Improving Domain-Specific Classification by Collaborative Learning with Adaptation Networks

Si Wu,[1,2] **Jian Zhong,**[1] **Wenming Cao,**[2] **Rui Li,**[2] **Zhiwen Yu,**[1] **Hau-San Wong**[2]
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
cswusi@scut.edu.cn, cszj@mail.scut.edu.cn, {wenmincao2-c, ruili52-c}@my.cityu.edu.hk,
zhwyu@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

For unsupervised domain adaptation, the process of learning domain-invariant representations could be dominated by the labeled source data, such that the specific characteristics of the target domain may be ignored. In order to improve the performance in inferring target labels, we propose a target-specific network which is capable of learning collaboratively with a domain adaptation network, instead of directly minimizing domain discrepancy. A clustering regularization is also utilized to improve the generalization capability of the target-specific network by forcing target data points to be close to accumulated class centers. As this network learns and specializes to the target domain, its performance in inferring target labels improves, which in turn facilitates the learning process of the adaptation network. Therefore, there is a mutually beneficial relationship between these two networks. We perform extensive experiments on multiple digit and object datasets, and the effectiveness and superiority of the proposed approach is presented and verified on multiple visual adaptation benchmarks, e.g., we improve the state-of-the-art on the task of MNIST→SVHN from 76.5% to 84.9% without specific augmentation.

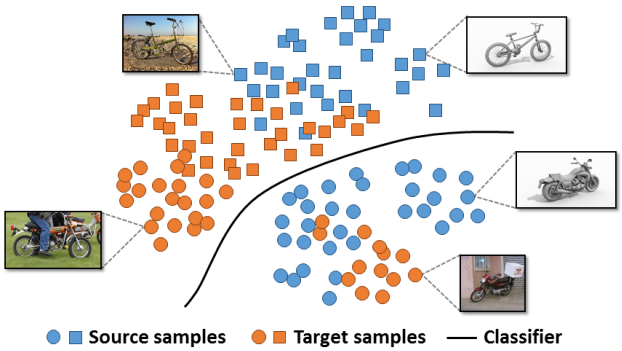

Figure 1: An example which illustrates that the training process of unsupervised domain adaptation on VisDA-2017 may be dominated by the labeled source data. The learnt representation is indiscriminative, and the resulting classifier performs unsatisfactorily on the target data, although the source samples of classes 'bicycle' and 'motorcycle' can be well distinguished.

Introduction

Sufficient amount of labeled data is vital for machine learning applications. However, it may not always be feasible to expend significant human effort for collecting and labeling data in many tasks. The objective of domain adaptation is to utilize the data in a label-rich (source) domain for inferring the class labels in a label-scarce (target) domain. In this domain adaptation setting, the source data may be sampled from a related but different distribution. How to effectively transfer knowledge learnt from source data is crucial for facilitating learning tasks in the target domain. A successful domain adaptation strategy is to learn cross-domain representations in a common space, such that the instances from different domains cannot be distinguished in the feature space. Impressive progress has been achieved, especially the adoption of deep convolutional neural networks in recent years. To guide network learning, some measures of distribution variance, e.g., Maximum Mean Discrepancy (MMD) (Long et al. 2015), are used, and the domain discrepancy can be effectively alleviated by optimizing the networks. Another way is to utilize adversarial training (Ganin and Lempitsky 2015) in the form of a minimax game between domain classification and representation learning to make domain labels as indistinguishable as possible. As a result, the classification model trained under supervision from the source can also be applied to the target domain.

Most domain adaptation methods focus either on learning projection from one domain to the other, or on learning representations that are invariant to different domains. The methods are able to reduce, but not remove the discrepancy between domains. On one hand, the distribution matching constraint may become less useful when the networks for domain-invariant feature learning have high capacities. On the other hand, it is difficult to train a single classifier having low generalization error in the target domain as well as the source domain on challenging adaptation tasks (Shu et al. 2018). In this case, the specific characteristics of the target domain is ignored due to the reason that the label information available in the source domain only leads to its associated data dominating the process of learning the projection or shared representation as illustrated in Figure 1. Since the

final objective is to infer class labels as accurately as possible on the target data rather than the source data, we propose to leverage the discriminative information from an adaptation network, and softly label the target data for learning a target-specific classification model.

In this work, we explore a collaborative learning framework between a deep adaptation network and a target-specific network for improving unsupervised domain adaptation. The proposed approach aims at training a model specialized to the target domain only, instead of directly reducing domain discrepancy. Since the labels in the target domain are unavailable, we employ a deep domain adaptation network which is capable of inducing domain confusion as a teacher for guiding a target-specific network. Specifically, these two networks are jointly optimized by minimizing the difference in class probability predictions for the same target instance. Furthermore, the discriminative capability of the target-specific network can be improved by minimizing the distances of target data points to a set of accumulated class centers. This is based on the assumption that the data points in a cluster should come from the same class. To avoid degradation, an orthogonality constraint is included to make the class centers dissimilar. Different from the existing distillation methods (Hinton, Vinyals, and Dean 2014), there is a mutually beneficial relationship between the two networks in our model. Since the target-specific network specializes to the target domain, it is able to outperform the adaptation network as it learns, and can in turn improve the performance of the adaptation network. An overview of the proposed approach is shown in Figure 2. Extensive experimental results demonstrate that collaborative learning and clustering regularization can lead to significant performance gains, and the proposed approach can achieve state-of-the-art results on multiple unsupervised domain adaptation benchmarks. This work makes the following contributions:

