balaji, Y., and Chellappa, C. D. C. R. Generate to adapt: Unsupervised domain adaptation using generative adversarial networks.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

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

Tzeng, E., Hoffman, J., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4068–4076, 2015.

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

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.