Partial Optimal Transport for Universal Domain Adaptation

Zhijian Li, 22031019

Abstract

Recently, Optimal Transport has been adopted to solve the closed-set Domain Adaptation (CDA) problem, which aims at transfering the knowledge from labeled source domain to unlabeled target domain. However, in practice there may exists category shift between two domains, therefore methods need to detect "unknown" classes in the unlabeled target domain, which is known as the Universal Domain Adaptation problem (UniDA). As a result, vanilla Optimal Transport method failes to work in UniDA problem, since it may lead to negative transfer due to the existence of the extra "unknown" classes. In this report, we present Deep Partial Optimal Transport (DPOT) method for UniDA problem. Our idea is that partial optimal transport provides more flexible transportation plans only for pairs of samples from common classes between source and target domain, thus alleviate the negative transfer phenomenon. Besides, we introduce a novel strategy which automatically obtain the transportation mass, consequently our method adapts to different situations in the UniDA problem well. To our best knowledge, this is the first optimal transport method to handle UniDA problem, and extensive experiments demonstrate the superiority of our method.

I. Introduction

Deep convolutional neural network has achieved significant progress in computer vision field, such as image classification [13] and semantic segementation [12]. However, this technique requires extensive labeled training data and shows poor generalization under domain shift, which limits its application in real world task. To overcome this issue, Unsupervised Domain Adaptation (UDA) transfer knowledge from a labeled source domain to an unlabeled target domain. Early research assume that there is no category shift between source and target domain, i.e., they share completely the same label set, which is known as the closed-set Domain Adaptation (CDA) situation [33]. However, this assumption may not hold in complex real-world scenario, where category shift may exist between source and target domain.

Recently developed UDA methods consider several potential situations: open-set Domain Adaptation (ODA) [17, 25, 20] assumes that the target domain holds some private classes which do not exists in the source domain; partial Domain Adaptation (PDA) [3] consider the setting that only the source domain holds private classes; open-partial Domain Adaptation (OPDA) [26] is a mixture of ODA and PDA. The concept of Universal Domain Adaptation was proposed by [33, 26] to take into account the uncertainty of category shift, i.e. the four different situations including CDA, PDA, ODA and OPDA.

Although optimal transport has been adopted by extensive methods for the CDA problem and shown powerful ability of finding a common representation between the source and target domain [6, 7, 8, 16, 31], they fails to work in UniDA problem due to the lack of ability to reject private classes in the target domain and the inevitable negative transfer phenomenon. Specifically, these methods focus on reducing the domain shift between source and target domain via optimal transport, such that a single classifier can be used in a common feature space for datapoints from both source and target domain [8]. However, vanilla optimal transport provides a transportation plan that match all the samples between source and target domain, as a result it easily leads to a phenomenon called negative transfer. Consider the settings of PDA, ODA and OPDA where category shift exists, this transportation plan forces samples from private classes of target domain to match samples from common classes of source domain, which may degenerate the performance compared to a model without adaptation.

In this report, we aim at addressing these challenging problems via partial optimal transport, and propose Deep Partial Optimal Transport (DPOT) method for Universal Domain Adaptation problem. Unlike vanilla optimal transport, partial optimal transport method allows to transport only a fraction of samples from the whole discrete distribution, thus naturally relief the adverse influence caused by negative transfer. FurtherMore, we introduce a novel strategy to automatically determine the proportion of transportation mass in partial optimal transport, which brings our method great adaptability to different situations in the UniDA problem. Our contributions can be summarized as follows:

- As far as we know, DPOT firstly adopts partial optimal transport method to address the UniDA problem. With a novel technique of automatically determining the proportion of transportation mass, our method alleviate the negative transfer phenomenon arised in the previous optimal transport based methods and is suitable for the different situations in UniDA problem.
- We apply dynamic threshold to reject "unknown" samples in the target domain by utilizing the prediction entropys of source samples. Unlike other UniDA methods who determine the threshold by validation with labeled target samples [33, 10] or the number of classes [26], our strategy is more flexible and leads to a better performance on detecting unknown samples.

Extensive experiments conducted on benchmark datasets show that DPOT achieves competitive performance in UniDA tasks.