- ◇ Instead of directly learning on both source and target data, we exploit a deep adaptation network as a teacher to guide the training of a target-specific network on the target data only, such that the latter is able to specialize to the target domain.
- ◇ To enhance the generalization capability of the target-specific network, we introduce an effective clustering regularization to force the target data points to move to the accumulated class centers in the latent feature space.
- ◇ The adaptation and target-specific networks are trained in a collaborative learning fashion, and thus both networks can learn from each other. It is observed that the target-specific network outperforms the adaptation network in most cases.

Related Work

Domain adaptation has shown to be effective for bridging different domains (or tasks) while avoiding labor-intensive manual labeling (Pan and Yang 2010). Recently, CNN-based methods have become a mainstream technique for unsupervised visual domain adaptation. Many works focus on reducing the discrepancy between domains. Since the transferability of the learnt features reduce significantly in higher layers, Long et al. (Long et al. 2015) applied multiple kernel-based MMDs to regularizing the fully-connected layers of the networks, such that the mean embedding of data from different domains can be matched. Also based on MMD for domain confusion, Venkateswara et al. (Venkateswara et al. 2017) utilized both source and target data to learn deep hash codes to classify unseen target data. In addition to the distribution matching-based methods, reconstruction-based methods have shown effectiveness in regularizing domains. Ghifary et al. (Ghifary et al. 2016) proposed a deep reconstruction-classification network, which jointly learns supervised classification on source data and unsupervised reconstruction on target data. Based on the private-shared component hypothesis, Bousmalis et al. (Bousmalis et al. 2016) proposed domain separation networks for learning image representations through two subspaces: one which is private to each domain and another which is shared across domains. Based on the two subspaces, the images from both domains can be reconstructed. To learn a common feature space, French and Mackiewicz (French and Mackiewicz 2018) adopted a mutual learning strategy to train a student network on both source and target data under the supervision of a teacher network trained only on target data.

Maximizing domain confusion can also be achieved by utilizing adversarial training to learn representations which fool a domain classifier. Ganin et al. (Ganin and Lempitsky 2015) adopted a gradient reversal layer to train a domain classifier while learning domain-confusion representation. In (Pinheiro 2018), Pinheiro used an adversarial loss to learn domain-invariant features, and simultaneously train a similarity-based classifier by measuring the similarity of each sample to categorical prototypes. Tzeng et al. (Tzeng et al. 2017) proposed an adversarial discriminative domain adaptation method, in which a source encoder is combined with a target encoder for adversarial adaptation. To take into account the category of instances during matching marginal distributions, Saito et al. (Saito et al. 2018) utilized two task-specific classifiers to ensure that target samples are close to source samples by maximizing the classification discrepancy. In (Shu et al. 2018), Shu et al. adopted a virtual adversarial domain adaptation model to learn domain-confusion representations in the first stage, and then refined decision boundaries by minimizing the conditional entropy on target instances. To learn data distributions and generate samples across domains, Liu and Tuzel (Liu and Tuzel 2016) proposed a coupled GAN to learn the joint distribution of multi-domain data. Sankaranarayanan et al. (Sankaranarayanan et al. 2018) adopted an adversarial image generation model to address the domain-mismatch in the feature space learnt by an encoder, such that the target embeddings can be used to generate source-like images. In (Russo et al. 2018), Russo et al. proposed a bi-directional GAN to generate target-like images from the source and source-like images from the target.

In contrast to the methods above, we design a target-specific network which specializes to the target domain and learns from a domain adaptation network, instead of directly reducing domain discrepancy. A clustering regularization is formulated for improving the generalization capability of the

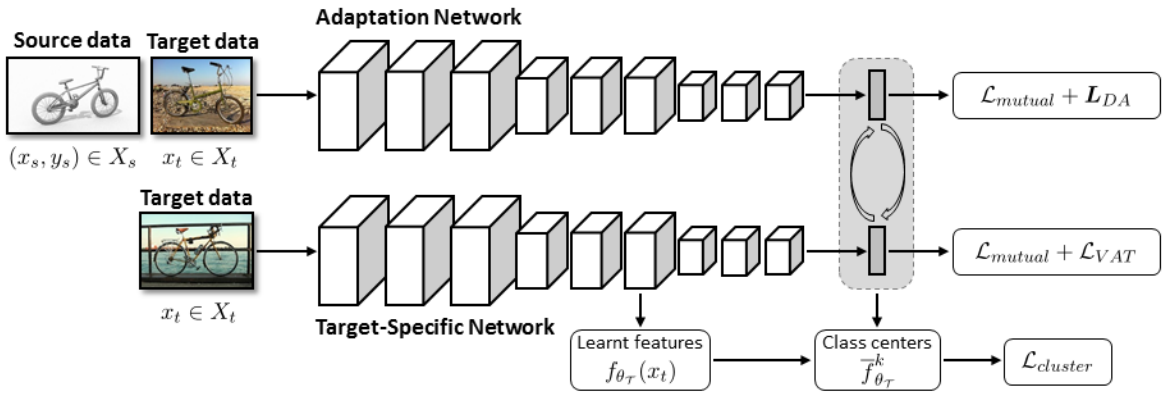

Figure 2: Overview of the proposed approach to train a target-specific network via collaborative learning with a domain adaptation network. The target-specific network learns to predict target labels only through the supervision of the adaptation network output and the accumulated class centers, so that it is able to specialize to the target domain without directly minimizing the domain discrepancy.

