Locality-Constrained Transfer Coding for Heterogeneous Domain Adaptation

Jingjing Li[1,2], Ke Lu[1], Lei Zhu[2], and Zhihui Li[3]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China.
[2]School of Information Technology and Electrical Engineering, University of Queensland, QLD 4067, Australia.
[3]Beijing Etrol Technologies Co., Ltd, Beijing, China.
lijin117@yeah.net, kel@uestc.edu.cn, leizhu0608@gmail.com, zhihuilics@gmail.com

Abstract. Currently, most of widely used databases are label-wise. In other words, people organize their data with corresponding labels, e.g., class information, keywords and description, for the convenience of indexing and retrieving. However, labels of the data from a novel application usually are not available, and labeling by hand is very expensive. To address this, we propose a novel approach based on transfer learning. Specifically, we aim at tackling heterogeneous domain adaptation (HDA). HDA is a crucial topic in transfer learning. Two inevitable issues, feature discrepancy and distribution divergence, get in the way of HDA. However, due to the significant challenges of HDA, previous work commonly focus on handling one of them and neglect the other. Here we propose to deploy locality-constrained transfer coding (LCTC) to simultaneously alleviate the feature discrepancy and mitigate the distribution divergence. Our method is powered by two tactics: feature alignment and distribution alignment. The former learns new transferable feature representations by sharing-dictionary coding and the latter aligns the distribution gaps on the new feature space. By formulating the problem into a unified objective and optimizing it via an iterative fashion, the two tactics are reinforced by each other and the two domains are drawn closer under the new representations. Extensive experiments on image classification and text categorization verify the superiority of our method against several state-of-the-art approaches.

Keywords: Domain adaptation, transfer learning, knowledge discovery

1 Introduction

From the perspective of general users, a database is a set of well organized data. For a piece of data in a database, it is usually organized with several keywords associated with it for the convenience of indexing and retrieving. These keywords can be seen as labels of the corresponding data. In real-world applications, however, labels are not always accessible. What should one do in this situation? Most people would say we can train a classifier to automaticly label samples, and some

may say why bother, we can label them by hand. Unfortunately, both of them are not practical in specific circumstances, because either we have insufficient training samples to train an accurate classifier for a new application, or labeling by hand is too expensive to be afforded. The endless stream of novel applications and pervasive scarcity of well-labeled data have stimulated a remarkable wave of research on transfer learning [1]. Transfer learning handles the problem where available labeled samples in the target domain are too scarce to train an accurate model. It borrows knowledge from related domains, i.e, the source domain, to facilitate training. However, a majority of machine learning approaches might fail to transfer knowledge because they commonly assume that training and test data are drawn from the same probability distribution. Unfortunately, it is always hard to find a semantically matched source domain which happens to share the same data distribution with the target domain.

Domain adaptation [2, 3, 5, 6] is proposed with the mission of knowledge transfer beyond the distribution gaps among different domains. Most of existing work [2, 7, 8] focus on the homogeneous domain adaptation problem where the two domains are sampled from different data distributions but the same feature representation. In practice, however, the two domains are often drawn from not only different probability distributions but also different feature representations. For instance, the target domain is sampled from convolutional neural network (CNN) features of images, whilst the source domain is drawn from bag-of-words (BoW) features of texts. Accordingly, HDA [9–11] is proposed and has received increasing attention.

To transfer knowledge among heterogeneous domains, HDA has to overcome two inevitable obstacles: feature discrepancy and distribution divergence. However, due to the intrinsic challenges of HDA, existing work generally select only one of them to tackle. They optimize either the feature discrepancy or the distribution divergence individually. For instance, previous work [9, 12, 8, 10] focus on minimizing the distribution divergence. A fraction of existing approaches [13, 11, 14] propose to learn new feature representations to mitigate the feature discrepancy. Nevertheless, it is easy to know that the distribution gaps would be smaller if the two domains are represented by more similar feature representations. Meanwhile, minimizing the distribution gaps also leads to more optimal feature representations (feature representations which can alleviate the distribution divergence are preferred). In a nutshell, the two objectives can reinforce each other and jointly optimizing them would be more beneficial.

