When Domain Adaptation Meets Semi-supervised Learning Through Optimal Transport

Mourad El Hamri, Younès Bennani, Issam Falih

When Domain Adaptation Meets Semi-Supervised Learning Through Optimal Transport

Mourad El Hamri[1,2], Younès Bennani[1,2], and Issam Falih[2,3]

[1] LIPN - CNRS UMR 7030, Université Sorbonne Paris Nord, France
[2] LaMSN - La Maison des Sciences Numériques, France
[3] LIMOS - CNRS UMR 6158, Université Clermont Auvergne, France
{Firstname.Lastname}@{sorbonne-paris-nord,uca}.fr

Abstract. This paper deals with the problem of unsupervised domain adaptation that aims to learn a classifier with a slight target risk while labeled samples are only available in the source domain. The proposed approach, called DA-SSL (Domain Adaptation meets Semi-Supervised Learning) attempts to find a joint subspace of the source and target domains using Linear Discriminant Analysis, such that the projections of the data into this latent subspace can be both domain invariant and discriminative. This aim, however, can be rather difficult to accomplish because of the missing labeled data in the target domain. To defeat this challenge, we use an incremental semi-supervised approach based on optimal transport theory, that conducts selective pseudo-labeling for unlabeled target instances. The selected pseudo-labeled target data are then combined with the source data to incrementally learn a robust classifier in a self-training fashion after the subspace alignment. Experiments show the competitiveness of the proposed approach over contemporary state-of-the-art methods on two benchmark domain adaptation datasets. We make our code publicly available. [4]

Keywords: Domain Adaptation, Semi-supervised learning, Optimal Transport, Self-training, Label Propagation

1 Introduction

Deep learning has achieved spectacular performances in a variety of supervised learning applications. Nevertheless, the tremendous performance growth often relies on the assumption that both training and test data are drawn from the same probability distribution. In many real-world applications, drastic degradation can affect the generalization ability of these models when applying to new shifted domains, where the distributions between training and test data are different. This domain shift is due to many factors like environments, acquisition devices, time, locations, etc. Fine-tuning on labeled target data can

[4] Code is available at: https://github.com/MouradElHamri/DA-SSL

be considered as a feasible solution, but in numerous practical scenarios, data labeling is tremendously grueling, very expensive and time-consuming. To overcome these issues, domain adaptation transfers knowledge from a relevant well labeled source domain in which the model is trained, to a different yet related unlabeled target domain in which the model is deployed. In particular, domain adaptation seeks to build a robust classifier with a low target risk by leveraging labeled source samples to tackles domain shifts in machine learning applications: healthcare diagnostic systems should be adapted to new physical human variations, industrial quality inspection systems must be accurate for new products, self-driving cars have to be able to adapt to new geographical environments and weather conditions, etc. There are two variants of domain adaptation: unsupervised domain adaptation, where labels are only available in the source domain and all target data are unlabeled and semi-supervised domain adaptation, where few labeled data are available in the target domain. In this work, we focus on unsupervised domain adaptation.

Multiple unsupervised domain adaptation approaches have been suggested, a well-known technique emphasizes aligning the source and target domains by learning a lower-dimensional joint subspace. Nevertheless, the latent subspace may not only push the source and target domains closer, but also confuse instances with different class labels. To enhance the discriminative power in the target domain, further directions were pursued by incorporating additional information contained in the unlabeled target data, such as pseudo-labeling. Pseudo-labeling methods are a form of self-training, which is a common algorithmic paradigm for leveraging unlabeled data in the target domain. Self-training methods train a model to fit pseudo-labels, that is, predictions on unlabeled data made by a previously-learned model. However, the distributional shift between source and target domains makes pseudo-labeling quit difficult, since it is subject to error accumulation. To address the powerlessness of the correctly-pseudo-labeled samples to reduce the bias caused by falsely pseudo-labeled data, selective pseudo-labeling can be an effective way to take into consideration the confidence in the target domain, by selecting a subset of target data to be assigned with pseudo labels using confidence threshold or re-weighting, and only these selected pseudo-labeled target data are jointly combined with source labeled data to train the model.

