TMDA: Task-Specific Multi-Source Domain Adaptation via Clustering Embedded Adversarial Training

Haotian Wang*, Wenjing Yang*, Zhipeng Lin*, Yue Yu†‡
*State Key Laboratory of High Performance Computing (HPCL), National University of Defense Technology, China
†National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, China
‡Peng Cheng Laboratory, Shenzhen, China
{wanghaotian13,wenjing.yang,linzhipeng13,yuyue}@nudt.edu.cn

***Abstract*—Beyond classical domain-specific adversarial training, a recently proposed task-specific framework has achieved a great success in single source domain adaptation by utilizing task-specific decision boundaries. However, compared to single-source-single-target setting, multi-source domain adaptation (M-DA) shows more powerful capability to handle with most real-life cases. To align target domain with diverse multi-source domains using task-specific decision boundaries, we provide a deep insight of task-specific framework on MDA for the first time. Accordingly, we propose a novel task-specific multi-source domain adaptation method (TMDA) with a clustering embedded adversarial training process. Specifically, the proposed TMDA detects and refines less discriminative target representations through a max-min optimization over two adversarial task-specific classifiers. Moreover, our analysis implies that scattered multi-source representations disturb the adversarial training under the task-specific framework. To tight up the dispersed source representations, we embeds a relationship-based domain clustering into TMDA. Empirical results demonstrate that our TMDA outperforms state-of-the-art methods on toy dataset, sentiment analysis and digit classification.**

***Index Terms*—Multi-source domain adaptation, Adversarial Training, Task-specific, Domain Clustering**

Despite the rapid developments in domain adaptation, most existing methods transfer knowledge from single *source domain* to single *target domain* [1]–[3]. However, the realistic source data commonly possesses an underlying multi-mode structure, which tends towards being sampled from different resources [4], [5]. Therefore adaptation from single source domain cannot fit most real-life cases. Consequently, *unsupervised multi-source domain adaptation* (MDA) methods have been actively researched due to their flexibility and practicality in realistic applications [6], [7]. Since the adversarial training has achieved a great success in single source domain adaptation, a number of recent studies have extended the adversarial framework to MDA [8]. More specifically, most existing MDA methods train multiple domain discriminators to distinguish which domain the features belong to [6], together with training a feature extractor to mimic the discriminators. When the training process converges, the model can well generate and distinguish feature representations from different domains (We call this kind of framework as *domain-specific* throughout this paper.)

However, simply aligning the features via the domain boundaries leads to the mismatch of source and target domains in the class level [9]. In cases of single-source domain adaptation, a classifier-based adversarial training method has been proposed to address this issue (We call this kind of framework as *task-specific* throughout this paper), while the multi-source domain adaptation situation becomes much more complicated under these circumstances. As shown in Fig 1(b), the scatter of multi-source representations degrades the effectiveness of task-specific adversarial training.

In order to align the multi-source domains with target domain by utilizing task-specific decision boundaries rather than domain boundaries, we provide a deep insight of task-specific framework on MDA for the first time. Accordingly, we propose a novel task-specific multi-source domain adaptation (TMDA) method with a clustering embedded adversarial training process in this paper. More specifically, the proposed TMDA trains the feature extractor for feature embedding and trains two task-specific classifiers for class prediction. We then present the adversarial training between the two classifiers and the feature extractor: (a) Maximizing the classifier discrepancy via the optimization over two classifiers to estimate less discriminative target representations that are away from the support of multi-source domains (b) Minimizing the discrepancy via the optimization over the feature extractor to push target representations close to source manifolds. To provide a flexible and accurate measurement of the divergence between two classifiers, we combined the sliced Wasserstein distance to compute the classifier discrepancy [10]. Nevertheless, our analysis implies that the scatter of multi-source representations degrades the accuracy of the max-min optimization on classifier discrepancy. To tight up the multi-source representations, we embeds a relationship-based domain clustering into the adversarial training process by utilizing a domain adaptor [8] in TMDA. With the clustering embedding, we decrease the gap across multi-source representations in the embedding space,

This work was partially supported by the National Key Research and Development Program of China (No. 2017YFB1001900), the Youth Talent Lifting Project (No.17-JCJQ-QT-047), the National Natural Science Foundation of China (No. 91648204), and the National Science and Technology Major Project.

which guarantees the performance of task-specific adversarial training. Moreover, we evaluate the proposed TMDA on sentiment analysis task and digits classification task, respectively. Empirical results demonstrate that TMDA achieves the state-of-the-art performance on benchmarks.