Motivated by above discussions, we propose a novel approach, named as locality-constrained transfer coding (LCTC). It simultaneously alleviates the feature discrepancy and mitigates the distribution divergence in a unified optimization problem. Specifically, our objective is formulated by the following motivations: 1) To alleviate the feature discrepancy, we learn new transferable feature representations for the two domains with a shared codebook. By sharing the codebook, the learned new features are interconnected and, thus, the knowledge can be transfered to the target domain. 2) To mitigate the distribution divergence, we further minimize the marginal distribution gaps between the two

domains on the learned new feature space. 3) It is worth noting that samples with the same semantic label tend to stay close (as shown in Fig. 1). Therefore, we also exploit to preserve the manifold structure and local consistency in our formulation. An iterative updating algorithm is presented to optimize the objective. Experimental results on image classification and text categorization verify the superiority of our method.

2 Related Work

Many approaches [19, 20, 7] have been proposed to handle domain adaptation problems. A majority of them focus on homogeneous domain adaptation. For instance, Pan et al. [8] propose to learn transferable components across domains in a reproducing kernel Hilbert space (RKHS) by using maximum mean discrepancy (MMD) [15]. Ding et al. [5] deploy low-rank coding in a deep structure to learn a latent space shared by the two domains. However, those work concentrate on either minimizing the data distribution gaps or mining the shared factors among the two domains. They did not give much attention to the feature discrepancy.

Recently, some HDA methods [21, 22] are proposed to handle the feature discrepancy problem. Wang and Mahadevan [9] align the data manifold of the two domains for adaptation. Li et al. [11] propose using augmented feature representations to effectively utilize the data from both domains with a SVM similar formulation. Thai et al. [10] re-weight the samples and select landmarks for classification. .

As stated before, the current issue is that there are two inevitable problems of which need to be taken care, but previous work exploit to optimize either the feature discrepancy or the distribution divergence separately. They catch one of them and lose another. This paper proposes a novel approach which aims to simultaneously alleviate the feature discrepancy and mitigate the distribution divergence in a unified optimization problem.

3 Locality-Constrained Transfer Coding

3.1 Notations

In this paper, we use bold lowercase letters to represent vectors, bold uppercase letters to represent matrices. A sample is denoted as a vector, e.g., $\mathbf{x}$, and the i-th sample in a set is represented by the symbol $\mathbf{x}_i$. For a matrix $\mathbf{M}$, its Frobenius norm is defined as $\|\mathbf{M}\|_F = \sqrt{\sum_i \delta_i(\mathbf{M})^2}$, where $\delta_i(\mathbf{M})$ is the i-th singular value of the matrix $\mathbf{M}$. The trace of matrix $\mathbf{M}$ is denoted by $\mathbf{tr}(\mathbf{M})$. For clarity, the frequently used notations and corresponding descriptions are shown in Table 1.

3.2 Definitions

Definition 1 *A domain $\mathbb{D}$ consists of two parts: feature space $\mathcal{X}$ and its probability distribution $P(\mathbf{X})$, where $\mathbf{X} \in \mathcal{X}$.*

Table 1: Frequently used notations and their descriptions.

Notation	Description	Notation	Description
$\mathbf{X_s} \in \mathbb{R}^{d_s * n_s}$, $\mathbf{y_s}$	source samples/labels	$\mathbf{X} \in \mathbb{R}^{m*n}$	$\mathbf{X}_s$ and $\mathbf{X}_t$
$\mathbf{X_t} \in \mathbb{R}^{d_t * n_t}$, $\mathbf{y_t}$	target samples/labels	$\mathbf{B} \in \mathbb{R}^{m*k}$	codebook
$\mathbf{M} \in \mathbb{R}^{n*n}$	MMD matrix	$\mathbf{S} \in \mathbb{R}^{k*n}$	coding matrix
$\mathbf{\Lambda}$	Lagrange multipliers	α, β	penalty parameters