target-specific network. As the performance of this network improves, it will in turn help the adaptation network through collaborative learning.

Deep Adaptation Networks

Before describing the proposed collaborative learning framework, we first introduce an adversarial domain adaptation network (Shu et al. 2018), which plays an important role in our framework. For unsupervised domain adaptation, there are labeled source data $(x_s, y_s) \in X_s$ and unlabeled target data $x_t \in X_t$. For the purpose of low source classification error and small feature divergence between domains, the overall loss function of this domain adaptation model is defined as follows:

$$\begin{aligned}\boldsymbol{L}_{DA}(\theta; X) = \mathcal{L}_{crosEnt}(\theta; X_s) + \mathcal{L}_{condEnt}(\theta; X_t) \\ + \mathcal{L}_{domain}(\theta; X) + \mathcal{L}_{VAT}(\theta; X_t),\end{aligned} \tag{1}$$

where $X = X_s \cup X_t$, θ denotes the parameters of the network, $\mathcal{L}_{crosEnt}$ denotes the standard cross entropy function on labeled source data, and $\mathcal{L}_{condEnt}$ denotes the conditional entropy function on unlabeled target data. To learn domain-confusion representations, the third term $\mathcal{L}_{domain}$ is formulated as follows:

$$\begin{aligned}\mathcal{L}_{domain}(\theta; X) = \lambda_d \sup_{\omega} \Bigg(\sum_{x_s \in X_s} [\ln h_\omega(f_\theta(x_s))] \\ + \sum_{x_t \in X_t} [\ln(1 - h_\omega(f_\theta(x_t)))] \Bigg),\end{aligned} \tag{2}$$

where $f_\theta(\cdot)$ denotes the representations learnt by the network, and $h_\omega(\cdot)$ denotes a domain discriminator parameterized by ω. Domain adversarial training is to minimize $\mathcal{L}_{domain}$ via mini-max optimization, which enforces the network to learn the representation which reduces the divergence between source and target domains. In order to obtain a reliable empirical estimate of conditional entropy, the term $\mathcal{L}_{VAT}$ for virtual adversarial training is defined as follows:

$$\mathcal{L}_{VAT}(\theta; X_t) = \lambda_v \sum_{x_t \in X_t} \left(\max_{\|\gamma\| \leq \epsilon} \mathbb{D}(p_\theta(x_t) \| p_\theta(x_t + \gamma)) \right), \tag{3}$$

where $p_\theta(\cdot)$ denotes the predicted class probability distribution of the network, $\mathbb{D}$ denotes the Kullback-Leibler (KL) divergence between two distributions, and ϵ is a hyperparameter controlling the intensity of the adversarial perturbation γ. As a result, adversarial samples are generated by including small perturbations to the input in the direction sensitive to the model prediction. Virtual adversarial training can be applied on both source and target data. In Eqs.(2-3), λ_d and λ_v are the weighting factors for achieving a balance among the terms in $\boldsymbol{L}_{DA}$.

Training Target-Specific Networks

The optimal classification model for source data does not often coincide with that for target data. Since the adaptation network is trained under the main supervision from the category information of the source data, we consider that a better classification model is achievable by introducing another network (target-specific network) specialized to the target data. Since the labels of target instances are unavailable, we adopt a collaborative learning framework to mutually reinforce the consistency of the two network predictions.

Our adaptation network parameterized by $\theta_\mathcal{A}$ is built based on the model described in Section III, and the overall loss function $\boldsymbol{L}_\mathcal{A}$ is defined as the sum of the loss function defined in Eq.(1) and an additional regularization term as follows:

$$\boldsymbol{L}_\mathcal{A}(\theta_\mathcal{A}; X) = \mathcal{L}_{mutual}(\theta_\mathcal{A}; X_t, \theta_\mathcal{T}) + \boldsymbol{L}_{DA}(\theta_\mathcal{A}; X), \tag{4}$$

where the target-specific network is parameterized by $\theta_\mathcal{T}$, and we define the term of mutual learning $\mathcal{L}_{mutual}$ as the expected similarity between the predictions of the adaptation

and target-specific networks as follows:

$$\mathcal{L}_{mutual}(\theta_{\mathcal{A}}; X_t, \theta_{\mathcal{T}}) = \lambda_m \sum_{x_t \in X_t} \mathbb{S}(p_{\theta_{\mathcal{A}}}(x_t); p_{\theta_{\mathcal{T}}}(x_t)), \tag{5}$$

where $\mathbb{S}(\cdot;\cdot)$ denotes a similarity measure between two distributions, such as the KL divergence, cross entropy, mean square error, and so on.