II. Related Work

a) Domain Adaptation: Unsupervised Domain Adaptation aims at learning a classifier for unlabeled target data given labeled source data. Denote L_s and L_t as the label set of source and target domain respectively. Early research focus on the closed-set Domain Adaptation (CDA) task ($L_s = L_t$), most of these methods measure the distance of discrete feature distributions between source and target domain [18, 29, 8, 16, 31, 27, 14], then train a single feature extractor model by minimizing this distance, such that one classifier can perform prediction for samples from both source and target domain. However, these methods can not easily generalize to complex situations, where category shift may exist between source and target domain. Partial Domain Adaptation (PDA) covers the case where target classes are a subset of source classes ($|L_s - L_t| > 0, |L_t \cap L_s| = |L_t|$). This task is solved by identifying common samples with similarity measure from the domain discriminator and performing importrance-weight on source samples for adversarial training [34, 4, 3]. Different from the situation of CDA and PDA, Open-set Domain Adaptation (ODA) task assumes that target domain holds some private classes that are unknown to the source domain ($|L_s - L_t| > 0, |L_t \cap L_s| = |L_s|$). Methods for ODA task need to perform prediction for samples from common classes and recognize "unknown" samples simultaneously during testing [25, 17, 20]. Note that ODA methods assume that there necessarily exists unknown samples in the target domain, thus they will fail in closed-set and partial Domain Adaptation tasks. Universal Domain Adaptation (UniDA) task was firstly proposed by [33] as a mixture of PDA and ODA, which is also named as open-partial Domain Adaptation (OPDA) by [26]. Saito et al. proposed a method DANCE to adapt well on CDA, PDA, ODA , and OPDA tasks simultaneously, they call this task Universal Domain Adaptation. Previous researches [10, 33, 24, 26] set a threshold, and samples from target domain with a lower score than the threshold will be regarded as "unknown". However, You et al. [33] and Fu et al. [10] determine a fixed threshold through validation [24], which may not be flexible enough to handle different situations in UniDA task. Saito et al. [26] set the threshold decided by the number of classes in source domain, which may not works well in ODA and OPDA tasks [24]. Our method applies dynamic threshold to reject "unknown" samples in the target domain by utilizing the prediction entropys of source samples, which is more flexible and leads to a better performance on detecting unknown samples.

b) Optimal Transport on Domain Adaptation: Courty et al. [6] firstly applied optimal transport in CDA problem to learn the transformation between domains with associated theoretical guarantees [8]. Based on the paradigm in [6], Courty et al. [7] proposed JDOT to minimize the optimal transport loss between the joint feature/label space, and finally learn a classifier on the target domain. Subsequently, DeepJDOT [8] was proposed to not only learn new data representations aligned between the source and target domain, but also simultaneously preserve the discriminative information used by the classifier via a measure of discrepancy on joint deep representations/labels based on optimal transport. Soon after, ETD [16] and RWOT [31] further improved DeepJDOT by weighting the transport distance to obtain a more precise transportation plan. However, none of these methods could easily apply to the UniDA task, since they transport all the samples which may cause negative transfer due to the exists of private classes. Our method adopts partial optimal transport to provide more flexible transportation plans only for pairs of samples from common classes between source and target domain, thus alleviate the negative transfer phenomenon. Besides, we introduce a novel strategy which automatically obtain the transportation mass, consequently our method adapts to different situations in the UniDA problem well.

III. PRELIMINARIES

A. Notation

Given a labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ with categories L_s and an unlabeled target domain $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{N_t}$ with categories L_t, the UniDA task focuses on labeling the target samples with either one of the L_s labels or rejecting it as "unknown". Following the conventions of [33], we denote $L = L_s \cap L_t$ as the common label set shared by the two domains, $\bar{L}_s = L_s \backslash L$ and $\bar{L}_t = L_t \backslash L$ as the private label sets of source and target domain respectively. Domain Adaptation problem assumes that data $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ are sampled from different probability distributions $\mathcal{P}_s$ and $\mathcal{P}_t$, hence there exists shift between the two domains $\mathcal{P}_s \neq \mathcal{P}_t$.