In this paper, we propose an incremental subspace alignment of the conditional distribution of the target domain with that of the source domain using Linear Discriminant Analysis. Nevertheless, this goal can be rather challenging to reach because of the absence of labeled samples in the target domain. To surmount this obstacle, we use an incremental semi-supervised technique based on optimal transport: Optimal Transport Propagation (OTP) [4], that conducts selective pseudo labeling in the target domain. The selected pseudo-labeled target instances are then used in combination with the source data to incrementally learn the subspace alignment and train the classifier in a self-training manner.

The rest of this paper is organized as follows: Section 2 introduces preliminary knowledge on unsupervised domain adaptation settings, self-training for unsupervised domain adaptation and optimal transport theory. Section 3 is devoted to the presentation of our proposed approach. In section 4 a comparative study with state-of-the-art methods is performed on two benchmark datasets. We conclude in section 5.

2 Preliminary Knowledge

2.1 Unsupervised domain adaptation

Unsupervised domain adaptation aims to improve the model generalization performance by transferring knowledge from a labeled source domain to an unlabeled target domain. Formally, we have an input space $\mathcal{X} = \mathbb{R}^d$, a discrete label space $\mathcal{Y} = \{c_1, ..., c_k\}$ composed of k classes and two different probability distributions $\mathcal{S}$ and $\mathcal{T}$ over $\mathcal{X} \times \mathcal{Y}$ called respectively the source and target domains. We observe a set $S = \{(x_i, y_i)\}_{i=1}^n$ of n labeled source data drawn i.i.d. from the joint distribution $\mathcal{S}$ and a set $T = \{x_j\}_{j=1}^m$ of m unlabeled target data drawn i.i.d. from the marginal distribution $\mathcal{T}_\mathcal{X}$ of $\mathcal{T}$ over $\mathcal{X}$:

$$S = \{(x_i, y_i)\}_{i=1}^n \sim (\mathcal{S})^n, \quad T = \{x_j\}_{j=1}^m \sim (\mathcal{T}_\mathcal{X})^m. \tag{1}$$

We assume that the source and target domains share the same label space $\mathcal{Y}$. The objective of unsupervised domain adaptation is to learn a classifier $\eta : \mathcal{X} \to \mathcal{Y}$ with a slight target risk:

$$\mathcal{R}_\mathcal{T}(\eta) = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{T}}(\eta(x) \neq y). \tag{2}$$

In the sequel, we denote by the source domain indifferently the distribution $\mathcal{S}$ and the labeled set S, and by the target domain, the distribution $\mathcal{T}$ and the unlabeled set T.

2.2 Self-training for unsupervised domain adaptation

Self-training is a popular technique that has proven to be very effective for learning with unlabeled data. Self-training algorithms train a model to fit synthetic labels predicted by another previously-learned model.

For unsupervised domain adaptation, pseudo-labeling methods are a form of self-training where the source labels are used to predict pseudo-labels on the unlabeled target data. These methods then train a fresh classifier to fit these pseudo-labels.

The empirical phenomenon that self-training on pseudo-labels often improves over the pseudo-labeler F_{pl} despite no access to true labels has been explained

in the work of [15] by Theorem 1. We first need the following definitions and assumptions.

Transformation set: Let $\mathbf{T}$ be the set of some transformations obtained via data augmentation, the transformation set of x is defined as:

$$\mathcal{B}(x) = \{x^{'} : \exists \, \mathrm{Tr} \in \mathbf{T} \text{ such that } \parallel x^{'} - \mathrm{Tr}(x) \parallel \leq r\} \tag{3}$$

$\mathcal{B}(x)$ is the set of points with distance r from some data augmentation of x.

Neighborhood: The neighborhood of x denoted by $\mathcal{N}(x)$ is the set of points whose transformation sets overlap with that of x:

$$\mathcal{N}(x) = \{x^{'} : \mathcal{B}(x) \cap \mathcal{B}(x^{'}) \neq \emptyset\} \tag{4}$$

For $S \subset \mathcal{X}$, the neighborhood of S is defined as the union of neighborhoods of its elements: $\mathcal{N}(S) = \underset{x \in S}{\cup} \mathcal{N}(x)$.