The target-specific network is a separate network which has the same architecture as the adaptation network. In order to avoid error propagation during pseudo-labeling in the target domain, the predictions produced by the adaptation network are used as soft labels to guide the training of the target-specific network. Adversarial training for domain-confusion is no longer necessary due to the reason that this network is trained on the target data only. More formally, the loss function $\boldsymbol{L}_{\mathcal{T}}$ for the target-specific network can be defined as follows:

$$\begin{aligned}\boldsymbol{L}_{\mathcal{T}}(\theta_{\mathcal{T}}; X_t) = \mathcal{L}_{mutual}(\theta_{\mathcal{T}}; X_t, \theta_{\mathcal{A}}) \\ + \mathcal{L}_{cluster}(\theta_{\mathcal{T}}, \mathcal{C}; X_t) + \mathcal{L}_{VAT}(\theta_{\mathcal{T}}; X_t),\end{aligned} \tag{6}$$

where $\mathcal{L}_{mutual}(\theta_{\mathcal{T}}; X_t, \theta_{\mathcal{A}})$ denotes the mutual learning term enforcing the target-specific network to learn from the adaptation network by using its predictions as soft labels, $\mathcal{L}_{cluster}$ denotes a clustering term encouraging the target data points to be close to the class centers $\mathcal{C}$, and the virtual adversarial term $\mathcal{L}_{VAT}$ helps to stabilize empirical estimates of class probabilities for target data. Specifically, $\mathcal{L}_{mutual}(\theta_{\mathcal{T}}; X_t, \theta_{\mathcal{A}})$ is defined as follows:

$$\mathcal{L}_{mutual}(\theta_{\mathcal{T}}; X_t, \theta_{\mathcal{A}}) = \lambda_m \sum_{x_t \in X_t} \mathbb{S}(p_{\theta_{\mathcal{T}}}(x_t); p_{\theta_{\mathcal{A}}}(x_t)). \tag{7}$$

Different from $\mathcal{L}_{mutual}(\theta_{\mathcal{A}}; X_t, \theta_{\mathcal{T}})$, the class probability prediction $p_{\theta_{\mathcal{T}}}(\cdot)$ in $\mathcal{L}_{mutual}(\theta_{\mathcal{T}}; X_t, \theta_{\mathcal{A}})$ will be improved under fixed $p_{\theta_{\mathcal{A}}}(\cdot)$. Also, different from minimization of the conditional entropy which moves the decision boundaries to low-density regions, our network encourages the unlabeled target data points to form tighter clusters, because the clustering assumption that the data points of the same class should concentrate together hold in most cases. For this purpose, the clustering regularization term is defined as follows:

$$\begin{aligned}\mathcal{L}_{cluster}(\theta_{\mathcal{T}}, \mathcal{C}; X_t) = \lambda_c \Bigg(\sum_{x_t \in X_t} \sum_{k=1}^{K} \delta(\widehat{y}_t^{\mathcal{A}}, k) \left\| f_{\theta_{\mathcal{T}}}(x_t) - c_k \right\|^2 \\ + \mu \| \mathcal{C}^T \mathcal{C} - I \|_F^2 \Bigg),\end{aligned} \tag{8}$$

where

$$\delta(\widehat{y}_t^{\mathcal{A}}, k) = \begin{cases} 1 & \text{if } \widehat{y}_t^{\mathcal{A}} = k, \\ 0 & \text{if } \widehat{y}_t^{\mathcal{A}} \neq k, \end{cases}$$

$\widehat{y}_t^{\mathcal{A}}$ denotes the predicted class of sample x_t by the adaptation network, $f_{\theta_{\mathcal{T}}}$ denotes the learnt representation on the middle hidden layer of the target-specific network, $\mathcal{C} = [c_1, c_2, \ldots, c_K]$ denotes a matrix with its columns corresponding to the accumulated class centers, $\| \cdot \|_F^2$ denotes the squared Frobenius norm, and I denotes the identity matrix. We require $\mathcal{C}^T\mathcal{C}$ to be close to I, such that the class centers become orthogonal. Minimizing this clustering regularization term ensures that the class centers will be dissimilar to each other and each target data point moves to a specific class center. In each mini-batch based update step, the accumulated class centers are also updated as follows:

$$c_k \leftarrow \alpha c_k + \frac{1-\alpha}{m_k} \sum_{x_t \in B_t} \delta(\widehat{y}_t^{\mathcal{A}}, k) f_{\theta_{\mathcal{T}}}(x_t), \tag{9}$$

where B_t denotes a mini-batch of target instances, and m_k denotes the number of the target instances predicted as class k by the adaptation network. As a result, the target-specific network is able to learn more discriminative representations on the target data, due to the reason that the adaptation network provides training targets as well as relationship information among target instances for representing the underlying structure.