B. Partial Optimal Transport

Denote $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^m$ and $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^n$ as two sample spaces representing the source and target respectively. Consider two discrete distributions $\mathbf{p} = \sum_{i=1}^m p_i \delta(\mathbf{x}_i)$ and $\mathbf{q} = \sum_{j=1}^n q_i \delta(\mathbf{y}_j)$ over $\mathcal{X}$ and $\mathcal{Y}$ s.t. $\mathbf{p} \in \sum^m$ and $\mathbf{q} \in \sum^n$, where $\sum^m$ and $\sum^n$ are histograms of $|\mathbf{p}| = m$ and $|\mathbf{q}| = n$ bins respectively [5]. Given C_{ij} defining the transportation cost between points $\mathbf{x}_i$ and $\mathbf{y}_j$, Optimal Transport (OT) sovles the problem of transporting distribution $\mathbf{p}$ to $\mathbf{q}$ with lowest total cost. When the cost matrix $\mathbf{C}$ is a distance matrix, the p-Wasserstein distance between $\mathbf{p}$ and $\mathbf{q}$ at the power of p is defined as

$$W_p^p(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \langle \mathbf{C}, \mathbf{T} \rangle = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i=1}^m \sum_{j=1}^n C_{ij} T_{ij},$$

where $\mathbf{T}$ is a coupling matrix with an element T_{ij} describes the transportation mass from $\mathbf{x}_i$ to $\mathbf{y}_j$, and $\Pi(\mathbf{p}, \mathbf{q})$ denotes the set of all admissible couplings which is given by

$$\Pi(\mathbf{p}, \mathbf{q}) = \{\mathbf{T} \in \mathbb{R}_+^{m \times n} | \mathbf{T}\mathbf{1} = \mathbf{p}, \mathbf{T}^T\mathbf{1} = \mathbf{q}\}.$$

The previous OT problem forces to transport all the mass between distribution $\mathbf{p}$ and $\mathbf{q}$, while the partial OT problem focuses on transporting only a fraction $0 \leq b \leq 1$ of the mass with lowest total cost [5, 2]. Under this circumstances, the set of admissible couplings becomes

$$\Pi^b(\mathbf{p}, \mathbf{q}) = \{\mathbf{T} \in \mathbb{R}_+^{m \times n} | \mathbf{T}\mathbf{1} \leq \mathbf{p}, \mathbf{T}^T\mathbf{1} \leq \mathbf{q}, \mathbf{1}^T\mathbf{T}\mathbf{1} = b\},$$

and the partial optimal transport problem is defined as

$$PW_p^p(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{T} \in \Pi^b(\mathbf{p}, \mathbf{q})} \langle \mathbf{C}, \mathbf{T} \rangle.$$

IV. METHODOLOGY

The model of DPOT method consists of two parts, a feature extractor F aims at mapping the input images $\mathbf{x}$ from either domain into vector representations $\mathbf{f}$, and a classifier G that maps the vector representation to the label space on the target domain. What distinguishes our work from the previous studies [31, 8] is that DPOT does not force complete transportation between the source and target distributions, thus naturally avoid the risk of misalignment.

The optimization of DPOT is based on the partial optimal transport problem to obtain a transportation plan between source and target domain:

$$\mathbf{T}^* = \underset{\mathbf{T} \in \Pi^b(\mu_s, \mu_t)}{\arg\min} \langle \mathbf{C} \cdot \mathbf{W}, \mathbf{T} \rangle = \underset{\mathbf{T} \in \Pi^b(\mu_s, \mu_t)}{\arg\min} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} C_{ij} W_{ij} T_{ij}, \tag{1}$$

where μ_s and μ_t represent the probability distribution of source and target domain respectively, and $\mathbf{W}$ is the weight matrix which incorporate the classification probability information of target samples to reduce the wrong pair-wise transportation. Denote $\mathbf{Z} = \mathbf{C} \cdot \mathbf{W}$ as the cost matrix in partial optimal transport (1), which has the following formulation

$$Z_{ij} = \|F(\mathbf{x}_i^s) - F(\mathbf{x}_j^t)\|_2^2 \cdot (1 - P(j, y_i^s)).$$

Here, $P(j, y_i^s)$ is the probability of the target samples j belongs to the label class y_i^s, which is defined as

$$P(j,c) = \begin{cases} p(\hat{y}^c|\mathbf{x}_j^t), & H(\mathbf{x}_j^t) \leq \omega_k \\ 0, & H(\mathbf{x}_j^t) > \omega_k \end{cases}$$

where $H(\mathbf{x}_j^t)$ denotes the entropy of the predicted probability vector of target sample $\mathbf{x}_j^t$, and ω_k is the threshold to reject "unknown" samples at iteration k. Denote $p(\hat{y}^k|\mathbf{x}_j^t)$ as the output probability of image $\mathbf{x}_j^t$ for the class k: $p(\hat{y}^k|\mathbf{x}_j^t) = \sigma(G(F(\mathbf{x}_j^t)))_k$, where σ represents the sigmoid activation function.