To summarize, we present our contributions as follows:

- To our best knowledge, we firstly incorporate multi-source domain adaptation into the task-specific framework and achieve a better feature alignment between multi-source and target domains.
- We propose a novel task-specific multi-source domain adaptation method with the clustering embedded adversarial training.
- Experiments on text analysis and image classification tasks verify the proposed method with the state-of-the-art performance.

I. Related Work

Drawbacks of Domain Adversarial Training Although domain-specific adversarial training framework aims to achieve a low target risk [11], researchers have pointed out one of its drawback that domain-specific adversarial training [12] focuses on distinguishing features by domain labels rather than class labels [9]. Accordingly, based on the above observations, Maximum Classifier Discrepancy (MCD) was proposed to build the task-specific framework by two classifiers instead of single domain discriminator [9]. However, this framework is incapable of MDA in some cases, as scattered representations from multi-source domains disturb the classifier discrepancy computation (Fig. 1). To tackle this issue, the proposed TMDA embeds the domain clustering mechanism via a domain adaptor to tight up the dispersed source features.

Unsupervised Multi-source Domain Adaptation Instead of domain specific framework adopted by most MDA methods, the proposed TMDA bridges the multiple source domains and the target domain by utilizing a novel task-specific framework. In addition, it is worth noting that a multi-domain matching framework was proposed with a embedded relation module [8], which provides the theoretical support for domain clustering in our proposed TMDA.

II. Task-specific Adversarial Training Model

In this section, we propose a novel task-specific multi-source domain adaptation model. First we make analysis via comparing the domain-specific framework with the task-specific framework for MDA in Section A. Then we formulate the loss functions and propose our model architecture in Section B.

A. Domain-Specific or Task-Specific: From Single Source Domain to Multi-source Domain

Assume we have access to labeled samples $\{(\mathbf{x}_s, y_s)\}$ from K source domains $S_1, S_2, \ldots, S_K$ and unlabeled samples $\{\mathbf{x}_t\}$ from single target domain T. Meanwhile, samples in each domain is classified into C classes. Moreover, the unsupervised multi-source domain adaptation aims to achieve desirable performance on T with a model trained on $S_1, S_2, \ldots, S_K, T$.

Most existing single-source domain adaptation methods tend to perform domain-specific framework, which contains a feature extractor F, a domain discriminator G and a classifier C [12]. Since the adversarial training process is conducted between F and G, domain-specific framework focuses on generating and distinguishing features in domain level. In contrary, task-specific framework shifted the training focus to features with class labels rather than domain labels [9]. More specifically, two task-specific classifiers C_1 and C_2 are proposed, together with a feature extractor F. On the one side, the discrepancy of C_1 and C_2 are maximized to detect the target samples away from the source support. On the other side, F is optimized to minimizing the discrepancy by pushing target representations close to source manifolds.

Existing studies have revealed that task-specific framework achieves better performance than domain-specific framework for single source domain adaptation [9], [12]. However, the situation becomes much more complicated when multi-source domains exist. Is task-specific framework necessary for MDA? Moreover, if it is necessary, how should we incorporate task-specific framework into MDA? To solve these two questions, we compare and analyze domain-specific framework and task-specific framework in cases where multi-source domains exist.