The adaptation and target-specific networks are jointly trained in our collaborative learning framework. These two networks learn from each other to correctly predict the true labels of target samples, as well as to match their predictions. During this training process, the target-specific network tends to produce more accurate results than the adaptation network because of its specialization to the target domain. In turn, we can take advantage of this improved accuracy during collaborative learning to guide the adaptation network by providing better targets. The implementation details are summarized in Algorithm 1.

Algorithm 1 Pseudo-code of collaborative learning between adaptation and target-specific networks.

Input: Labeled source data X_s and unlabeled target data X_t.
Initialize: Adaptation network $\theta_{\mathcal{A}}$ and target-specific network $\theta_{\mathcal{T}}$ with different initial conditions, learning rates $\zeta_{\mathcal{A}}$ and $\zeta_{\mathcal{T}}$, and class centers $\mathcal{C}$.
for $n = 1$ to N **do**
  Sample mini-batches B_s from X_s and B_t from X_t.
  for each mini-batch B_s and B_t **do**
    Evaluate adaptation network predictions $p_{\theta_{\mathcal{A}}}(x_s)$ and $p_{\theta_{\mathcal{A}}}(x_t)$, and target-specific network predictions $p_{\theta_{\mathcal{T}}}(x_s)$ and $p_{\theta_{\mathcal{T}}}(x_t)$.
    Update adaptation network: $\theta_{\mathcal{A}} \leftarrow Adam(\nabla_{\theta_{\mathcal{A}}} \boldsymbol{L}_{\mathcal{A}}, \theta_{\mathcal{A}}, \zeta_{\mathcal{A}})$.
    Compute the mid-level representations $f_{\theta_{\mathcal{T}}}(x_t)$.
    Update class centers c_k according to Eq.(9).
    Update target-specific network: $\theta_{\mathcal{T}} \leftarrow Adam(\nabla_{\theta_{\mathcal{T}}} \boldsymbol{L}_{\mathcal{T}}, \theta_{\mathcal{T}}, \zeta_{\mathcal{T}})$.
  end for
end for
Return $\theta_{\mathcal{A}}$ and $\theta_{\mathcal{T}}$.

Experiments and Discussion

In the experiments, we focus on visual domain adaptation tasks, and perform an extensive evaluation of the pro-

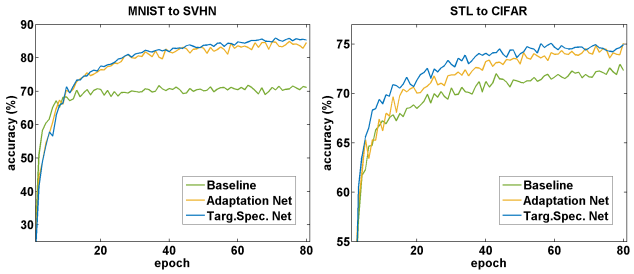

Figure 3: The training curves of the baseline model and the two networks in the proposed model on the MNIST→SVHN (left) and STL→CIFAR (right) tasks. The adaptation and target-specific networks have very similar performance, and significantly outperform the baseline model due to collaborative learning.

Figure 4: The t-SNE plots of the middle hidden layers of the baseline network (left) and target-specific network (right) in the proposed model on the task of MNIST→SVHN. In the right subfigure, each black circle denotes an accumulated class center, and the embedding of each class forms a relatively tight cluster. We use this comparison with the baseline network to highlight the improvement brought by the clustering regularization to the target-specific network.

posed approach on multiple standard adaptation tasks by using the digit datasets: MNIST (LeCun et al. 1998), USPS (Hull 1994), Syn-Digits (Ganin and Lempitsky 2015) and SVHN (Netzer et al. 2011), the object datasets: CIFAR-10 (Krizhevsky and Hinton 2009) and STL-10 (Coates, Lee, and Ng 2011), and the VisDA-2017 dataset (Peng et al. 2017). We use the standard evaluation protocol (Shu et al. 2018) for all the tasks, and report a comparison with state-of-the-art unsupervised domain adaptation methods.

For adaptation between the digit and object datasets, we adopt a relatively small CNN architecture, which is the same as in (Shu et al. 2018). For the experiments on the VisDA-2017 dataset, we use a larger CNN architecture, ResNet-50 (He et al. 2016), due to the high resolution of the training images. In our collaborative learning framework, the adaptation and target-specific networks have the same architecture but with different initializations and dropout. At each training step, two mini-batches of source and target instances are drawn from their respective datasets. We adopt the ADAM optimization method (Kingma and Ba 2015) to update the networks.

Results on Digit and Object Datasets

Comparison to Previous Work. We first compare the proposed model with state-of-the-art unsupervised domain adaptation methods in this sub-section. The results of the competed methods are shown in Table 1. Most of the competing methods were tested on three of the tasks: MNIST→USPS, USPS→MNIST and SVHN→MNIST. Since the discrepancy between MNIST and USPS is relatively small and SVHN is more challenging than MNIST, recent methods, such as 'DIRT-T' and 'Self-Ensembling', have high accuracies on the three tasks. The proposed approach also achieves comparable results in these tasks.