Assumption 1 ((a,c)-expansion): Let P be the distribution of unlabeled target data, and P_i for $i \leq k$ be the class-conditional distribution of $x \in \mathcal{X}$ conditioned on the class c_i. We say that the class-conditional distribution P_i satisfies (a, c)-expansion if for all $V \subset \mathcal{X}$ with $P_i(V) \leq a$, the following holds:

$$P_i(\mathcal{N}(\mathcal{V})) \geq \min\{cP_i(V), 1\} \tag{5}$$

If P_i satisfies (a, c)-expansion for all $i \leq k$, then we say P satisfies (a, c)-expansion.

Population consistency loss: We define the population consistency loss $R_{\mathcal{B}}(F)$ as the fraction of examples where a classifier F is not robust to input transformations:

$$R_{\mathcal{B}}(F) = \mathbb{E}_P[\mathbb{1}(\exists x^{'} \in \mathcal{B}(x) \text{ such that } F(x^{'}) \neq F(x))] \tag{6}$$

Assumption 2 (Separation): We assume P is $\mathcal{B}$-separated with probability $1 - \mu$ by ground-truth classifier F^*, as follows: $R_{\mathcal{B}}(F^*) \leq \mu$.

Assumption 3: Define $\bar{a} = \underset{i \leq k}{max}\{P_i(\mathcal{M}(F_{pl}))\}$ to be the maximum fraction of incorrectly pseudo-labeled examples in any class: $\mathcal{M}(F_{pl})) = \{x : F_{pl}(x) \neq F^*(x)\}$. We assume that $\bar{a} < \frac{1}{3}$ and P satisfies $(\bar{a}, \bar{c})$-expansion for $\bar{c} > 3$.

Theorem 1. *Define* $c = \min\{\frac{1}{\bar{a}}, \bar{c}\}$. *Suppose Assumptions 3 and 2 hold. Then for any minimizer* $\hat{F}$ *of* $\mathcal{L}(F) = \frac{c+1}{c-1} L_{0-1}(F, F_{pl}) + \frac{2c}{c-1} R_{\mathcal{B}}(F) - Err(F_{pl})$, *we have:*

$$Err(\hat{F}) \leq \frac{2}{c-1} Err(F_{pl}) + \frac{2c}{c-1}\mu. \tag{7}$$

Which explains the perhaps surprising fact that self-training with pseudo-labeling often improves over the pseudo-labeler F_{pl} even though no additional information about true labels is provided.

2.3 Optimal Transport

The birth of optimal transport is dated back to 1781, with the following problem introduced by Gaspard Monge [10]: Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ a measurable cost function, the problem of Monge aims at finding the transport map $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$, that transport the mass represented by the measure μ to the mass represented by the measure ν and which minimizes the total cost of this transportation, more formally:

$$\inf_{\mathcal{T}} \{ \int_{\mathcal{X}} c(x, \mathcal{T}(x)) d\mu(x) | \mathcal{T}\#\mu = \nu \}, \tag{8}$$

where $\mathcal{T}\#\mu$ denotes the push-forward operator of μ through the map $\mathcal{T}$.

A long period of sleep followed Monge's formulation until the relaxation of Leonid Kantorovitch in 1942 [8]. The relaxed formulation of Kantorovich, known as the Monge-Kantorovich problem, can be formulated in the following way:

$$\inf_{\gamma} \{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\gamma(x, y) \, | \, \gamma \in \mathit{\Pi}(\mu, \nu) \, \}, \tag{9}$$

where $\mathit{\Pi}(\mu, \nu)$ is the set of probability measures over the product space $\mathcal{X} \times \mathcal{Y}$ such that both marginals of γ are μ and ν.

In several real world applications, the access to the measures μ and ν is only available through finite samples $X = (x_1, ..., x_n) \subset \mathcal{X}$ and $Y = (y_1, ..., y_m) \subset \mathcal{Y}$, then, the measures μ and ν can be casted as the following discrete measures, $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, where $a \in \sum_n$ and $b \in \sum_m$ are probability vectors of size n and m respectively. The relaxation of Kantorovich becomes then the following linear program [12]:

$$\min_{\gamma \in U(a,b)} \langle \gamma, C_{XY} \rangle_F \tag{10}$$

where $U(a, b) = \{\gamma \in \mathcal{M}_{n \times m}(\mathbb{R}^+) \, | \, \gamma \mathbb{1}_m = a \text{ and } \gamma^{\mathbf{T}} \mathbb{1}_n = b\}$ is the transportation polytope which acts as a feasible set, C_{XY} is the cost matrix and $\langle \gamma, C_{XY} \rangle_F = trace(\gamma^{\mathrm{T}} C_{XY})$ is the Frobenius dot-product of matrices.