In combination with the annotations before, we build an intuitive example in Fig. 1 using the case $K = 2$ for analysis. Primarily, we answer the first question by analyzing the drawbacks of domain-specific training in our case. As domain-specific framework for MDA builds multiple domain discriminators for the pairs of each source and target domain [6], we use G_1, G_2 to discriminate (S_1, T) and (S_2, T) in Fig. 1(a), together with a classifier C trained on S_1 and S_2 (left side in Fig. 1(a)). When adaptation completes under the domain-specific framework (right side in Fig. 1(a)), S_1, S_2 and T are overlapped as the domain discrepancy is reduced and C appropriately classifies S_1 and S_2 with a low source error. However, for MDA, C is not a desirable classifier with a high target error [9]. During training process, F domain-specific framework tends to generate ambiguous features that are incorrectly classified by the task-decision boundaries because domain-specific framework only aims to align the marginal distributions by domain boundaries G_1 and G_2 in MDA. In contrary, task-specific adversarial training concentrates on aligning multi-source and target domains via task-specific decision boundaries C_1 and C_2 (1(c)). According to the above analysis, we answer the first question that building a task-specific adversarial training will achieve a more desirable target performance than domain-specific training for MDA.

However, the existence of multiple source domains in MDA increases the complexity to perform task-specific training. To answer the second question, we conduct further analysis to indicate that the scattering of multi-source representations degrades the effect of task-specific training for MDA. As shown in Fig. 1(b), the first step of task-specific adver-

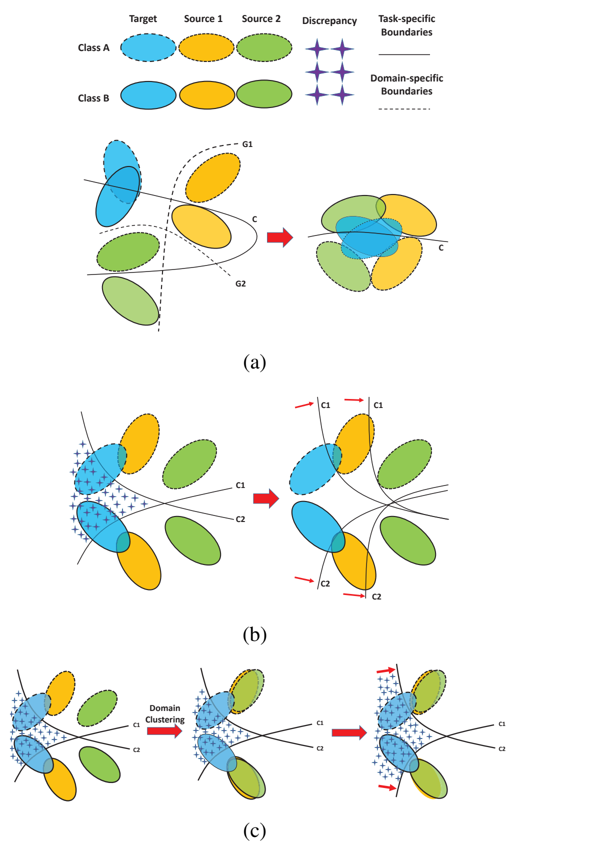

Fig. 1. Analysis of domain-specific and task-specific adversarial training for the situation of binary classification with two source domains and single target domain. (a) Domain-specific training for MDA. (b) Conventional task-specific training for MDA. (c) Task-specific training with domain clustering for MDA (our TMDA).

sarial training maximizes the classifier discrepancy between C_1, C_2 to detect the less discriminative target features outside the support of multi-source domains. Nevertheless, feature representations from multi-source domains might scatter in embedding space (left side in Fig. 1(b)). Consequently, the scattered representations disturb the estimation of the less discriminative target features (right side in Fig. 1(b)). More specifically, the support of multi-source domains becomes fuzzy and the max-min optimization of classifiers C_1 and C_2 is easily misleaded in this case. To tight up multi-source representations, we embed the domain clustering into the task-specific adversarial training for MDA. As shown in Fig. 1(c), the domain clustering decreases the gap across multi-source representations in the embedding space. The max-min optimization of the classifier discrepancy can then be conducted accurately.

