Learning Discriminative Correlation Subspace for Heterogeneous Domain Adaptation

Yuguang Yan[1*], **Wen Li**[2*], **Michael Ng**[3], **Mingkui Tan**[1], **Hanrui Wu**[1], **Huaqing Min**[1], **and Qingyao Wu**[1†]
[1]School of Software Engineering, South China University of Technology, China
[2]Computer Vision Laboratory, ETH Zurich, Switzerland
[3]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
yan.yuguang@mail.scut.edu.cn, liwen@vision.ee.ethz.ch, qyw@scut.edu.cn

Abstract

Domain adaptation aims to reduce the effort on collecting and annotating target data by leveraging knowledge from a different source domain. The domain adaptation problem will become extremely challenging when the feature spaces of the source and target domains are different, which is also known as the *heterogeneous domain adaptation* (HDA) problem. In this paper, we propose a novel HDA method to find the optimal discriminative correlation subspace for the source and target data. The discriminative correlation subspace is inherited from the canonical correlation subspace between the source and target data, and is further optimized to maximize the discriminative ability for the target domain classifier. We formulate a joint objective in order to simultaneously learn the discriminative correlation subspace and the target domain classifier. We then apply an alternating direction method of multiplier (ADMM) algorithm to address the resulting non-convex optimization problem. Comprehensive experiments on two real-world data sets demonstrate the effectiveness of the proposed method compared to the state-of-the-art methods.

1 Introduction

In traditional machine learning, people usually have to collect as more labeled data as possible to achieve good classification performance in a target domain of interest. However, the collection and annotation of data can be very expensive and time-consuming. To tackle this issue, domain adaptation (DA) is proposed to utilize auxiliary data from another domain, a.k.a., source domain, to enhance the performance of the target task [Pan and Yang, 2010; Patel *et al.*, 2015]. DA has been successfully applied in many real-world applications, including text classification [Chen and Zhang, 2013], visual recognition [Duan *et al.*, 2010; 2012b],

To build the connection between two different feature spaces, a usual way is to learn a feature transformation mapping to project the heterogeneous source and target data into the same space [Duan *et al.*, 2012a; Li *et al.*, 2014; Hubert Tsai *et al.*, 2016]. Existing state-of-the-art methods mainly focus on reducing the domain distribution mismatch when learning the feature transformation and the target classifier. For example, SHFA [Li *et al.*, 2014] employs feature augmentation strategy, while CDLS [Hubert Tsai *et al.*, 2016] learns landmarks based on the maximum mean discrepancy.

When learning the feature transformation to bridge the two heterogeneous feature spaces, one critical issue is how to preserve the useful information due to the feature transformation. However, if important information is lost after the feature transformation, it will be very difficult to boost the learning performance over the target data using the transformed source features, even the domain distribution mismatch is reduced by some efforts in the existing HDA approaches [Li *et al.*, 2014; Hubert Tsai *et al.*, 2016]. To this end, in this paper, we propose a new algorithm called *Discriminative Correlation Analysis (DCA)* to deal with the heterogeneous domain adaptation problem by learning a discriminative subspace that preserves the maximum feature correlation between the source and target domain data. On one hand, the learned subspace should maximize the feature correlation as measured in the canonical correlation analysis (CCA) method. On the other hand, it also ensures good discriminative ability such that the empirical training loss over target domain data is minimized.

We formulate a joint objective to learn the target classifier and the correlation subspace simultaneously, which, however, leads to a non-trivial optimization problem. To address it, we take advantage of the characteristic of general solutions to CCA, and represent the set of optimal projection matrices based on an orthogonal matrix and two arbitrary matrices (See Section 3.3 for more details). By optimizing a transformed objective, we seek to find the best projection matrix with discriminative power so that the feature correlation between two domains can be guaranteed to be maximized when learning the target domain classifier. Interestingly, the resultant optimization problem becomes much easier to address, since we have only a single orthogonal constraint in the reformulated objective. We then use an alternating direction method of multipliers (ADMM) algorithm to solve the resulting problem. We evaluate our proposed DCA approach on

[*]The co-first authors.

[†]The corresponding author.

two real-world benchmark data sets.

2 Related Work

Homogeneous domain adaptation has been extensively studied in the last decade [Pan and Yang, 2010; Patel *et al.*, 2015]. Recently, heterogeneous domain adaptation (HDA) has been attracting more and more attention. The principal challenge of HDA is that the feature spaces of the source and target domains are completely different, making it difficult to leverage the source data to assist the target task [Zhou *et al.*, 2014a; 2016]. Generally, the existing HDA algorithms can be divided into two categories.

The first group of HDA algorithms utilizes auxiliary data to bridge source and target domains [Yang *et al.*, 2015]. Text-image pairs are used to perform image classification tasks in [Zhu *et al.*, 2011]. In [Wu *et al.*, 2014], researchers constructed transition probabilities via co-occurrence data to enhance classification tasks. Tan *et al.* [Tan *et al.*, 2015] proposed to select informative intermediate domains to connect source and target domains. However, collecting auxiliary data brings extra cost, and it might be difficult to obtain co-occurrence data in some applications like object recognition.