We use subscripts s and t to indicate the source domain and the target domain, respectively. This paper focuses on the following problem:

Problem 1 *Given a well-labeled source domain $\mathbb{D}_s$ and a mostly unlabeled target domain $\mathbb{D}_t$, where $\mathbb{D}_s \neq \mathbb{D}_t$, $\mathcal{X}_s \neq \mathcal{X}_t$ and $P(\mathbf{X}_s) \neq P(\mathbf{X}_t)$. Simultaneously align the feature discrepancy and distribution divergence between $\mathbb{D}_s$ and $\mathbb{D}_t$.*

3.3 Problem Formulation

As we stated in the introduction, our model supposes to learn new transferable feature representations for $\mathbf{X}_s$ and $\mathbf{X}_t$ with a shared codebook $\mathbf{B}$. It also minimizes the distribution gaps between the two domains on the new feature space. Thus, our objective can be formulated as follows:

$$\min_{\mathbf{B},\mathbf{S}} \underbrace{\mathcal{C}_1(\mathbf{X_s}, \mathbf{X_t}, \mathbf{B}, \mathbf{S})}_{feature\ alignment} + \underbrace{\alpha \mathcal{C}_2(\mathbf{S_s}, \mathbf{S_t})}_{distribution\ alignment} + \underbrace{\beta \Omega(\mathbf{S})}_{constraint} \tag{1}$$

where $\mathbf{S}$ ($\mathbf{S}_s$ and $\mathbf{S}_t$ for $\mathbf{X}_s$ and $\mathbf{X}_t$ respectively) is the new feature representation, $\mathcal{C}_1$ is the feature alignment part, $\mathcal{C}_2$ is the distribution alignment part and Ω is the constraint. Notice that $\mathcal{C}_2$ is deployed on $\mathbf{S}$ rather than $\mathbf{X}$. It is a progressive alignment based on the results of $\mathcal{C}_1$. $\alpha > 0$ and $\beta > 0$ are two hyper-parameters. In the remainder of this section, we will present each part in detail and show how to optimize Eq. (1).

Feature Alignment. Suppose that we can learn a new feature representation $\mathbf{S}$ by which the feature discrepancy between the two domains can be alleviated. For the source domain data $\mathbf{X}_s$, we can learn $\mathbf{S}_s$ through

$$\begin{gathered} \min_{\mathbf{b}_s, \mathbf{s}_s} \sum_{i=1}^{n_s} \left\| \mathbf{x}_{s,i} - \sum_{j=1}^{k} \mathbf{b}_{s,j} \mathbf{s}_{s,i}^{j} \right\|_2^2 + \beta \sum_{i=1}^{n_s} \left\| \mathbf{s}_{s,i} \right\|_1, \\ s.t. \|\mathbf{b}_{s,j}\|^2 \leq c, \quad \forall j, \end{gathered} \tag{2}$$

where c is a constant, we keep it as 1 in this paper. Eq. (2) can be rewritten as the following form in matrix,

$$\min_{\mathbf{B_s}, \mathbf{S_s}} \|\mathbf{X_s} - \mathbf{B_s}\mathbf{S_s}\|_F^2 + \beta \sum_{i=1}^{n_s} \|\mathbf{s}_{s,i}\|_1, \ s.t. \ \|\mathbf{b}_{s,j}\|^2 \leq c, \ \forall j, \tag{3}$$

where $\mathbf{B_s}$ is the codebook, also known as dictionary matrix, learned on $\mathbf{X}_s$, and $\mathbf{S}_s$ is the coding matrix. $\mathbf{s}_{s,i}$ is a sparse representation for the corresponding data

point in $\mathbf{X}_s$. In a similar fashion, we can learn a new feature representation for the target domain data by optimizing:

$$\min_{\mathbf{B_t},\mathbf{S_t}} \|\mathbf{X_t}-\mathbf{B_t}\mathbf{S_t}\|_F^2+\beta\sum_{i=1}^{n_t}\|\mathbf{s}_{t,i}\|_1,\ s.t.\ \|\mathbf{b}_{t,j}\|^2\leq c,\ \forall j. \tag{4}$$