Through the above series of analysis, we have answered the two questions raised before. In the next section, we detail the architecture of our framework, along with the formulation of loss function.

B. Formulations and Architectures

1) Source Classification Loss: We first define the classification error of C_1 and C_2 on multi-source domains with cross-entropy loss. As multiple source domains exist, we adopt the max-loss among $S_1, S_2, \ldots, S_K$ for each classifier [6]. Analogously, we use $p_1^s(y|\mathbf{X})$ and $p_2^s(y|\mathbf{X})$ to denote the source outputs from C_1 and C_2. Then we sum the max source error on C_1, C_2 and obtain the total source classification loss:

$$\mathcal{L}_S = \sum_{i=1}^{2} \max_k \left(-\mathbb{E}_{(\mathbf{x},y)\sim P_k} \sum_{c=1}^{C} \mathbb{I}_{[c=y]} \log p_i^s\left(\mathbf{y}|\mathbf{x}\right) \right). \tag{1}$$

2) Sliced Wasserstein Discrepancy Loss: To simplify the annotations, we use $p_1^t(y|\mathbf{x})$ and $p_2^t(y|\mathbf{x})$ to denote the C-dimensional probabilistic outputs for target samples from C_1 and C_2, respectively. As p_1^t and p_2^t are probabilistic distributions, we combine the Wasserstein distance in order to provide more flexible and accurate measurement of their divergence [13]. Assume Ω is a probability space with two probability measures p_1^t and p_2^t. Based on Monge's map [14] and Kantorovich-dual theory [15], the commonly applied 1-Wasserstein distance is formulated as follows:

$$W(\mu, \nu) = \inf_{\gamma \in \Pi\left(p_1^t, p_2^t\right)} \int_{\Omega\times\Omega} c\left(\mathbf{d}_1, \mathbf{d}_2\right)^q d\gamma\left(\mathbf{d}_1, \mathbf{d}_2\right), \tag{2}$$

where $c : \Omega \times \Omega \to \mathbb{R}^+$ is a measurement on the manifold, $\#$ denotes to the push-forward distribution and $\Pi(p_1^t, p_2^t) = \{\gamma \in \mathcal{P}(\Omega \times \Omega) | \pi_{1\#}\gamma = p_1^t, \pi_{2\#}\gamma = p_1^t\}$ with π_1 and π_2 as two marginal distributions in Ω. In order to further reduce the computational cost, we perform the sliced Wasserstein discrepancy [16]:

$$\begin{aligned} SW\left(p_1^t, p_2^t\right) &= \int_{\theta \in \mathbb{S}^{K-1}} W_p^p\left(\mathcal{R}_\theta p_1^t, \mathcal{R}_\theta p_2^t\right) d\theta \\ &\approx \frac{1}{M} \sum_{i=1}^{M} W\left(\mathcal{R}_{\theta_i} p_1^t, \mathcal{R}_{\theta_i} p_2^t\right), \end{aligned} \tag{3}$$

where $\mathbb{S}^{K-1}$ represents the K-dimensional unit ball, $\mathcal{R}_\theta$ denotes to the projection to hyperplane θ on $\mathbb{S}^{K-1}$. Note that we sample M hyperplanes from $\mathbb{S}^{K-1}$ to approximate the integral and the setting of this hyper-parameter will be discussed later. Based on the N observations on each training batch, we formulate the sliced Wasserstein discrepancy loss for outputs from C_1 and C_2 as follows:

$$\mathcal{L}_{sd} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} c\left(T_1(\mathcal{R}_{\theta_i} p_1^t)_j, T_2(\mathcal{R}_{\theta_i} p_2^t)_j\right), \tag{4}$$

where N is the batch size, T_1 and T_2 are repermutations over the batch such that $T_1(\mathcal{R}_{\theta_i} p_1^t)_j \leq T_1(\mathcal{R}_{\theta_i} p_1^t)_{j+1}$, $T_2(\mathcal{R}_{\theta_i} p_2^t)_j \leq T_2(\mathcal{R}_{\theta_i} p_2^t)_{j+1}$ [10]. Moreover, we sets the manifold measurement c to a common quadratic loss in this paper.