The second group of HDA algorithms performs feature transformation to connect two different feature spaces [Shi *et al.*, 2010; Wang and Mahadevan, 2011; Kulis *et al.*, 2011; Wu *et al.*, 2013; Hoffman *et al.*, 2014; Zhou *et al.*, 2014b; Xiao and Guo, 2015]. Among them, most algorithms were designed for the supervised HDA scenario, where only labeled training instances are available in the target domain. The recent proposed SHFA [Li *et al.*, 2014] and CDLS [Hubert Tsai *et al.*, 2016] methods show that the unlabeled target domain instances are helpful for HDA tasks, and achieve state-of-the-art performance on the benchmark data sets. In particular, SHFA [Duan *et al.*, 2012a; Li *et al.*, 2014] augments labeled source and target data based on two projection matrices, and trained a SVM classifier on the augmented data. CDLS [Hubert Tsai *et al.*, 2016] finds representative cross-domain landmarks to derive a domain-invariant feature subspace, and then train a classifier in the learned subspace. Nevertheless, those two methods did not consider the feature correlation between the source and target domain. In contrast, we find the optimal discriminative subspace that is ensured to maximize the feature correlation, thus our method is more effective for transferring the source domain knowledge for learning the target domain classifier.

3 Methodology

3.1 Problem Definition and Notation

In the heterogeneous domain adaptation problem, we have a source domain and a target domain that are represented by different features. The task is to learn a classifier for predicting the target domain instances.

Let us denote by $\mathcal{X}_S = \{(\mathbf{x}_{S,i}, y_{S,i})|_{i=1}^{n_S}\}$ the source domain, where n_S is the number of source domain instances, $\mathbf{x}_{S,i} \in \mathbb{R}^{d_S}$ is the i-th training instance with d_S being the source feature dimension, and $y_{S,i} \in \mathcal{Y}$ is the label of $\mathbf{x}_{S,i}$ with $\mathcal{Y} = \{1, \ldots, K\}$ being the label space and K being the number of categories.

Similarly, we denote by $\mathcal{X}_L = \{(\mathbf{x}_{L,i}, y_{L,n_L})|_{i=1}^{n_L}\}$ the target domain labeled instances, where n_L is the number of labeled instances, $\mathbf{x}_{L,i} \in \mathbb{R}^{d_T}$ with d_T being the target feature dimension, and $y_{L,i} \in \mathcal{Y}$ is the label of $\mathbf{x}_{L,i}$. Following [Li *et al.*, 2014; Hubert Tsai *et al.*, 2016], we also consider the semi-supervised situation where some unlabeled instances are available in the target domain, which is denoted by $\mathcal{X}_U = \{(\mathbf{x}_{U,i})|_{i=1}^{n_U}\}$, where $\mathbf{x}_{U,i} \in \mathbb{R}^{d_T}$. We also denote by $n_T = n_L + n_U$ the total number of instances in the target domain.

In the reminder of this paper, for any matrix $\mathbf{A}$, $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$, and the MATLAB style notation $\mathbf{A}(:, a : b)$ is the submatrix including the elements of all the rows and the a-th to b-th columns.

3.2 Proposed Model

The core issue of the HDA problem is how to learn a feature transformation, such that the source domain training data can be effectively utilized for the learning tasks in the heterogeneous target domain. To ensure that sufficient information is preserved during this process, inspired by the canonical correlation analysis (CCA), we aim to maximize the feature correlation when learning the feature transformation for HDA tasks.

Given two kinds of features, canonical correlation analysis (CCA) finds two orthogonal projection matrices $\mathbf{P} \in \mathbb{R}^{d_T \times d_C}$ and $\mathbf{Q} \in \mathbb{R}^{d_S \times d_C}$, such that the correlations after projection are maximized, where d_C is the dimension after projection. Since CCA usually considers the same number of two sets of data, we use n source instances $\mathbf{X}_S = [\mathbf{x}_{S,1}, \ldots, \mathbf{x}_{S,n}] \in \mathbb{R}^{d_S \times n}$ from $\mathcal{X}_S$ and n target instances $\mathbf{X}_T = [\mathbf{x}_{T,1}, \ldots, \mathbf{x}_{T,n}] \in \mathbb{R}^{d_T \times n}$ from $\mathcal{X}_L \cup \mathcal{X}_U$ to perform CCA, where $n = \min\{n_S, n_T\}$.

Mathematically, the optimization problem of CCA is

$$\begin{aligned} \max_{\mathbf{P},\mathbf{Q}} \quad & \mathrm{trace}\big(\mathbf{P}^\top \mathbf{X}_T \mathbf{X}_S^\top \mathbf{Q}^\top\big), \\ \text{s.t.} \quad & \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I}, \end{aligned} \tag{1}$$

where $\mathbf{I}$ is an identity matrix. It can be equivalently written as a minimization problem as follows,

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q}} \quad & ||\mathbf{X}_T^\top \mathbf{P} - \mathbf{X}_S^\top \mathbf{Q}||_F^2, \\ \text{s.t.} \quad & \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I}. \end{aligned} \tag{2}$$

Although CCA finds the projection matrices that maximize the correlation, it does not utilize the label information of the target data. Therefore, the discriminative capacity of the projected features cannot be guaranteed, which is critical for learning a good classifier in the target domain. To this end, we propose to maximize the canonical correlation between two domains, and simultaneously minimize the regularized training loss on the labeled target data, which is given as follows.

$$\begin{aligned} \min_{\mathbf{w},\mathbf{P},\mathbf{Q}} & C \sum_{i=1}^{n_L} \xi_{L,i} + \frac{1}{2}||\mathbf{w}||^2 + \frac{\gamma}{2}||\mathbf{X}_T^\top \mathbf{P} - \mathbf{X}_S^\top \mathbf{Q}||_F^2, \\ \text{s.t.} \ & \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I}, \end{aligned} \tag{3}$$

where $\mathbf{w}$ is the parameters for the target classifier, $\xi_{L,i}$ is the loss on the i-th labeled target instance, C and γ are the trade-off parameters.