This linear program, can be solved with the simplex algorithm or interior point methods. However, optimal transport problem scales cubically on the sample size, which is often too costly in practice, especially for machine learning applications that involve massive datasets. Entropy-regularization [3] has emerged as a solution to the computational burden of optimal transport. The entropy-regularized discrete optimal transport problem reads:

$$\min_{\gamma \in U(a,b)} \langle \gamma, C_{XY} \rangle_F - \varepsilon \mathcal{H}(\gamma) \tag{11}$$

where $\mathcal{H}(\gamma) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}(\log(\gamma_{ij})-1)$ is the entropy of γ. This regularized problem can be solved efficiently via an iterative procedure: Sinkhorn-Knopp algorithm.

3 Proposed Approach

The proposed method aims to learn a joint subspace from the source and target domains such that the projected data into the subspace are domain invariant and well separated. To accomplish this aim, linear discriminant analysis (LDA) appears to be a good candidate for many reasons, principally for its capacity to find a linear combination of features, which separates two or more classes of data no matter the domain they come from, providing an appropriate approach for the unsupervised domain adaptation problem, where the source and target data come from different distributions. Nonetheless, LDA needs labeled data to learn the projection matrix. To surmount this challenge, we use pseudo-labels in the target domain produced by a semi-supervised technique (OTP). The reason for choosing OTP is its ability to capture the geometry of data thanks to optimal transport. Furthermore this technique falls into the class of selective pseudo-labeling methods. The benefit of these methods is that they avoid mislabeled target instances from impeding the subspace learning process by spreading the errors to the next iteration which can reduce the robustness of the learned classifier. Thus, we use the labeled source data and the selected pseudo-labeled target data provided by OTP to incrementally learn a robust classifier in a self-training fashion after the subspace alignment.

3.1 Domain Alignment via Linear discriminant analysis

To learn a domain-invariant and discriminative subspace $\tilde{\mathcal{X}}$ from $\mathcal{X}$, we employ Linear Discriminant Analysis (LDA), which is a common technique used for dimensionality reduction. LDA can also provides class separability by drawing a decision region between the different classes.

Let $X \in \mathcal{M}_{d,N}(\mathbb{R})$ be a labeled data matrix composed of N samples. Basically, LDA seeks to find a projection matrix W for which the low-dimensional projection of X yields a cloud of points that are close when they are in the same class relative to the overall spread. This projection matrix can be found by maximizing the Rayleigh quotient of the within scatter matrix S_w and between scatter matrix S_b:

$$W = \underset{V}{\operatorname{argmax}} \frac{|V^T S_b V|}{|V^T S_w V|} \tag{12}$$

The maximization problem in Eq(12) is equivalent to the following generalized eigenvalue problem:

$$S_b w = \lambda S_w w \tag{13}$$

The eigenvectors of Eq(13) represent the directions of the lower-dimensional feature space learned by LDA, and the corresponding eigenvalues represent the

ability of the eigenvectors to discriminate between different classes, i.e. increase the between-class variance, and decreases the within-class variance of each class. The eigenvectors with the d_1 highest eigenvalues give us the LDA projection matrix $W_{=}[w_1, ..., w_{d_1}] \in \mathcal{M}_{d,d_1}(\mathbb{R})$, from which we can learn the lower-dimensional discriminant representation $\tilde{X} \in \mathcal{M}_{d_1,N}(\mathbb{R})$:

$$\tilde{X} = W^T X \tag{14}$$

3.2 Self-Training via Optimal Transport Propagation

To learn a domain-invariant and discriminative subspace $\tilde{\mathcal{X}}$ from $\mathcal{X}$ using the projection matrix W of LDA we need labeled data as stated above. Nevertheless, in unsupervised domain adaptation setting, labeled data in the target domain are unavailable. To addresse this limitation, we propose to use a semi-supervised learning approach called Optimal Transport Propagation (OTP), able to perform selective pseudo-labeling in the target domain.