Inspired by the Hard Negative Classifer Sampling (HNCS) technique in [24], we propose to update the feature extractor F and classifier G by gradient descent to reduce sample-wise distance of the same classes with fixed coupling γ^* obtained at the previous step (1) for the following loss

$$\mathcal{L}_a = \sum_{i,j} \gamma^* \left(\alpha \|F(\mathbf{x}_i^s) - F(\mathbf{x}_j^t)\|_2^2 - \beta \left(\log\left(p(\hat{y}^{y_i^s}|\mathbf{x}_j^t)\right) + \min_{c \neq y_i^s} \log\left(1 - p(\hat{y}^c|\mathbf{x}_j^t)\right) \right) \right), \tag{2}$$

where parameter α and β are used to balance the effects of the loss terms.

According to the theoretical result of [1], the classification error on the target domain is bounded by three terms, the domain discrepancy, the classification error on the domain classifier, and a shared error which is usually viewed as constant. Therefore, we also need to minimize the classification error and extract discriminant features on the source domain, which is a supervised learning task only performed on the source domain with the following loss

$$\mathcal{L}_c = -\frac{1}{N_s} \sum_{i=1}^{N_s} \log\left(p(\hat{y}^{y_i^s}|\mathbf{x}_i^s)\right) + \min_{c \neq y_i^s} \log\left(1 - p(\hat{y}^c|\mathbf{x}_i^s)\right). \tag{3}$$

To extract discriminant feature on the target domain and seperate "unknown" samples with the threshold ω_k, we introduce the following entropy loss based on the transportation plan γ^*

$$\mathcal{L}_e = \frac{1}{|B_k|} \sum_{j \in B_k} |H(\mathbf{x}_j^t) - \omega_k|,$$

where B_k is the set of "unmatched" target samples at iteration k, i.e. $B_k = \{j | \gamma^*_{k,i,j=0}, \forall i\}$.

Combining the optimization objective functions described above, the whole loss fucntion of the model consists of three terms, the source classification loss $\mathcal{L}_c$, and domain alignment loss $\mathcal{L}_a$, and the target entropy loss $\mathcal{L}_e$, which can be written as

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_a + \lambda \mathcal{L}_e,$$

where the parameter λ is used to balance the effect of loss functions. To render our method scablable to large-scale datasets, we propose to minimize $\mathcal{L}$ with a stochastic approximation using minibatches from both the source and target domain.

NOTE: extra subsection to introduce One-vs-All classifier and the overconfidence phenomenon in Softmax classifier.

V. Experiment

NOTE: writting based on DANCE

A. Experimental Settings

The goal of the experiments is to compare DPOT with other UniDA methods across different situations of UniDA (i.e., CDA, PDA, ODA, and ODPA) under the three object classification datasets. We following the settings of [15] in our experiments.

a) Datasets: . We use Office31 [23] as the first dataset, which has three domains (Amaon (A), DSLR (D), Webcam (W)) and 31 classes for each domain. The second benchmark dataset OfficeHome [30] contains four domains and 65 classes. The third dataset VisDA [22] contains 12 classes from two domains: synthetic and real images. We show the class split in each setting ($|L_s \cap L_t|/|L_s - L_t|/|L_t - L_s|$) in each table of the results.

Method	Office31(31/0/0)						
	A2W	D2W	W2D	A2D	D2A	W2A	Avg
SO	74.1	95.3	99.0	80.1	54.0	56.3	76.5
DANN	86.7	97.2	99.8	86.1	72.5	72.8	85.9
CDAN	93.1	98.2	100	89.8	70.1	68.0	86.6
UAN	86.5	97.0	100	84.5	69.6	68.7	84.4
DANCE	88.6	97.5	100	89.4	69.5	68.2	85.8
DCC	89.1	96.8	100	87.2	74.4	76.8	87.4
OVANet	-	-	-	-	-	-	-
DPOT	83.5	84.9	94.6	86.1	37.8	41.5	71.4

TABLE I: Classification accuracy on closed-set domain adaptation of dataset Office31.