3) Domain Clustering Loss: According to our analysis, clustering on multi-source domains is necessary to perform task-specific adversarial training for MDA. Inspired from a recently proposed approach [8], we define the domain gap across the multi-source domains $S_1, S_2, \dots, S_K$ with their distributions $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K$. Let $\hat{\mathcal{P}}_k = \frac{1}{K-1}\sum_{l=1}^{K} w_{kl}\mathcal{P}_l$ be the distribution of the complement for P_k, where the w_{kl} represents the relationship between the source domain S_k and S_l [8]. Note that $w_{kk} = 0$ for $k = 1, 2, \dots, K$. Meanwhile, let μ_k represent the proportions of samples from domain S_k in a training batch. On a batch with N observations $\{(\mathbf{x}_j, s_j)\}$ (s_j is the domain label), we formulate the gap across $S_1, S_2, \dots, S_K$ to describe the scatter of the multi-source domains in $\mathcal{H}$ as follows:

$$\begin{aligned}\mathcal{L}_{cl} &= \sum_{k=1}^{K} \frac{1}{K} d\left(\mathcal{P}_k, \hat{\mathcal{P}}_k\right) \\ &= \frac{1}{K}\sum_{k=1}^{K} \sup_{\|D\|_L \le 1} \left(\mathbb{E}_{\mathbf{x}\sim\mathcal{P}_k}\left[D(E(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{x}\sim\hat{\mathcal{P}}_k}\left[D(E(\mathbf{x}))\right]\right) \\ &\approx \frac{1}{KN}\sum_{j=1}^{N} \mu_{s_j}^{-1} \mathbf{W}_{s_j}^{T} D\left(F\left(\mathbf{x}_j\right)\right),\end{aligned} \tag{5}$$

where $\sup_{\|D\|_L \le 1}$ means that the domain adaptor D is a 1-Lipschitz function, $\mathbf{W}_k$ is written as $W_{kl} = -\frac{1}{K}w_{kl}$ and the output of domain adaptor $D(F(\mathbf{x}))$ is a K-dimensional vector. The second equation in (5) is facilitated by the Kantorovitch-dual theory of the 1-Wasserstein distance [15]. Moreover, to satisfy the dual condition, the Lipschitz constraint $\sup_{\|D\|_L \le 1}$ on D is achieved by adding a penalty on gradients when training D [17]. Furthermore, the relationship weight w_{kl} is also obtained from the output from domain adaptor D as well [8]:

$$\mathbf{w}'_{kl} = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_k}\left[D(F(\mathbf{x}))\right] - \mathbb{E}_{\mathbf{x}\sim\hat{\mathcal{P}}_l}\left[D(F(\mathbf{x}))\right]. \tag{6}$$

Furthermore, to constrain the weights $\mathbf{w}_k$ for domain S_k, a softmax operation is applied: $\mathbf{w}_k = \mathrm{softmax}(\mathbf{w}'_k)$.

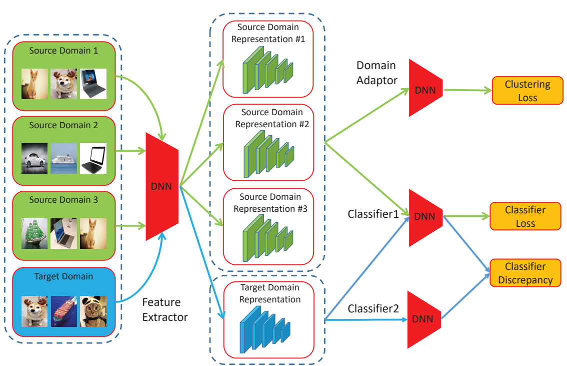

Fig. 2. Architecture of TMDA model.