In particular, the cross-entropy loss is used in our model due to its smoothness and the ease of dealing with multi-class problems. To deal with the domain distribution mismatch, inspired by [Li *et al.*, 2014], we augment each labeled target instance by its projected feature vector when learning the target classifier, *i.e.*, the i-th augmented target instance is given by $\tilde{\mathbf{x}}_{L,i} = [\mathbf{x}_{L,i}^\top, (\mathbf{P}^\top \mathbf{x}_{L,i})^\top]^\top \in \mathbb{R}^{d_C+d_T}$, where $\mathbf{P}$ is the projection matrix. Then, the cross-entropy loss function is given by

$$\xi_{L,i} = \log\big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \tilde{\mathbf{x}}_{L,i} - \mathbf{w}_{y_{L,i}}^\top \tilde{\mathbf{x}}_{L,i})\big), \quad (4)$$

where $\mathbf{w}_k \in \mathbb{R}^{d_C+d_T}$ is the parameter vector for the k-th class. Also, the parameter $\mathbf{w}$ in Eq. (3) for target classifier can be written as

$$\mathbf{w} = [\mathbf{w}_1^\top, \ldots, \mathbf{w}_k^\top, \ldots, \mathbf{w}_K^\top]^\top \in \mathbb{R}^{(d_C+d_T)K}. \quad (5)$$

After $\mathbf{w}$ is learned, the label of the i-th unlabeled target instance $\mathbf{x}_{U,i}$ can be predicted by

$$\hat{y}_{U,i} = \arg\max_k \mathbf{w}_k^\top \tilde{\mathbf{x}}_{U,i}, \quad (6)$$

where $\tilde{\mathbf{x}}_{U,i}$ is an unlabeled augmented target instance.

To utilize the unlabeled instances in the target domain, we further use the current $\mathbf{w}$ to predict the unlabeled target instances in the training process. Then, we also minimize the cross-entropy loss for the unlabeled target instances with the pseudo labels, which is given as

$$\begin{aligned} \min_{\mathbf{w},\mathbf{P},\mathbf{Q}} \; & C\sum_{i=1}^{n_L} \xi_{L,i} + C\sum_{i=1}^{n_U} \xi_{U,i} + \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\|\mathbf{X}_T^\top \mathbf{P} - \mathbf{X}_S^\top \mathbf{Q}\|_F^2, \\ \text{s.t.} \; & \mathbf{P}^\top \mathbf{X}_T \mathbf{X}_T^\top \mathbf{P} = \mathbf{I}, \quad \mathbf{Q}^\top \mathbf{X}_S \mathbf{X}_S^\top \mathbf{Q} = \mathbf{I}, \end{aligned} \quad (7)$$

where $\xi_{U,i}$ is the loss on the i-th unlabeled target instance with the predicted label $\hat{y}_{U,i}$. We iteratively refine the pseudo labels, thus improve the learned subspace and the classifier.

3.3 Discriminative Correlation Subspace

Problem (7) is nontrivial to solve because of the non-convex objective function and the complicated orthogonal constraints. A suboptimal solution might be obtained, but good feature correlation and discriminative capacity can hardly be guaranteed.

To this end, we first find the set of subspaces that maximizes the canonical correlation term in (7), and then seek for the most discriminative subspace in the feasible set. To deal with the complicated orthogonal constraints, we take advantage of the characteristic of general solutions to CCA [Chu *et al.*, 2013], which shows that $\mathbf{P}$ and $\mathbf{Q}$ can be represented by an orthogonal matrix, and two arbitrary matrices. Thus, we reformulate the problem (7) as a new one with only a simple constraint.

In particular, the characteristic of CCA is provided in the following theorem,

Theorem 1 *Let the reduced SVDs of* $\mathbf{X}_T$ *and* $\mathbf{X}_S$ *be*

$$\mathbf{X}_T = [\mathbf{U}_{T,1}\mathbf{U}_{T,2}] \begin{bmatrix} \mathbf{\Sigma}_T \\ \mathbf{0} \end{bmatrix} \mathbf{V}_T^\top = \mathbf{U}_{T,1}\mathbf{\Sigma}_T\mathbf{V}_T^\top, \quad (8)$$

and

$$\mathbf{X}_S = [\mathbf{U}_{S,1}\mathbf{U}_{S,2}] \begin{bmatrix} \mathbf{\Sigma}_S \\ \mathbf{0} \end{bmatrix} \mathbf{V}_S^\top = \mathbf{U}_{S,1}\mathbf{\Sigma}_S\mathbf{V}_S^\top, \quad (9)$$

respectively. In addition, let

$$\mathbf{Z}_1 \mathbf{\Sigma} \mathbf{Z}_2^\top = \mathbf{V}_T^\top \mathbf{V}_S \quad (10)$$

be the SVD of $\mathbf{V}_T^\top \mathbf{V}_S$.