Method	OfficeHome(65/0/0)												
	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
SO	37.0	62.2	70.7	46.6	55.1	60.3	46.1	32.0	68.7	61.8	39.2	75.4	54.6
DANN	46.8	68.4	76.6	54.7	63.9	69.7	57.1	44.7	75.7	64.9	51.3	78.7	62.7
CDAN	49.0	69.3	74.5	54.4	66.0	68.4	55.6	48.3	75.9	68.4	55.4	80.5	63.8
UAN	45.0	63.6	71.2	51.4	58.2	63.2	52.6	40.9	71.0	63.3	48.2	75.4	58.7
DANCE	54.3	75.9	78.4	64.8	72.1	73.4	63.2	53.0	79.4	73.0	58.2	82.9	69.1
OVANet	-	-	-	-	53.0	-	-	-	-	-	-	-	-
DPOT	44.3	63.4	69.2	59.8	66.8	68.9	49.3	40.5	68.0	63.0	45.2	73.1	59.3

TABLE II: Classification accuracy on closed-set domain adaptation of dataset OfficeHome.

b) Evaluation: We use the same evaluation metrics as previous works. In CDA and PDA, we simply calculate the accuracy over all target samples. In ODA and OPDA, target-private classes are grouped into a single "unknown" class, and we report the H-score metric, which is the harmonic mean on accuracy of common samples and private ones [10]. For all the settings, we activate "unknown" sample rejection method for all methods since we assume that we do not know the kind of category shift.

c) Implementation: All experiments are implemented in Pytorch [21], and we employ ResNet50 [11] pre-train on ImageNet [9] with an additional full conneted layer as the feature extractor in all experiments. As for the classifier, we use a single full conneted layer. We optimize the model using Nesterov momentum SGD with momentum of 0.9 and weight decay of 5×10^{-4}. The initial learning rate is set to 0.01 and the batch size is set to 96 for all the experiments. We set hyper-parameters $\alpha = 0.002$, $\beta = 0.01$, and $\lambda = 0.01$.

B. Results

We summarize the experimental results in the following Tables including four sub-cases of UniDA task, the propoed method DPOT achieves comparable performance among the previous state-of-the-arts methods.

VI. Discussion

a) NOTE: **This a brief report of my recent work, plese keep it private.**

Method	OfficeHome(25/40/0)												
	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
SO	37.1	64.5	77.1	52.0	51.3	62.4	52.0	31.3	71.6	66.6	42.6	75.1	57.0
DANN	35.5	48.2	51.6	35.2	35.4	41.4	34.8	31.7	46.2	47.5	34.7	49.0	40.9
ETN	52.1	74.5	83.1	69.8	65.2	76.5	69.1	50.6	82.5	76.3	53.8	79.1	69.4
UAN	24.5	35.0	41.5	34.7	32.3	32.7	32.7	21.1	43.0	39.7	26.6	46.0	34.2
DANCE	53.6	73.2	84.9	70.8	67.3	82.6	70.0	50.9	84.8	77.0	55.9	81.8	71.1
DCC	54.2	47.5	57.5	83.8	71.6	86.2	63.7	65.0	75.2	85.5	78.2	82.6	70.9
OVANet	-	-	-	-	-	-	-	-	-	-	-	-	-
DPOT	47.7	64.4	72.1	64.7	64.6	71.2	56.0	44.4	70.2	66.9	49.6	73.0	62.1

TABLE III: Classification accuracy on partial domain adaptation of dataset OfficeHome.

Method	VisDA(6/6/0)
SO	46.3
DANN	38.7
ETN	59.8
UAN	39.7
DANCE	73.7
OVANet	-
DPOT	46.5

TABLE IV: Classification accuracy on partial domain adaptation of dataset VisDA.

Method	Office31(10/0/11)						
	A2W	D2W	W2D	A2D	D2A	W2A	Avg
OSBP	82.7	97.2	91.1	82.4	75.1	73.7	83.7
ROS	82.1	96.0	99.7	82.4	77.9	77.2	85.9
UAN	46.8	68.8	53.0	38.9	68.0	54.9	55.1
DCC	54.8	89.4	80.9	58.3	67.2	85.3	72.6
DANCE	-	-	-	-	-	-	-
OVANet	-	-	-	-	-	-	-
DPOT	90.6	93.6	94.9	87.3	77.3	79.8	87.3

TABLE V: H-score on open-set domain adaptation of dataset Office31.