4) Model Architecture: Combined with the aforementioned analysis and loss functions, we propose an overall framework named "Task-specific Multi-Source Domain Adaptation" (T-MDA) in this paper. Fig. 2 details the architecture of the proposed TMDA. With reference to the earlier annotations, we build the architecture of TMDA as follows:

- A feature extractor F generates the features in the embedding space with the input samples from $S_1, S_2, \dots, S_K$ and T.
- On the one side, two task-specific classifiers C_1, C_2 calculate classification error $\mathcal{L}_S$ on source features. On the other side, the sliced Wasserstein discrepancy $\mathcal{L}_{sd}$ between C_1 and C_2 is estimated to estimate the divergence between target and source manifold.
- A domain adaptor D computes the clustering loss across multi-source domains $\mathcal{L}_{cl}$, which aims to tight up the scattered source representations in embedding space.

Now that the architecture of TMDA is built, we present the details of its optimization in next section.

III. Optimization

In this section, we propose a clustering embedded adversarial training algorithm to optimize TMDA model. More specifically, we divide the whole optimization process into four steps:

Optimizing Source Risk: First, we optimize C_1 and C_2 on samples from multi-source domains. Similar to [9], we combine the classification loss with the clustering loss as the objective function and train F, C_1 and C_2:

$$\min_{F, C_1, C_2} \mathcal{L}_S + \rho_1 * \mathcal{L}_{cl}, \tag{7}$$

where ρ_1 is the regularization parameter to balance the classification and clustering.

Clustering Multi-source Domains: In this step, we train the domain adaptor D to maximize the clustering loss according to the derivation in (5):

$$\max_{D} \mathcal{L}_{cl} + \mathcal{L}_{pal}. \tag{8}$$

The $\mathcal{L}_{pal}$ in (8) denotes to the gradient penalty term, which smooths the domain adaptor D in order to satisfy the 1-Lipschitz constraint in the dual condition [15].

Maximizing Classifier Discrepancy: We then train the two task-specific classifiers C_1, C_2 to maximize their disagreement. In this step, the target samples away from the overlapped manifolds of multi-source domains are detected. Combined with the previous empirical results [9], we optimize C_1, C_2 with the source risk as follows:

$$\min_{C_1, C_2} \mathcal{L}_S - \rho_2 * \mathcal{L}_{sd}, \tag{9}$$

where ρ_2 is the regularization parameter to control the discrepancy maximization.

Refining Target Representations: Since the less discriminative target representations in embedding space are detected, we finally train the feature extractor F to pushing the these target representations close to source manifolds [10]:

$$\min_{F} \mathcal{L}_{sd}. \tag{10}$$

IV. Experiments

In this section, we evaluate TMDA and compare it with several state-of-the-art methods on two real-life tasks: natural language sentiment analysis and digit image classification.

A. Natural language sentiment analysis

Here we validate TMDA on the Amazon Review [18] dataset. As the commonly adopted benchmark for MDA [8], [19], this dataset contains four kinds of product (Books, DVDs, Electronics and Kitchen appliances), each of which is considered as a domain. Since most reviews are labeled either positive or negative, the sentiment analysis can be considered as a binary classification problem. In our experimental settings, the text feature for each review is cropped into a 5,000-dimensional vector. Then we set each product as the target domain and the other three products are combined as the source domains. Each source domain contains 2,000 samples and the target domain contains 3,000-6000 samples [6]. In order to evaluate TMDA, we compare its performance with five baselines (TCA [20], SA [21], ITL [22], MSDA [18], DANN [23]) and two state-of-the-art methods (MDAN [6], MDMN [8])

Meanwhile, we detail the parameters setting of TMDA on Amazon Review dataset. The whole deep learning architecture of TMDA is built with fully connected layers. More specifically, the feature extractor F contains three hidden layers with the units $\{1000, 500, 100\}$, while the two classifiers C_1, C_2 and the domain adaptor D consist of two layers with units $\{100, 100\}$. The *relu* is applied after each layer as the activation function and the Adadelta optimizer is applied to train the whole model. The number of hyperplanes for discrepancy computation M is set to 128, while the regularization parameters ρ_1 and ρ_2 are set to 0.01 and 0.1, respectively.