3.2.1 Optimal Transport Propagation

Optimal transport propagation (OTP) [4] [5] is a transductive semi-supervised method which relies on optimal transport theory to propagate labels between the vertices of a complete bipartite edge-weighted graph in two phases.

Let X_L be a finite ordered set of l labeled samples $\{(x_1, y_1), ..., (x_l, y_l)\}$. Each example (x_i, y_i) of this set consists of a sample x_i from an input space $\mathcal{X}$, and its corresponding label $y_i \in \mathcal{Y} = \{c_1, ..., c_k\}$ where $\mathcal{Y}$ is a discrete label set composed of k classes, and X_U a larger collection of u instances $\{x_{l+1}, ..., x_u\}$, whose labels Y_U are unknown. In the graph construction phase, the first part of the graph $\mathcal{L}$ is composed of labeled data $\mathcal{L} = X_L$ and the second part $\mathcal{U}$ is composed of unlabeled data $\mathcal{U} = X_U$. To compute the edge-weights of the graph, authors suggest to solve the regularized optimal transport problem between the empirical distribution of X_L and X_U:

$$\gamma_\varepsilon^* = \underset{\gamma \in U(a,b)}{\operatorname{argmin}} \ \langle \gamma, C \rangle_F - \mathcal{H}(\gamma), \tag{15}$$

where C denotes the cost matrix defined by: $c_{i,j} = \|x_i - x_j\|^2, \ \forall (x_i, x_j) \in \mathcal{L} \times \mathcal{U}$. The optimal transport plan γ_ε^* can be interpreted as a similarity matrix between the two parts $\mathcal{L}$ and $\mathcal{U}$ of the graph $\mathcal{G}$. To have a class probability interpretation, the matrix γ_ε^*, is normalized to get the affinity matrix $\mathcal{W}$ defined as follows:

$$w_{i,j} = \frac{\gamma^*_{\varepsilon_{i,j}}}{\sum_i \gamma^*_{\varepsilon_{i,j}}}, \ \forall i, j \in \{1, ..., l\} \times \{l+1, ..., l+u\}, \tag{16}$$

where $w_{i,j}, \ \forall i, j \in \{1, ..., l\} \times \{l+1, ..., l+u\}$ is then, the probability of jumping from the vertex $x_i \in \mathcal{L}$ to $x_j \in \mathcal{U}$.

In the second phase, a label matrix U is constructed from the affinity matrix $\mathcal{W}$ to denotes the probability of each unlabeled data x_j to belong to every class c_h. This probability is defined as the sum of the similarity of x_j with the representatives of the class c_h, formally:

$$u_{j,h} = \mathbb{P}(x_j \in c_h) = \sum_{i/x_i \in c_h} w_{i,j}, \forall j, h \in \{l+1, ..., l+u\} \times \{1, ..., k\}, \tag{17}$$

To avoid hard assignment of labels directly from the label matrix U, and the consequent neglect of the different degrees of certainty for each prediction. A certainty score is associated with each pseudo label in the following way:

$$s_j = 1 - \frac{H(U_j)}{log_2(k)}, \quad \forall j \in \{l+1, ..., l+u\}, \tag{18}$$

where U_j is the j^{th} row of the label matrix U, which corresponds to the stochastic vector that encodes the probability of x_j to belong to the different classes, and H is Shannon's entropy.

To endow OTP with the selectivity property during the propagation process, a comparison is made between the certainty score corresponding to each unlabeled instance x_j and a confidence threshold $\alpha \in [0, 1]$. If the value of the score s_j is superior to α, x_j is then labeled in the following way:

$$\hat{y_j} = \underset{c_h \in \mathcal{C}}{\operatorname{argmax}}\, u_{j,h}, \ \forall j \in \{l+1, ..., l+u\}, \tag{19}$$

Thus, the unlabeled instance x_j will belong to the most likely class c_h. Otherwise, x_j does not receive any label. This process corresponds to one iteration of the incremental approach OTP. In each iteration, the labeled set X_L is enriched with new points from X_U, and the number of samples in X_U is reduced, until convergence, i.e. when all the data initially in X_U are labeled, or, in other words when X_U is reduced to the empty set $\emptyset$.