Method	OfficeHome(25/0/40)												
	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
OSBP	55.1	65.2	72.9	64.3	64.7	70.6	63.2	53.2	73.9	66.7	54.5	72.3	64.7
ROS	60.1	69.3	76.5	58.9	65.2	68.6	60.6	56.3	74.4	68.8	60.4	75.7	66.2
UAN	0.0	0.0	0.2	0.0	0.2	0.2	0.0	0.0	0.2	0.2	0.0	0.1	0.1
DCC	56.1	67.5	66.7	49.6	66.5	64.0	55.8	53.0	70.5	61.6	57.2	71.9	61.7
DANCE	-	-	-	-	-	-	-	-	-	-	-	-	-
OVANet	-	-	-	-	-	-	-	-	-	-	-	-	-
DPOT	56.7	66.5	72.6	62.0	63.7	68.2	62.3	55.6	71.2	69.6	58.8	72.1	64.9

TABLE VI: H-score on open-set domain adaptation of dataset OfficeHome.

Method	Office31(10/10/11)						
	A2W	D2W	W2D	A2D	D2A	W2A	Avg
UAN	58.6	70.6	71.4	59.7	60.1	60.3	63.5
CMU	67.3	79.3	80.4	68.1	71.4	72.2	73.1
DANCE	71.5	91.4	87.9	78.6	79.9	72.2	80.3
DCC	78.5	79.3	88.6	88.5	70.2	75.9	80.2
OVANet	79.4	95.4	94.3	85.8	80.1	84.0	86.5
DPOT	81.1	94.3	88.9	64.1	86.5	88.8	84.0

TABLE VII: H-score on open partial domain adaptation of dataset Office31.

Method	OfficeHome(10/5/50)												
	A2C	A2P	A2R	C2A	C2P	C2R	P2A	P2C	P2R	R2A	R2C	R2P	Avg
UAN	51.6	51.7	54.3	61.7	57.6	61.9	50.4	47.6	61.5	62.9	52.6	65.2	56.6
CMU	56.0	56.9	59.1	66.9	64.2	67.8	54.7	51.0	66.3	68.2	57.8	69.7	61.6
DANCE	-	-	-	-	-	-	-	-	-	-	-	-	49.2
DCC	58.0	56.9	58.0	74.6	70.6	77.5	64.3	73.6	75.0	81.0	75.1	80.4	70.2
OVANet	62.8	75.6	78.6	70.7	68.8	75.0	71.3	58.6	80.5	76.1	64.1	78.9	71.8
DPOT	58.4	77.8	84.8	71.8	69.4	77.8	75.6	58.3	84.6	79.2	61.0	80.7	73.2

TABLE VIII: H-score on open partial domain adaptation of dataset OfficeHome.

Method	VisDA(6/3/3)
UAN	30.5
CMU	34.6
DANCE	4.4
DCC	43.0
OVANet	53.1
DPOT	52.6

TABLE IX: H-score on open partial domain adaptation of dataset VisDA.

b) Use softmax as activation function to calculate the entropy: Softmax cross-entropy tends to produce more overconfient incorrect predictions [19], as a result, without the entropy seperation loss $\mathcal{L}_e$ the samples from target domain will have low entropy value even for those belong to "unknown". Thus I think it is not a good choice to use softmax.

c) Use sigmoid as activation function to calculate the entropy: Without softmax normalization, the prediced vector of sigmoid function is not a probability distribution, thus we normalize the output $p_c = \sigma(z_c)/\sum_{c=1}^{K}\sigma(z_c)$ before calculating the entropy. According to [19], a one-vs-all classifier captures the notion of "none of the above", thus an "unknown" samples from target domain may be rejected by all the classifiers which leads to high entropy.

d) Why not use the prediced probability (confidence) as the threshold ω_k: As we discussed above, the softmax cross-entropy tends to produce overconfient predictions for samples from target domain, which makes it difficult to detect "unknown" samples. On the other hand, it might be a good choice to define the threshold as the confidence with one-vs-all classifier, since it naturally capture the notion of "none of the above". Actually, entropy measurement and confidence measurement achieve similar performance in most cases, except for the OfficeHome dataset under closed-set domain adaptation setting. Specifically, the samples $\mathbf{x}$ from OfficeHome dataset have high entropy $H(\mathbf{x})$, which leads to a high threshold value consequently. As a result, few samples from target domain will be rejected as "unknown", thus contribute to obtain a better classification accuracy.