Under mild conditions, the solutions to Problem (2) can be represented by the following equations

$$\begin{aligned} \mathbf{P} &= \mathbf{A}_T\mathbf{\Theta} + \mathbf{B}_T\mathbf{\Phi}_T, \quad (11) \\ \mathbf{Q} &= \mathbf{A}_S\mathbf{\Theta} + \mathbf{B}_S\mathbf{\Phi}_S, \quad (12) \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_T &= \mathbf{U}_{T,1}\mathbf{\Sigma}_T^{-1}\mathbf{Z}_1(:, 1:d_C), \quad \mathbf{B}_T = \mathbf{U}_{T,2}, \\ \mathbf{A}_S &= \mathbf{U}_{S,1}\mathbf{\Sigma}_S^{-1}\mathbf{Z}_2(:, 1:d_C), \quad \mathbf{B}_S = \mathbf{U}_{S,2}. \end{aligned}$$

Here, $\mathbf{\Theta}$ *is an orthogonal matrix, and* $\mathbf{\Phi}_T$ *and* $\mathbf{\Phi}_S$ *are arbitrary matrices.*

For the proof, please refer to Section 3 in [Chu *et al.*, 2013].

Remark 1 *Theorem 1 states that given the matrices* $\mathbf{A}_T, \mathbf{B}_T, \mathbf{A}_S$ *and* $\mathbf{B}_S$*, as long as* $\mathbf{\Theta}$ *is orthogonal, the matrices* $\mathbf{P}$ *and* $\mathbf{Q}$ *obtained by Eqs. (11) and (12) are the solutions to Problem (2) for any* $\mathbf{\Phi}_T$ *and* $\mathbf{\Phi}_S$*. In particular, these solutions have the same objective value of Problem (2).*

Based on Theorem 1, instead of solving Problem (7) directly, we optimize $\mathbf{P}$ and $\mathbf{Q}$ from the feasible set of the CCA problem, such that the learned $\mathbf{P}$ and $\mathbf{Q}$ are beneficial for the classification task in the target domain. According to Eqs. (11) and (12), the problem of finding $\mathbf{P}$ and $\mathbf{Q}$ can be equivalently converted to the problem of finding $\mathbf{\Theta}$, $\mathbf{\Phi}_T$ and $\mathbf{\Phi}_S$. Moreover, as all $\{\mathbf{\Theta}, \mathbf{\Phi}_S, \mathbf{\Phi}_T\}$ give the same objective value of Problem (2), we further remove the CCA term in Problem (7), which leads to the following problem,

$$\begin{aligned} \min_{\mathbf{w},\mathbf{\Theta},\mathbf{\Phi}_T} \; & C\sum_{i=1}^{n_L} \xi_{L,i} + C\sum_{i=1}^{n_U} \xi_{U,i} + \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \; & \mathbf{\Theta}^\top\mathbf{\Theta} = \mathbf{I}. \end{aligned} \quad (13)$$

The new model leverages both source and target data to find a discriminative correlation subspace by exploiting CCA and label information of the target instances. With $\mathbf{\Theta}$ and $\mathbf{\Phi}_T$ observed, we can obtain $\mathbf{P}$ according to Eq. (11), thus can project the target data into the found subspace. Since we only consider the training loss on the target data, we do not need to learn $\mathbf{Q}$ in the above problem.

3.4 Optimization

Now we discuss how to optimize Problem (13). In order to handle the orthogonal constraint on $\mathbf{\Theta}$, we construct a new matrix $\tilde{\mathbf{\Theta}} \in \mathbf{\Omega}$, where the set $\mathbf{\Omega} = \{\tilde{\mathbf{\Theta}} | \tilde{\mathbf{\Theta}}^\top \tilde{\mathbf{\Theta}} = \mathbf{I}\}$. Then

Problem (13) can be reformulated as an equality-constrained problem as

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\Theta},\boldsymbol{\Phi}_T,\tilde{\boldsymbol{\Theta}}} \quad & C\sum_{i=1}^{n_L}\xi_{L,i} + C\sum_{i=1}^{n_U}\xi_{U,i} + \frac{1}{2}||\mathbf{w}||^2, \\ \text{s.t.} \quad & \boldsymbol{\Theta} = \tilde{\boldsymbol{\Theta}}, \ \tilde{\boldsymbol{\Theta}} \in \boldsymbol{\Omega}. \end{aligned} \tag{14}$$

We apply the alternating direction method of multipliers (ADMM) to solve this problem [Boyd *et al.*, 2011]. Specifically, the augmented Lagrangian function of Problem (14) can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \boldsymbol{\Theta}, \boldsymbol{\Phi}_T, \tilde{\boldsymbol{\Theta}}, \boldsymbol{\Lambda}) = & C\sum_{i=1}^{n_L}\xi_{L,i} + C\sum_{i=1}^{n_U}\xi_{U,i} \\ & + \frac{1}{2}||\mathbf{w}||^2 + \frac{\rho}{2}||\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}} + \boldsymbol{\Lambda}||_F^2, \end{aligned} \tag{15}$$

where $\tilde{\boldsymbol{\Theta}} \in \boldsymbol{\Omega}$, $\boldsymbol{\Lambda}$ is the dual variable matrix, ρ is the penalty parameter.