In order to transfer knowledge from the source domain to the target domain, we advocate $\mathbf{X_s}$ and $\mathbf{X_t}$ sharing the same codebook $\mathbf{B}$. Thus, their corresponding new feature representations $\mathbf{S_s}$ and $\mathbf{S_t}$ would be interconnected and can be directly compared. It means the new feature representations are transferable. To this end, we have the following problem:

$$\begin{gathered}\min_{\mathbf{B},\mathbf{S_s},\mathbf{S_t}} \|\mathbf{X_s}-\mathbf{B}\mathbf{S_s}\|_F^2+\|\mathbf{X_t}-\mathbf{B}\mathbf{S_t}\|_F^2+\beta\sum_{i=1}^{n}\|\mathbf{s}_i\|_1,\\ s.t.\|\mathbf{b}_j\|^2\leq c,\quad\forall j.\end{gathered} \tag{5}$$

Note that $\mathbf{S_s}$ and $\mathbf{S_t}$ may have different dimensionalities in practice. In this paper, to reduce the computational costs and filter out high-dimensional noises, samples are aligned to the same dimensionality by PCA. One can also align them in a RKHS, or by learning two projections, one for each domain. Eq. (5) can be rewritten as the following equivalent equation after some algebraic manipulations,

$$\begin{gathered}\min_{\mathbf{B},\mathbf{S}} \|\mathbf{X}-\mathbf{B}\mathbf{S}\|_F^2+\beta\sum_{i=1}^{n}\|\mathbf{s}_i\|_1,\\ s.t.\|\mathbf{b}_j\|^2\leq c,\quad\forall j,\end{gathered} \tag{6}$$

where $\mathbf{X}$ and $\mathbf{S}$ are defined as,

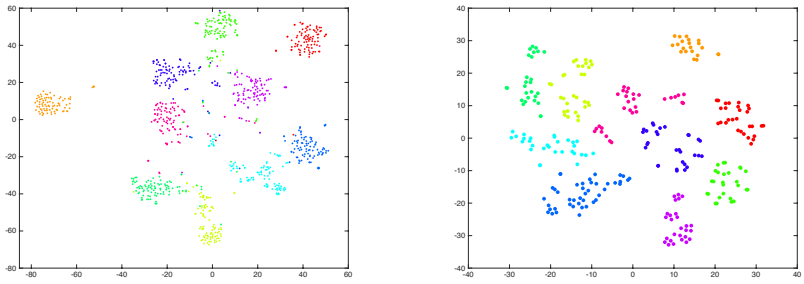
$$\mathbf{X}=[\mathbf{X}_s\ \mathbf{X}_t],\ \mathbf{S}=[\mathbf{S}_s\ \mathbf{S}_t].$$

Distribution Alignment. Please notice that in HDA tasks, one needs to take care of not only feature discrepancy but also distribution divergence. Recently, MMD [15] has been introduced to estimate the distance between distributions because of its non-parametric merit. The MMD between two datasets $\mathbf{X}_s$ and $\mathbf{X}_t$ can be computed as:

$$\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\phi(\mathbf{x}_{s,i})-\frac{1}{n_t}\sum_{j=1}^{n_t}\phi(\mathbf{x}_{t,j})\right\|^2, \tag{7}$$

where $\phi(\cdot)$ is a feature mapping. Since we have learned the new feature representations $\mathbf{S_s}$ and $\mathbf{S_t}$ for $\mathbf{X_s}$ and $\mathbf{X_t}$ respectively, we further mitigate the distribution divergence on the new feature space. The minimization of distribution divergence between $\mathbf{S}_s$ and $\mathbf{S}_t$ can be formulated as follows:

$$\min_{\mathbf{S_s},\mathbf{S_t}}\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{S}_{s,i}-\frac{1}{n_t}\sum_{j=1}^{n_t}\mathbf{S}_{t,j}\right\|_2^2=\min_{\mathbf{S}}\mathbf{tr}(\mathbf{S}\mathbf{M}\mathbf{S}^\top), \tag{8}$$