TABLE I
PERFORMANCE OF NATURAL LANGUAGE SENTIMENT ANALYSIS ON AMAZON REVIEW DATASET (%). IN THE TABLE, B MEANS BOOK, D MEANS DVD, E MEANS ELECTRONICS AND K MEANS KITCHEN.

Method	**D+E+K-B**	**B+E+K-D**	**B+D+K-E**	**B+D+E-K**
Source	79.52	81.54	83.49	85.78
Baselines				
TCA	79.64	79.75	82.49	84.81
SA	79.04	81.96	83.37	85.55
ITL	79.60	81.90	82.75	85.25
Domain Adaptation				
MSDA	76.98	78.61	77.32	78.86
DANN	79.60	80.51	84.12	85.84
Multi-Source Domain Adaptation				
MDAN	80.76	**82.74**	84.54	86.16
MDMN	81.21	82.37	84.63	86.56
TMDA	**82.85**	81.90	**86.01**	**87.21**

Results and Analysis The classification results of different methods are presented in Table I. It is clear that the proposed TMDA achieves better performance than other methods. When DVD is set as the target domain, MDAN performs slightly better than TMDA and all the methods deliver roughly the same performance. For three other adaptations, TMDA outperforms the state-of-the-art methods with a significant improvement. This is mainly attributed to the fact that TMDA aligns the multi-source and target domains by task-specific decision boundaries. More specifically, TMDA pushes the less discriminative target representations close to the task-specific decision boundaries such that these features are more easily to be classified. Other deep MDA methods such as MDAN and MDMN, only optimize the target features in embedding space by domain-specific discriminators, which leads to a suboptimal. In summary, based on the observations and analysis on Amazon Review dataset, we successf verify the effectiveness of our TMDA for unsupervised multi-source domain adaptation.

B. Digit image classification

We then further verify the proposed TMDA on the digit image classification task. This task contains four public benchmarks: MNIST [1], MNIST-M [23], SVHN [2] and USPS [3]. Analogously, we choose each benchmark as the target domain and the combination of the other three benchmarks as the source domains.

Similar to sentiment analysis, we choose TCA, SA and DANN as the three baselines to evaluate their performance on four digit benchmarks. For fair comparison, we first extract the features from the same Lenet-5 model trained on the source domains and then feed the representations to TCA and SA as inputs. Meanwhile, when considering the computational cost of TCA and SA, the data resources from each source domain are limited. Moreover, we compare TMDA with six recently proposed methods: ADDA [24], MDAC [7], MTAE [25], MCD [9], MDAN and MDMN. ADDA is a recently proposed domain adaptation method based on discriminative model and GAN-based loss. MDAC is a novel MDA approach that builds causal models to extract the relationship between features and class labels. MTAE is an improved autoencoder for object recognition across multi-domains. Note that we perform DANN, ADDA, MCD and MTAE on each of the three source-target domain pairs and record the best accuracies.

The mode setup of TMDA on digit image classification is different from sentiment analysis. Since network input comprises grayscale or colored images, we use the commonly applied Lenet-5 network [4] to extract the image features. More specifically, F is built with two convolutional layers, two pooling layers and one fully connected layer. Moreover, for the classifiers C_1, C_2 and domain adaptor D, we build two layer fully connected networks with $\{100, 100\}$ hidden units. Here, we adopt the Adam optimizer to train TMDA and set the hyperplane number M to 256. Meanwhile, the regularization parameters ρ_1 and ρ_2 are set to 0.1 and 0.1, respectively.

Results and Analysis As shown in Table II, we report the performance of different methods on MNIST, MNIST-M, USPS and SVHN. We observe that most MDA methods achieve better performance than the Source-only baseline.