The ADMM algorithm involves three main steps, which are detailed as follows.

- Update $(\mathbf{w}, \boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$ by solving the following problem

$$\begin{aligned} (\mathbf{w}, \boldsymbol{\Theta}, \boldsymbol{\Phi}_T) := \arg \min_{\mathbf{w},\boldsymbol{\Theta},\boldsymbol{\Phi}_T} & C\sum_{i=1}^{n_L}\xi_{L,i} + C\sum_{i=1}^{n_U}\xi_{U,i} \\ & + \frac{1}{2}||\mathbf{w}||^2 + \frac{\rho}{2}||\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}} + \boldsymbol{\Lambda}||_F^2, \end{aligned} \tag{16}$$

- Update $\tilde{\boldsymbol{\Theta}}$ by solving the following problem

$$\tilde{\boldsymbol{\Theta}} := \arg \min_{\tilde{\boldsymbol{\Theta}} \in \boldsymbol{\Omega}} \frac{\rho}{2}||\boldsymbol{\Theta} + \boldsymbol{\Lambda} - \tilde{\boldsymbol{\Theta}}||_F^2, \tag{17}$$

- Update $\boldsymbol{\Lambda}$ using a simple rule

$$\boldsymbol{\Lambda} := \boldsymbol{\Lambda} + \boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}. \tag{18}$$

The proposed **D**iscriminative **C**orrelation **A**nalysis (DCA) algorithm is summarized in Algorithm 1. Since the objective is non-convex and $\boldsymbol{\Omega}$ is also non-convex, Problem (14) is a non-convex optimization problem. However, as stated in Remark 1, as long as $\boldsymbol{\Theta}$ is orthogonal, the matrices $\mathbf{P}$ and $\mathbf{Q}$ obtained by Eqs. (11) and (12) are the solutions to Problem (2) for any $\boldsymbol{\Phi}_T$ and $\boldsymbol{\Phi}_S$. As a result, our method, which is inherited from CCA, is stable to the initialization of $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$. Next, we are ready to describe how to solve Problems (16) and (17) efficiently.

Solution to Subproblem (16)

Subproblem (16) is still a non-convex optimization problem. To solve this, we alternatively update $\mathbf{w}$ and $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$. Detailedly, we firstly fix $\mathbf{w}$ and update $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$ by

$$\begin{aligned} (\boldsymbol{\Theta}, \boldsymbol{\Phi}_T) := \arg \min_{\boldsymbol{\Theta},\boldsymbol{\Phi}_T} & C\sum_{i=1}^{n_L}\xi_{L,i} + C\sum_{i=1}^{n_U}\xi_{U,i} \\ & + \frac{\rho}{2}||\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}} + \boldsymbol{\Lambda}||_F^2. \end{aligned} \tag{19}$$

Algorithm 1 Discriminative Canonical Correlation Analysis

```
Input:  Unlabeled source instances $\mathbf{X}_S = \{\mathbf{x}_{S,i}\}_{i=1}^{n_S}$,
        labeled target instances $\mathbf{X}_L = \{\mathbf{x}_{L,i}, y_{L,i}\}_{i=1}^{n_L}$,
        unlabeled target instances $\mathbf{X}_U = \{\mathbf{x}_{U,i}\}_{i=1}^{n_U}$.
Initialize $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$ such that $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \mathbf{I}$.
Initialize $\mathbf{w}$ by Softmax Regression.
repeat
    Obtain $\{\hat{y}_{U,i}\}_{i=1}^{n_U}$ by Eq. (6),
    Update $(\mathbf{w}, \boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$ by solving Problem (16),
    Update $\tilde{\boldsymbol{\Theta}}$ by solving Problem (17),
    Update $\boldsymbol{\Lambda}$ according to Eq. (18),
until converge.
```

Next, we fix $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$ and update $\mathbf{w}$ by solving the following problem,

$$\mathbf{w} := \arg \min_{\mathbf{w}} C\sum_{i=1}^{n_L}\xi_{L,i} + C\sum_{i=1}^{n_U}\xi_{U,i} + \frac{1}{2}||\mathbf{w}||^2, \tag{20}$$

which is a standard softmax regression problem on the augmented data.

We use the LBFGS algorithm to solve the above two subproblems [Schmidt, 2005]. The warm start strategy is used to solve these subproblems. Specifically, we update the parameters from the results that are obtained by the last iteration of ADMM. Note that the label information of the target instances is exploited to update $(\boldsymbol{\Theta}, \boldsymbol{\Phi}_T)$, which assists to find the discriminative correlation subspace that is beneficial for the target classification task.

Solution to Subproblem (17)

Subproblem (17) seeks to find an orthogonal matrix that is closest to $\boldsymbol{\Theta} + \boldsymbol{\Lambda}$. The solution can be calculated by Singular Value Decomposition. Specifically, let

$$\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \boldsymbol{\Theta} + \boldsymbol{\Lambda} \tag{21}$$

be the SVD of $\boldsymbol{\Theta} + \boldsymbol{\Lambda}$, the solution of Problem (17) is $\mathbf{U}\mathbf{V}^\top$.