Compared with the above three tasks, the remaining tasks including SVHN→MNIST, DIGITS→SVHN and STL→CIFAR are more difficult and important. A few of the competing methods had been tested on these tasks. The MNIST→SVHN adaptation performance of the proposed approach is the best in the comparison. Specifically, the accuracy of our approach attains 84.9%, which is higher than 'DIRT-T' and 'Self-Ensembling' by about 8.4 and 42.9 percentage points, respectively. On the task of DIGITS→SVHN, we achieve an accuracy of 96.4%, which is the highest in the comparison. In addition, the proposed approach achieves an accuracy of 75.4% on the task of STL→CIFAR, and outperforms 'Self-Ensembling' by about 6.2 percentage points. The main reason the proposed approach outperforms the competing methods on these difficult adaptation tasks is that the collaborative learning and clustering regularization play an important role in training the target-specific network by specializing the learning process to the target domain.

Comparison to Baseline and Model Variants. We perform experimental analysis to better understand the proposed approach. We report the performance of the 'Source Only' model using the same network architecture as the proposed approach but without adaptation. The 'Source Only' results, which are the accuracies obtained by training the CNN in a supervised fashion on the source and being applied to the target data directly, serve as the lower bound. In addition, we train the adaptation network separately, and consider the resulting model as 'Baseline'. To investigate the performance improvements brought by the proposed collaborative learning, we build a variant of our model by disabling the clustering regularization 'Our Model w/o CR'. As expected, both 'Baseline' and 'Our Model w/o CR' outperform the 'Source Only' model on all the adaptation tasks. When adopting collaborative learning, 'Our Model w/o CR' performs better than 'Baseline' in most cases. On the difficult adaptation task of MNIST→SVHN, 'Our Model w/o CR' reaches an accuracy of 83.9%, which is much higher than those of 'Source Only' and 'Baseline' by about 43.7 and 12.3 percentage points, respectively. The improvement margins over these two models are significant, which demonstrates the effectiveness of our proposed collaborative learning. We consider that the target-specific network guided by the adaptation network can achieve better performance. We also investigate the relative contribution of the clustering regularization in Table 1. For this purpose, we compare our model with its variant 'Our Model w/o CR'. When including the clustering regularization term in the joint loss function,

Table 1: Classification accuracy (%) of the proposed approach and previous works in the experiments of unsupervised domain adaptation on the digit and object datasets. w/o CR indicates 'without clustering regularization'.

Method	MNIST→USPS	USPS→MNIST	SVHN→MNIST	MNIST→SVHN	DIGITS→SVHN	STL→CIFAR
MMD(Long et al. 2015)	81.1	-	71.1	-	88.0	-
RevGrad(Ganin et al. 2016)	91.1	74.0	74.0	35.7	-	57.0
DANN(Ganin and Lempitsky 2015)	77.1	73.0	71.1	35.7	90.3	-
DRCN(Ghifary et al. 2016)	91.8±0.1	73.7±0.1	82.0±0.2	40.1±0.1	-	58.9±0.1
DSN(Bousmalis et al. 2016)	91.3	-	82.7	-	91.2	-
ADDA(Tzeng et al. 2017)	89.4±0.2	90.1±0.8	76.0±1.8	-	-	-
CoGAN(Liu and Tuzel 2016)	91.2±0.8	89.1±0.8	-	-	-	-
ATT(Saito, Ushiku, and Harada 2017)	-	-	86.2	52.8	92.9	-
PixelDA(Bousmalis et al. 2017)	95.9	-	-	-	-	-
TRUDA(Sener et al. 2016)	-	-	78.8	40.3	-	-
UNIT(Liu, Breuel, and Kautz 2017)	95.9	93.5	90.5	-	-	-
AssocDA(Haeusser et al. 2017)	-	-	95.7±1.5	-	91.3±0.2	-
GenToAdapt(Sankaranarayanan et al. 2018)	92.5±0.7	90.8±1.3	84.7±0.9	36.4±1.2	-	-
CyCADA(Hoffman et al. 2018)	94.8±0.2	95.7±0.2	88.3±0.2	-	-	-
MCDUDA(Saito et al. 2018)	94.2±0.7	94.1±0.3	96.2±0.4	-	-	-
SimNet(Pinheiro 2018)	96.4	95.6	-	-	-	-
SBADA-GAN(Russo et al. 2018)	97.6	95.0	76.1	61.1	-	-
DIRT-T(Shu et al. 2018)	-	-	**99.4**	76.5	96.2	73.3
Self-Ensembling(French and Mackiewicz 2018)	**98.3±0.1**	**99.5±0.1**	99.2±0.3	42.0±5.7	96.0±0.1	69.2±0.4
Source Only	92.9±0.2	90.0±3.5	78.8±2.0	40.2±0.6	83.8±0.9	61.5±0.9
Baseline	97.7±0.1	98.8±0.4	97.2±1.3	71.6±0.5	95.5±0.4	71.9±0.5
Our Model w/o CR	97.7±0.1	99.2±0.1	99.0±0.3	83.9±0.6	96.3±0.1	74.5±0.9
Our Model	98.0±0.2	99.4±0.1	99.3±0.1	**84.9±0.4**	**96.4±0.2**	**75.4±0.2**

Figure 5: Visualization of the accumulated class centers learnt by the proposed model on the task of SVHN→MNIST. One can observe that each center represents the prototype of a specific class, which confirms that the clustering regularization is beneficial to exploring target data structures.

the performance is improved in all the cases. On the tasks of MNIST→SVHN and STL→CIFAR, the accuracies of the proposed approach are increased to 84.9% and 75.4%, respectively. We consider that the clustering regularization is of benefit to improving the network generalization capability.