[1] http://yann.lecun.com/exdb/mnist/
[2] http://ufldl.stanford.edu/housenumbers/
[3] https://www.kaggle.com/bistaumanga/usps-dataset
[4] http://yann.lecun.com/exdb/lenet/a35.html

TABLE II

PERFORMANCE OF DIGIT IMAGES CLASSIFICATION ON MNIST, MNIST-M, USPS, SVHN (%). THE SOURCE-ONLY APPROACH MEANS TRAINING DIRECTLY ON SOURCE DOMAINS WITHOUT ADAPTATION.

Method	**MNIST**	**MNIST-M**	**USPS**	**SVHN**
Source	94.6	60.8	89.4	43.7
Baselines				
TCA	88.4	55.2	85.4	39.7
SA	90.8	59.9	86.3	40.2
Domain Adaptation				
MTAE	86.7	62.1	71.0	38.2
DANN	97.1	67.0	90.4	51.9
MCD	96.4	72.3	94.2	44.1
ADDA	89.0	80.3	85.2	43.5
Multi-Source Domain Adaptation				
MDAC	85.5	55.3	72.4	41.8
MDAN	97.2	68.5	90.1	50.5
MDMN	98.0	83.8	84.5	**53.1**
TMDA	**98.5**	**86.1**	**96.6**	52.2

This proves that it is necessary for a trained model to perform adaptation when facing an unfamiliar target domain. Meanwhile, the performances of TCA and SA show that the negative transfer occurs in classical subspace learning methods when facing large-scale and complex data distributions. Deep single domain adaptation methods including ADDA, DANN and MTAE obtain undesirable adaptation on MNIST-M and SVHN, where MDAN, MDMN and TMDA achieve better accuracies. This implies that the combination of multi-source domains improves the target performance. Compared to other methods, the proposed TMDA achieves significant improvement on MNIST, MNIST-M and USPS. This can mainly be attributed to the task-specific framework, which provides a better feature multiple domain alignment by utilizing task-specific decision boundaries. Meanwhile, the embedding of domain clustering decreases the scatter representations and ensures the performance of task-specific adversarial training.

V. CONCLUSIONS

In this paper, we propose a novel unsupervised multi-source domain adaptation method, named TMDA, to align multi-source and target domains via task-specific decision boundaries. Meanwhile, the clustering of multi-source domains in the embedding space ensures the accuracy of adversarial training process. Although the experimental results verify the effectiveness of TMDA, our method is still lack of an theoretical explanation of task-specific framework under MDA. Moreover, the clustering of embedded multi-source features requires further improvement. These issues will be considered in our future work.

REFERENCES

[1] R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, "Deep cocktail network: Multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3964–3973.

[2] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 1013–1026, 2018.

[3] Y. Luo, Y. Wen, and D. Tao, "Heterogeneous multitask metric learning across multiple domains," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 9, pp. 4051–4064, 2017.

[4] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[5] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-gan for object transfiguration in wild images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 164–180.

[6] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 8559–8570.

[7] K. Zhang, M. Gong, and B. Schölkopf, "Multi-source domain adaptation: A causal view," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[8] Y. Li, D. E. Carlson *et al.*, "Extracting relationships by multi-domain matching," in *Advances in Neural Information Processing Systems*, 2018, pp. 6798–6809.

[9] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[10] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," *arXiv preprint arXiv:1903.04064*, 2019.

[11] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[13] L. Mi, W. Zhang, X. Gu, and Y. Wang, "Variational wasserstein clustering," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 322–337.

[14] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie royale des sciences de Paris*, 1781.

[15] L. V. Kantorovich, "On the translocation of masses," *Journal of Mathematical Sciences*, vol. 133, no. 4, pp. 1381–1382, 2006.

[16] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde, "Generalized sliced wasserstein distances," *arXiv preprint arXiv:1902.00434*, 2019.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[18] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012.

[19] H. Zhao, R. T. d. Combes, K. Zhang, and G. J. Gordon, "On learning invariant representation for domain adaptation," *arXiv preprint arXiv:1901.09453*, 2019.

[20] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[21] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2960–2967.

[22] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," *arXiv preprint arXiv:1206.6438*, 2012.

[23] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[24] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[25] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2551–2559.