4 Experiments

4.1 Data Sets

- **Office Data Set**. The Office data set [Saenko *et al.*, 2010] includes 4106 images with 31 categories from three domains: amazon (A), dslr (D), webcam (W). We use the publicly available SURF [Bay *et al.*, 2006] feature and the DeCAF$_6$ [Donahue *et al.*, 2014] feature in our experiments. The dimensions of SURF and DeCAF$_6$ features are 800 and 4096, respectively.
- **Office-Caltech Data Set**. The Caltech-256 data set [Griffin *et al.*, 2007] includes 256 categories, among which 10 categories are overlapped with the Office data set. The same 800-d SURF feature is publicly available for those 10 categories from the Caltech-256 data set.

We use these 10 overlapping categories between the Office data set and Caltech-256 in our experiments. To extensively evaluate our proposed method, we design experiments by exploiting all possible combinations across different domains and different types of features.

Table 1: Results on the Office data set using 3 labeled target instances

S → T	SVM	Softmax	CCA+Softmax	SHFA	CDLS	DCA
DeCAF$_6$ → SURF						
A → D	56.62 ± 3.39	56.67 ± 2.92	58.11 ± 3.89	56.93 ± 3.79	57.09 ± 3.62	**59.13 ± 3.47**
A → W	57.86 ± 2.76	61.56 ± 3.75	62.33 ± 4.14	60.65 ± 7.11	60.57 ± 4.34	**63.35 ± 5.56**
D → A	42.97 ± 3.20	44.44 ± 2.35	44.05 ± 2.29	42.71 ± 1.75	44.12 ± 1.90	**44.78 ± 1.63**
D → W	57.86 ± 2.76	61.56 ± 3.75	**64.86 ± 2.93**	63.10 ± 2.36	60.82 ± 1.78	64.41 ± 2.70
W → A	42.97 ± 3.20	44.44 ± 2.35	46.76 ± 2.94	46.54 ± 2.49	47.80 ± 2.14	**48.99 ± 2.12**
W → D	56.62 ± 3.39	56.67 ± 2.92	57.29 ± 4.32	**62.43 ± 2.59**	60.00 ± 6.97	59.25 ± 4.79
Average	52.48	54.22	55.57	55.39	55.07	**56.65**
SURF → DeCAF$_6$						
A → D	87.76 ± 4.98	86.68 ± 3.89	85.33 ± 5.12	88.79 ± 5.25	90.84 ± 3.46	**93.46 ± 2.09**
A → W	86.19 ± 2.56	86.41 ± 2.46	86.49 ± 3.48	90.37 ± 2.17	88.08 ± 2.82	**93.06 ± 1.41**
D → A	81.26 ± 2.65	83.85 ± 2.18	83.83 ± 2.27	84.76 ± 1.75	82.38 ± 2.24	**88.57 ± 1.72**
D → W	86.19 ± 2.56	86.41 ± 2.46	86.61 ± 1.83	90.04 ± 2.64	87.10 ± 3.40	**92.41 ± 1.06**
W → A	81.26 ± 2.65	83.85 ± 2.18	83.03 ± 2.93	85.57 ± 2.38	84.07 ± 2.53	**89.27 ± 1.09**
W → D	87.76 ± 4.98	86.68 ± 3.89	87.01 ± 3.80	92.90 ± 2.25	90.09 ± 4.97	**93.27 ± 2.59**
Average	85.07	85.65	85.38	88.74	87.10	**91.67**

In particular, the Office data set contains both SURF and DeCAF$_6$ features for all three domains (*i.e.*, A, D and W), so for each pair of domains, we perform the HDA tasks by using one type of feature for the source domain, and the other one for the target domain, which leads to two groups of tasks (*i.e.*, DeCAF$_6$ → SURF, and SURF → DeCAF$_6$). Each group contains 6 tasks, *i.e.*, A→D, A→W, D→A, D→W, W→A, and W→D.

For the Office-Caltech data set, since the Caltech domain contains only SURF feature, we perform experiments by using Caltech as the source domain, and one of the three domains in the Office data set as the target domain, and vice versa. This also leads to two groups of HDA tasks. The first group takes DeCAF$_6$ as the source feature, and SURF as the target feature, thus includes A → C, D → C and W → C three tasks. Similarly, the second group takes SURF as the source feature, and DeCAF$_6$ as the target feature, thus includes C → A, C → D and C → W three tasks.

Similar to the experimental settings in [Hubert Tsai *et al.*, 2016], we randomly choose 3 target instances per category as the labeled target data for all the target domains, and the rest instances as the test data. For the source domains constructed on A, C and W, we randomly choose 20 source instances per category for training. For the source domains constructed on D, we randomly choose 5 source instances per category since the number of instances in D is much smaller. All experiments are conducted for 10 trials, and the average classification accuracy is reported for comparison.

For our DCA method, we empirically set the subspace dimension as $d_C = 20$, the trade-off parameter as $C = 1$, and the penalty parameter of ADMM as $\rho = 5$. For the baseline method, we follow the previous work [Li *et al.*, 2014] to report their best results on each data set. The parameter sensitivity of our DCA method is provided in Section 4.5.