e) The performance of DANCE: DANCE shows superior performance in OfficeHome and VisDA under CDA setting, but poor performance in OPDA situation with H-score metric. DANCE set a threshold decided by the number of classes in the source $\omega = \frac{1}{2}\log K$, which does not always works well. Here are some of my viewpoints: (1) Static threshold can not adapt well to different situation, for two datasets with the same number of classes but different underlying distribution, the entropy $H(\cdot)$ of images from the two datasets may differ a lot. (2) It is not suitable to apply entropy as threshold. The value of threshold $\omega = \frac{1}{2}\log K$ will be high with a large number classes K. As a result, most of the samples from target domain will not be rejected since the softmax cross-entropy tends to produce overconfient predictions. Here are some conclusions from ablation study of DANCE: (1) Entropy Seperation loss significantly improves performance even in CDA and PDA setting, it seems that lower the entropy of target predictions is critical. (2) Clustering on the target domain marginally improves performance.

f) Dynamic Threshold?: DOC [28] builds a multi-class classifier with a one-vs-all final layer of sigmoids rather than softmax to reduce the open space risk. It reduces the open space risk further for rejection by tightening the decision boundaries of sigmoid functions with Gaussian fitting. We set the threshold using "matched" target samples from the same minibatch, e.g. *entropy_list.mean() + 3 * entropy_list.std()* or *prob_list.mean() - 3 * prob_list.std()* as our first attempt. However, this proposal does not work well due to the mismatch of "unknown" target samples. As an alternative, we set the threshold using source samples from the same minibatch, which leads to a better performance. However, there still exists the risk that the distribution of entropy (probability) of samples from source domain differs from those belong to target domain.

g) L2 normalization layer?: DANCE adopts the architecture that has an L2 normalization layer before the last linear layer. One can regard the weight vectors (without bias) in the last linear layer as prototype features of each class. According to [32], progressively adapting the feature norms of the two domains to a large range of values can result in significant transfer gains, implying that those task-specific features with larger norms are more transferable.

h) estimate parameter b (based on L2 normalization layer): Here are some of my viewpoints: (1) Consider a weight vector $\mathbf{w}_i$ of class i, two feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ that belong to class i and j respectively, the diffience between inner product $\mathbf{w}_i^T \cdot \mathbf{f}_i$ and $\mathbf{w}_i^T \cdot \mathbf{f}_j$ is not significant ($\frac{1}{1+e^{-x/0.05}}$).

i) misestimation of parameter b: Some of the "unknown" samples obtain high predicted confidence, thus parameter b might be inevitably overestimated in some cases.

j) DCC method: DCC method adopts K-means to group features of target domain into K clusters and obtain corresponding centers $\{\mu_1^t, ..., \mu_K^t\}$. For each cluster center, DCC searches for its nearest cluster center in the source domain. If two clusters reach consensus, i.e., both act as the other's nearest center simultaneously, such a pair of clusters is reconized as common clusters. The intuition is simple: cluster centers from the same class usually lie close enough to be associated compared to the clusters representing private classes. DCC achieves comparable performance in different situations, including CDA, PDA, ODA, and OPDA.

REFERENCES

[1] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

[2] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[3] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[4] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, 2019.

[5] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. *arXiv preprint arXiv:2002.08276*, 2020.

[6] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[7] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *arXiv preprint arXiv:1705.08848*, 2017.

[8] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *European Conference on Computer Vision*, pages 567–583. Springer, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[14] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019.

[15] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9757–9766, 2021.

[16] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13944, 2020.

[17] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019.

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[19] Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. *arXiv preprint arXiv:2007.05134*, 2020.

[20] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[22] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[23] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[24] Kuniaki Saito and Kate Saenko. Ovanet: One-vs-all network for universal domain adaptation. *arXiv preprint arXiv:2104.03344*, 2021.

[25] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018.

[26] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *arXiv preprint arXiv:2002.07953*, 2020.

[27] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] Lei Shu, Hu Xu, and Bing Liu. Doc: Deep open classification of text documents. *arXiv preprint arXiv:1709.08716*, 2017.

[29] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[30] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[31] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020.

[32] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1426–1435, 2019.

[33] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019.

[34] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018.