Visualization. To further illustrate the process of collaborative learning, Figure 3 shows the accuracy curves of the adaptation and target-specific networks during training, from which we can observe the relationship between these two networks on the tasks of MNIST→SVHN and STL→CIFAR. Compared to the baseline, the accuracies of both adaptation and target-specific networks are improved as collaborative learning proceeds. It is noted that collaborative learning consistently improves the target-specific network. Since only target data is fed to this network, it specializes to the target domain and produces better predictions than the adaptation network. This result confirms that specialization to the target can lead to better adaptation.

In addition, we analyze the effectiveness of the clustering regularization by showing the t-SNE (Maaten and Hinton 2008) embedding of the learnt features of our network when adapting from MNIST to SVHN. Figure 4 visualizes the features of the target instances on the middle hidden layers of the 'Baseline' network and the target-specific network in our model, respectively. Each class is encoded by a color, and the corresponding accumulated class center is shown as a black circle. For the 'Baseline' network, the target data points are not clustered strongly, due to the reason that this network is trained without any clustering regularization. In contrast, our target-specific network is able to improve clustering of the target data points in the latent feature space. As expected, the accumulated class centers lie in the center regions of the clusters, and the target data points are better separated.

To further demonstrate what accumulated class centers represent, we visualize them via an auxiliary network mapping the data points in the latent space to the corresponding original images. For simplicity, this network consists of four deconvolution layers reconstructing a 32×32 grey-scale images x, conditioned on the representation $f_{\theta_T}(x)$. Since the MNIST images have simple background facilitating visualization, we train this auxiliary network to visualize the class centers learnt by the our model on the task of SVHN→MNIST, and the result is shown in Figure 5. It is noted that the resulting class centers represent a set of prototypes for the classes. We consider that these visualized centers directly reveal the reason of the clustering regularization stabilizing and improving the classification performance of our model in the target domain.

Results on VisDA-2017

Comparison to Previous Work. Since the resolution of the VisDA-2017 images is substantially higher than that of the

Table 2: Classification accuracy (%) of the proposed approach and previous works in the experiment of unsupervised domain adaptation on the VisDA-2017 dataset. w/o CR indicates 'without clustering regularization'.

Method	plane	bcycl.	bus	car	horse	knife	mcycl.	person	plant	sktbrd.	train	truck	**mean**
MMD(Long et al. 2015)	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
RevGrad(Ganin et al. 2016)	75.9	70.5	65.3	17.3	72.8	38.6	58.0	77.2	72.5	40.4	70.4	44.7	58.6
DANN(Ganin and Lempitsky 2015)	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
SimNet(Pinheiro 2018)	94.5	80.2	69.5	43.5	89.5	16.6	76.0	81.1	86.4	76.4	79.6	41.9	69.6
Self-Ensembling(French and Mackiewicz 2018)	94.9	84.1	71.1	40.9	88.9	43.6	64.6	73.2	87.1	64.5	83.7	47.9	70.4
MCDUDA(Saito et al. 2018)	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
Source Only	78.8	57.5	54.4	52.5	79.9	21.9	82.7	58.8	86.5	51.6	86.5	36.9	62.3
Baseline	72.7	73.4	72.7	49.7	86.9	33.4	80.0	71.8	87.3	66.0	78.4	37.4	67.5
Our Model w/o CR	91.5	83.5	77.0	48.5	91.9	41.3	85.0	75.0	83.5	63.7	79.3	42.6	71.9
Our Model	91.5	80.4	82.3	59.2	90.1	62.9	85.3	75.0	87.3	70.6	78.5	28.4	**74.3**

images in the datasets used in the previous experiments, we use a larger CNN architecture, ResNet-50, in our model, and most competing methods are also based on the same backbone architecture in this experiment. Due to the high resolution, we reduce the batch size to 44. Table 2 shows the performance of different unsupervised adaptation methods tested on the VisDA-2017 task. The results reported are in the form of the average per-class classification accuracy and the corresponding mean over all the classes. In this comparison, we train the proposed model with the minimum augmentation (French and Mackiewicz 2018), and present the result achieved by the target-specific network in our model as it outperforms the adaptation network. 'SimNet' and 'Self-Ensembling' perform better than the other competing methods, and the proposed approach achieves the best performance. In some difficult classes such as 'knife' and 'sktbrd.', 'SimNet', 'Self-Ensembling' and 'MCDUDA' perform less satisfactory, and the proposed model produces more accurate prediction.