4.2 Baseline Methods

We evaluate the effectiveness of our proposed method by comparing with the state-of-the-art HDA methods SHFA [Li *et al.*, 2014] and CDLS [Hubert Tsai *et al.*, 2016], as well as a few baselines as follows,

- **SVM**. We perform SVM [Chang and Lin, 2011] on labeled target data as a baseline without domain adaptation.
- **Softmax**. Since we adopt the cross-entropy loss in our proposed model, we also conduct softmax regression on labeled target data. This is also a baseline method without domain adaptation.
- **CCA+Softmax**. CCA+Softmax firstly performs CCA on source and target data, and then conducts softmax regression on the augmented labeled target data.
- **SHFA**. SHFA [Li *et al.*, 2014] learns augmented features for source and target data, and trains an SVM classifier in the semi-supervised fashion simultaneously.
- **CDLS**. CDLS [Hubert Tsai *et al.*, 2016] selects representative cross-domain instances to derive a proper feature subspace for domain adaptation. CDLS is also a semi-supervised algorithm using unlabeled target data.

4.3 Results on Office Data Set

Table 1 presents the average classification accuracy and the standard deviation for all the methods on the Office data set. We have several observations as follows.

- SHFA and CDLS are better than the baseline SVM method, and our newly proposed DCA method is also better than the baseline Softmax method. Considering both SHFA and CDLS are designed based on SVM, and our DCA method adopts the same corss-entropy loss with the Softmax method, this clearly verifies that leveraging heterogeneous source data is beneficial to enhance the performance of the target classifier.
- CCA+Softmax performs comparably or slightly better than the baseline Softmax method. This indicates that the unsupervised method CCA cannot guarantee the good discriminative capacity of the learned subspace, which is critical for learning the target classifier.

Table 2: Results on the Office-Caltech data set using 3 labeled target instances

S → T	SVM	Softmax	CCA+Softmax	SHFA	CDLS	DCA
			$\text{DeCAF}_6 \rightarrow \text{SURF}$			
A → C	29.74 ± 2.15	31.64 ± 2.92	31.46 ± 3.09	30.38 ± 2.09	31.84 ± 3.11	**32.79 ± 3.21**
D → C	29.74 ± 2.15	31.64 ± 2.92	32.02 ± 2.51	31.61 ± 2.46	32.00 ± 2.99	**33.03 ± 2.88**
W → C	29.74 ± 2.15	31.64 ± 2.92	31.74 ± 2.73	30.83 ± 2.54	33.35 ± 2.64	**34.95 ± 3.24**
Average	29.74	31.64	31.74	30.94	32.40	**33.59**
			$\text{SURF} \rightarrow \text{DeCAF}_6$			
C → A	80.37 ± 3.04	84.00 ± 2.27	82.95 ± 2.00	84.30 ± 1.72	82.56 ± 2.17	**89.93 ± 0.70**
C → D	88.69 ± 2.70	89.53 ± 3.76	88.32 ± 4.12	94.21 ± 3.64	92.15 ± 3.83	**95.70 ± 4.00**
C → W	86.20 ± 3.35	86.20 ± 2.37	86.41 ± 2.16	90.04 ± 3.52	89.71 ± 4.01	**90.94 ± 2.05**
Average	85.09	86.58	85.89	89.51	88.14	**92.19**

- In terms of mean accuracy, our proposed DCA algorithm outperforms all other methods, which demonstrates the effectiveness of our proposed method for finding an optimal discriminative correlation subspace, and learning the robust target classifier. Specifically, comparing with the second best method SHFA, our method gets better results on 11 of those 12 tasks, which shows that it is beneficial to ensure the canonical correlation between the source and target domain when learning the target classifier.
- Compared to DeCAF_6, SURF is relatively weak to represent objects, thus learning tasks based on SURF features are more challenging. As a result, the performance improvement of DCA can be marginal with auxiliary features.

4.4 Results on Office-Caltech Data Set

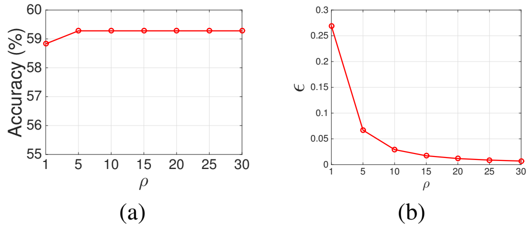
The results of all the methods on the Office-Caltech data set are reported in Table 2. We have similar observations as on the Office data set. Our proposed DCA method achieves the best performance in all of six tasks. This clearly demonstrates the effectiveness of our proposed DCA method for heterogeneous domain adaptation.

4.5 Sensitivity Study

We take the task $\text{A}(\text{DeCAF}_6) \rightarrow \text{D}(\text{SURF})$ as an example to show the parameter sensitivity of our proposed DCA method. In particular, our method mainly involves the penalty parameter ρ that controls the residual of the equality constraint in Eq. (14). We vary $\rho \in \{1, 5, 10, 15, 20, 25, 30\}$, and plot the results in Figure 1. In general, the performance of DCA is not sensitive to ρ. We plot the relative residual $\epsilon = \frac{||\mathbf{\Theta} - \tilde{\mathbf{\Theta}}||_F}{||\mathbf{\Theta}||_F}$ w.r.t. ρ in Figure 1(b). When $\rho = 1$, the relative residual is large, which means that the equality constraint $\mathbf{\Theta} = \tilde{\mathbf{\Theta}}$ is not satisfied well. Nevertheless, when $\rho \geq 5$, ϵ becomes very small, and the classification accuracy is also stable.