3.2.2 Domain Adaptation via Optimal Transport Propagation

Once the LDA projection matrix W is learned (at the first iteration, the projection matrix is learned using only the labeled source data), the projection of both source samples S and target samples T in the joint subspace can be obtained as follows:

$$\tilde{S} = W^T S \ \text{ and } \ \tilde{T} = W^T T \tag{20}$$

Pseudo-labeling in the target domain can then be performed using OTP considering that:

$$X_L = \tilde{S} \ \text{ and } \ X_U = \tilde{T} \tag{21}$$

The intuition behind the use of OTP as a pseudo-labeling technique is its capability to capture the geometry of the underlying subspace and its selective

ability based on the incorporated certainty score which make it closely related to entropy minimization, where the model's predictions are encouraged to be low-entropy (i.e., high-confidence) on unlabeled data. Thus, instead of using all the pseudo-labeled target samples to learn the next projection, we incrementally select a subset $\tilde{T}_p \subset \tilde{T}$ that contains an amount of p pseudo-labeled target samples with the highest certainty score. Nevertheless, this technique has the potential risk to only select instances from particular classes and to overlook the other classes. To prevent this issue, we conduct a class-wise selection in order to ensure that pseudo-labeled target samples of each class have an equal opportunity to be selected. Precisely, for each class c_h, $\forall h \in \{1, ..., k\}$ we select $\frac{p}{k}$ target samples pseudo-labeled as class c_h.

Thereafter, the projected source data is combined with the selected pseudo-labeled target data to form a new augmented source domain, simultaneously, the pseudo-labeled target data must be retired from the target domain in the following way:

$$S \leftarrow \tilde{S} \cup \tilde{T}_p \quad \text{and} \quad T \leftarrow \tilde{T} \setminus \tilde{T}_p \tag{22}$$

Equations 22 are used to incrementally update the source and target domains. At each iteration, a classifier η is trained on the augmented source samples in a self-training manner. The intuition behind this idea is that at each iteration the classifier becomes more and more robust, since it is trained on both the source data and the selected pseudo-labeled target data, so that in the last iteration, it will be trained on the source samples and the totality of pseudo-labeled target instances, allowing it to improve its accuracy according to Theorem 1.

Algorithm 1: DA-SSL

Parameters: Dimensionality of LDA d_1, sampling rate p
Input : Labeled source data S, Unlabelled target data T
while *not converged* **do**
  Learn the projection W using source data S
  Get the projected source and target samples $\tilde{S}$ and $\tilde{T}$
  Assign pseudo labels for the projected target data $\tilde{T}$ using OTP
  Select a subset of pseudo-labeled target data $\tilde{T}_p$
  Update the source domain $S \leftarrow \tilde{S} \cup \tilde{T}_p$
  Update the target domain $T \leftarrow \tilde{T} \setminus \tilde{T}_p$
  Learn a classifier η on S
end
return *Predicted labels of the original target data T using η*

The overall algorithm called DA-SSL (Domain Adaptation meets Semi-Supervised Learning) is summarized in Algorithm 1.

4 Experiments

In this section, we provide empirical experimentation for the proposed algorithm.

4.1 Datasets

We adopt two datasets that are benchmarks in domain adaptation: ImageCLEF-DA and Office31 .

ImageCLEF-DA dataset [2] consists of four domains. We use three of them in our experiments: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). There are 12 classes and 50 images for each class in each domain.

Office31 dataset [13] composed of 4110 images. The dataset consists of three domains: Amazon, Webcam and DSLR, 31 common classes from the three domains are used.

4.2 Experimental Protocol

We use ResNet50 [7] features (d = 2048) for ImageCLEF-DA and Office31 datasets. Our proposed approach consists of two hyper-parameters, the dimensionality d_1 of LDA that we set equal to 128 and the sampling rate p that we set equal to 48 for ImageCLEF-DA and 62 for Office31 dataset. We use an SVM with a Gaussian kernel as classifier [1]. The width parameter of the SVM was chosen as $\sigma = \frac{1}{2\mathbb{V}}$, where $\mathbb{V}$ is the variance of the source samples.
Following the standard protocol [6], the comparison is conducted using three deep learning models RTN [9], MADA [11] and iCAN [16], and with a manifold embedded distribution alignment technique based on deep features MEDA [14]. We use the average accuracy as the evaluation metric in all our experiments.