Comparison to Baseline and Model Variants. In addition, we also report the results of the variants of the proposed model including 'Source Only', 'Baseline', and 'Our Model w/o CR'. As expected, 'Baseline', the proposed model and its variant 'Our Model w/o CR' improve the 'Source Only' model and achieve higher classification accuracies on all the class, and the overall average accuracy is improved by about 5.2, 12 and 9.6 percentage points, respectively. The significant performance gains are attributable to domain adaptation. The proposed model also outperforms 'Baseline' and 'Our Model w/o CR' on most classes, since the collaborative learning and clustering regularization lead to more accurate classification in the target domain. Furthermore, we compare the adaptation and target-specific networks of our model in Figure 6. It is noted that the target-specific network is able to achieve higher classification accuracies than the adaptation network on most classes, which confirms that the specialization of the target-specific network to the target domain benefits unsupervised domain adaptation.

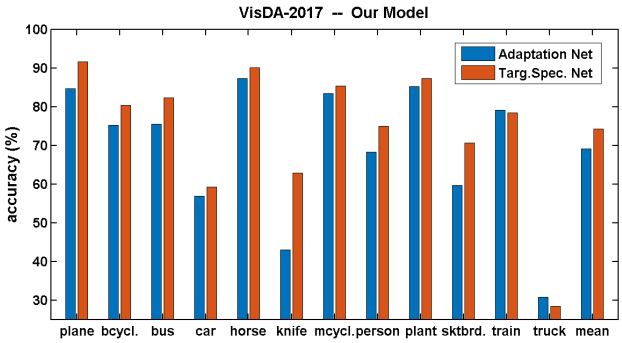

Figure 6: Per-class average classification accuracy and the corresponding mean over all the classes of the adaptation and target-specific networks of the proposed model in the experiment on the VisDA-2017 dataset. The target-specific network outperforms the adaptation network on most classes due to its specialization to the target domain.

Conclusion

In this work, we explore how to improve unsupervised domain adaptation without directly reducing the discrepancy between domains with different data distributions. To learn a target-specific network, we exploit a separate adaptation network to guide its training on the target data. The clustering regularization, which enforces target data points to be close to accumulated class centers in the latent feature space, is applied to further improve the generalization capability of the target-specific network. Since this network specializes to the target domain, it captures the domain-specific information to facilitate more effective inference of class labels on the target data. In addition, we adopt a collaborative learning framework to mutually reinforce the performance of the target-specific and adaptation networks. The effectiveness of the proposed approach is verified by extensively evaluating the performance improvement in inferring target labels on multiple visual adaptation benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 61502173, U1611461), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11300715), in part by City University of Hong Kong (Project No. 7004884), in part by the Natural Science Foundation of Guangdong Province (Project No. 2016A030310422), and in part by the Fundamental Research Funds for the Central Universities (Project No. 2018ZD33).

References

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

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 95 – 104.

Coates, A.; Lee, H.; and Ng, A. 2011. An analysis of single layer networks in unsupervised feature learning. In *AISTATS*.

French, G., and Mackiewicz, M. 2018. Self-ensembling for domain adaptation. In *ICLR*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180 – 1189.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain adversarial training of neural networks. *JMLR* 17:1 – 35.

Ghifary, M.; Kleijn, W.; Zhang, M.; Balduzzi, D.; and Li, W. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 597 – 613.

Haeusser, P.; Frerix, T.; Mordvintsev, A.; and Cremers, D. 2017. Associative domain adaptation. In *ICML*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770 – 778.

Hinton, G.; Vinyals, O.; and Dean, J. 2014. Distilling the knowledge in a neural network. In *NIPS Workshop on Deep Learning and Representation Learning*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.; Isola, P.; Efros, A.; and Darrel, T. 2018. CyCADA: cycle-consistent adversarial domain adaptation. In *ICML*.

Hull, J. 1994. A database for handwritten text recognition research. *IEEE TPAMI* 16(5):550 – 554.

Kingma, D., and Ba, J. 2015. Adam: a method for stochastic optimization. In *ICLR*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. In *Univ. Toronto, Toronto, ON, Canada, Tech. Rep.*

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

Liu, M., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*, 469 – 477.

Liu, M.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NIPS*, 700 – 708.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97 – 105.

Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR* 9(11):2579 – 2605.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

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

Peng, X.; Usman, B.; Kaushik, N.; Hoffman, J.; Wang, D.; and Saenko, K. 2017. VisDA: the visual domain adaptation challenge. In *arXiv:1710.06924*.

Pinheiro, P. 2018. Unsupervised domain adaptation with similarity learning. In *CVPR*.

Russo, P.; Carlucci, F.; Tommasi, T.; and Caputo, B. 2018. From source to target and back: symmetric bi-directional adaptive GAN. In *CVPR*.

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*.

Saito, K.; Ushiku, Y.; and Harada, T. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2988 – 2997.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C.; and Chellappa, R. 2018. Generate to adapt: aligning domains using generative adaversarial networks. In *CVPR*.

Sener, O.; Song, H.; Saxena, A.; and Savarese, S. 2016. Learning transferable representations for unsupervised domain adaptation. In *NIPS*, 2110 – 2118.

Shu, R.; Bui, H.; Narui, H.; and Ermon, S. 2018. A DIRT-T approach to unsupervised domain adaptation. In *ICLR*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrel, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 2962 – 2971.

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 5385 – 5394.