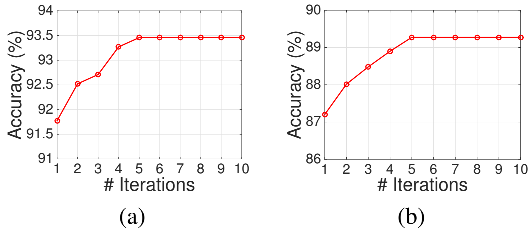
4.6 Performance *w.r.t.* Iterations

We study the performance of DCA *w.r.t.* the iteration number on two example tasks in Figure 2. Specifically, Figure 2(a) shows the results on the task of $\text{A}(\text{SURF}) \rightarrow \text{D}(\text{DeCAF}_6)$, and Figure 2(b) shows the results on the task of $\text{W}(\text{SURF}) \rightarrow \text{A}(\text{DeCAF}_6)$. From the figures, the accuracy of the proposed method improves with increasing iterations, and keeps relatively stable within 10 iterations. In other words, our method converges to a solution with good performance very quickly.



(a)
(b)

Figure 1: Sensitivity study DCA on the A → D task from DeCAF_6 to SURF. (a) Accuracy *w.r.t.* ρ. (b) ϵ *w.r.t.* ρ.



(a)
(b)

Figure 2: Accuracy *w.r.t.* number of iterations of DCA on two example tasks.

5 Conclusions

In this paper, we proposed a new heterogeneous domain adaptation method to learn a discriminative correlation subspace between source and target data. We formulated a unified objective to learn the target classifier, and simultaneously optimize the discriminative correlation subspace. An ADMM algorithm was applied for solving the proposed learning problem. Experiments on two real-world data sets clearly demonstrated the effectiveness of the proposed method.

Acknowledgments

This work is supported by the Guangzhou Key Laboratory of Robotics and Intelligent Software (15180007), NSFC (61502177 and 61602185), HKBU (12306616 and 12302715).

References

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *TIST*, 2(3):27, 2011.

[Chen and Zhang, 2013] Zheng Chen and Weixiong Zhang. Domain adaptation with topical correspondence learning. In *IJCAI*, 2013.

[Chu *et al.*, 2013] Delin Chu, Li-Zhi Liao, Michael K Ng, and Xiaowei Zhang. Sparse canonical correlation analysis: New formulation and algorithm. *T-PAMI*, 35(12):3050–3065, 2013.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

[Duan *et al.*, 2010] L Duan, D Xu, I Tsang, and J Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[Duan *et al.*, 2012a] Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, pages 711–718, 2012.

[Duan *et al.*, 2012b] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *T-PAMI*, 34(9):1667–1680, 2012.

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[Hoffman *et al.*, 2014] Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *IJCV*, 109(1-2):28–41, 2014.

[Hubert Tsai *et al.*, 2016] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. In *CVPR*, pages 5081–5090, 2016.

[Kulis *et al.*, 2011] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pages 1785–1792, 2011.

[Li *et al.*, 2014] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *T-PAMI*, 36(6):1134–1148, 2014.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010.

[Patel *et al.*, 2015] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[Schmidt, 2005] Mark Schmidt. minfunc: unconstrained differentiable multivariate optimization in matlab. *http://www.cs.ubc.ca/ schmidtm/Software/minFunc.html*, 2005.

[Shi *et al.*, 2010] Xiaoxiao Shi, Qi Liu, Wei Fan, S Yu Philip, and Ruixin Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *ICDM*, pages 1049–1054, 2010.

[Tan *et al.*, 2015] Ben Tan, Yangqiu Song, Erheng Zhong, and Qiang Yang. Transitive transfer learning. In *KDD*, pages 1155–1164, 2015.

[Wang and Mahadevan, 2011] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, volume 22, page 1541, 2011.

[Wu *et al.*, 2013] Xinxiao Wu, Han Wang, Cuiwei Liu, and Yunde Jia. Cross-view action recognition over heterogeneous feature spaces. In *ICCV*, pages 609–616, 2013.

[Wu *et al.*, 2014] Qingyao Wu, Michael K. Ng, and Yunming Ye. Cotransfer learning using coupled markov chains with restart. *IEEE Intelligent Systems*, 29(4):26 – 33, 2014.

[Xiao and Guo, 2015] Min Xiao and Yuhong Guo. Feature space independent semi-supervised domain adaptation via kernel matching. *T-PAMI*, 37(1):54–66, 2015.

[Yang *et al.*, 2015] Liu Yang, Liping Jing, Jian Yu, and Michael K Ng. Learning transferred weights from co-occurrence data for heterogeneous transfer learning. *T-NNLS*, 2015.

[Zhou *et al.*, 2014a] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, pages 2213–2220, 2014.

[Zhou *et al.*, 2014b] Joey Tianyi Zhou, Ivor W Tsang, Sinno Jialin Pan, and Mingkui Tan. Heterogeneous domain adaptation for multiple classes. In *AISTATS*, pages 1095–1103, 2014.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Shen-Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. In *AAAI*, pages 2400–2406, 2016.

[Zhu *et al.*, 2011] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.