4.3 Results

We use bold and underlined fonts to indicate the best and the second best results respectively. The classification accuracy of our proposed approach and other baseline methods are illustrated in Table 1 and Table 2, from which we can see that our proposed approach achieves the highest average accuracy over the two benchmark datasets. Specifically, DA-SSL achieves an average accuracy of 89.4% on ImageCELF-DA dataset (Table 1), slightly better than MEDA which has an average accuracy of 89.0%. On the Office31 dataset (Table 2), DA-SSL outperforms all other baseline models with an average accuracy of 87.6% against 85.7% by MEDA and 87.2% by iCAN, besides, DA-SSL achieves the best performance in three out of six tasks and the second-best results in two other tasks. In summary, the proposed approach is highly competitive compared to several state-of-the-art methods, and can outperform competitors on many tasks of the two domain adaptation problems. This results are mainly attributed to the capacity of OTP to capture mush more information than the other methods of

Table 1. Classification Accuracy (%) on ImageCELF-DA dataset (ResNet50 features).

Task	RTN	MADA	iCAN	MEDA	DA-SSL
I → P	75.6	75.0	79.5	**79.7**	78.9
P → I	86.8	87.9	89.7	**92.5**	91.8
I → C	95.3	96.0	94.7	95.7	**97.8**
C → I	86.9	88.8	89.9	92.2	**92.6**
C → P	72.7	75.2	**78.5**	**78.5**	78.2
P → C	92.2	92.2	92.0	95.5	**95.8**
average	84.9	85.8	87.4	89.0	**89.4**

Table 2. Classification Accuracy (%) on Office31 dataset (ResNet50 features).

Task	RTN	MADA	iCAN	MEDA	DA-SSL
A → W	84.5	90.0	92.5	86.2	**93.3**
D → W	96.8	97.4	98.8	97.2	**99.0**
W → D	99.4	99.6	**100.0**	99.4	99.6
A → D	77.5	87.8	90.1	85.3	**90.7**
D → A	66.2	70.3	72.1	72.4	71.9
W → A	64.8	66.4	69.9	**74.0**	71.3
average	81.6	85.2	87.2	85.7	**87.6**

pseudo-labeling thanks to the enhanced affinity matrix constructed by optimal transport and to its intrinsic property of selectivity which make it a good candidate for pseudo-labeling target data.

5 Conclusion

In this paper, a novel selective pseudo-labeling approach for unsupervised domain adaptation called DA-SSL is proposed. DA-SSL learn a domain-invariant and discriminative subspace by LDA using labelled source data and pseudo-labeled target data. The pseudo-labeling is performed by OTP, a certainty-aware semi-supervised method that selects samples with high confidence to participate in the next iteration of the incremental learning process. In each iteration of the incremental domain adaptation process, a classifier is learned using the augmented source data composed of the samples in the original source domain and the accumulation of the pseudo-labeled target data in the previous iterations. The proposed approach outperforms several state-of-the-art methods on two benchmark datasets.

References

1. Khalid Benabdeslem and Younès Bennani. Dendrogram-based svm for multi-class classification. *Journal of Computing and Information Technology*, 14(4):283–289, 2006.
2. Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, et al. Imageclef 2014: Overview and analysis of the results. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211. Springer, 2014.
3. Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.
4. Mourad El Hamri, Younés Bennani, and Issam Falih. Label propagation through optimal transport. In *2021 International Joint Conference on Neural Networks*.
5. Mourad El Hamri, Younès Bennani, and Issam Falih. Inductive semi-supervised learning through optimal transport. In *International Conference on Neural Information Processing*, pages 668–675. Springer, 2021.
6. Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073, 2012.
7. 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.
8. Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.
9. Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016.
10. Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.
11. Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
12. Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019.
13. 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.
14. Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018.
15. Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.
16. Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3801–3809, 2018.