(a) Amazon (b) Webcam

Fig. 1: Visualization of the Amazon and Webcam dataset from Office dataset [4]. The figure is generated by t-SNE [23] with DeCAF_6 features [24]. Each color denotes one class.

where $\mathbf{M}$ is the MMD matrix computed as:

$$\mathbf{M}_{ij} = \begin{cases} \frac{1}{n_s n_s} & , \; if \; \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_s \\ \frac{1}{n_t n_t} & , \; if \; \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_t \\ \frac{-1}{n_s n_t} & , \; otherwise \end{cases} . \tag{9}$$

Local Consistency. Suppose that the label information of the target domain is known, we visualize the target samples as shown in Fig. 1. An interesting observation is that the data samples with the same label stay in a compact cluster. More formally, a sample tends to have the same label with its k-nearest neighbors. Traditionally, the local consistency is formulated as a graph Laplacian regularization [25, 26]. However, to formulate a compact and efficient objective, we advocate a locality constraint instead of the graph regularization in this paper. The locality constraint can be seamlessly incorporated into our sharing-dictionary coding framework. And, fortunately, it gives rise to a serendipity that the locality constraint must lead to sparsity [32]. Thus, we can remove the ℓ_1 norm from Eq. (6). The locality constraint is defined as follows:

$$\| \mathbf{d}_i \odot \mathbf{s}_i \|^2, with \;\; \mathbf{d}_i = \exp(\tfrac{\text{dist}(\mathbf{x}_i, \mathbf{B})}{\sigma}) \tag{10}$$

where $\odot$ denotes the element-wise multiplication, $\mathbf{d}_i$ is the locality adaptor which measures the distance between an instance $\mathbf{x}_i$ and the codebook. $\text{dist}(\mathbf{x}_i, \mathbf{b}_j)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{b}_j$. σ is used for adjusting the weight decay speed [32]. We set $\sigma = 1$ in this paper.

At last, by taking all the considerations into account, we have the overall objective shown as follows:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{S}} \|\mathbf{X}-\mathbf{BS}\|_F^2+\alpha\mathbf{tr}(\mathbf{SMS}^\top)+\beta\sum_{i=1}^{n}\|\mathbf{d}_i\odot\mathbf{s}_i\|^2, \\ s.t. \|\mathbf{b}_j\|^2 \leq c, \quad \forall j. \end{aligned} \tag{11}$$

3.4 Problem Optimization

It is easy to know that Eq. (11) is not convex for $\mathbf{B}$ and $\mathbf{S}$ simultaneously. However, it is convex for each of them when the other is fixed. We, therefore, deploy an alternative strategy to optimize it as shown in the following steps.

Step 1. Optimize coding matrix S. Optimizing Eq. (11) with respect to $\mathbf{S}$ when $\mathbf{B}$ is fixed can be reformulated as optimizing the following problem:

$$\min_{\mathbf{S}} \|\mathbf{X}-\mathbf{BS}\|_F^2+\alpha\mathbf{tr}(\mathbf{SMS}^\top)+\beta\sum_{i=1}^{n}\|\mathbf{d}_i\odot\mathbf{s}_i\|^2. \tag{12}$$

To make the problem easier to be optimized, we introduce an auxiliary variable $\mathbf{D}_i = \mathrm{diag}(\mathbf{d}_i)$. As a result, Eq. (12) can be rewritten as,

$$\min_{\mathbf{S}} \|\mathbf{X}-\mathbf{BS}\|_F^2+\alpha\mathbf{tr}(\mathbf{SMS}^\top)+\beta\sum_{i=1}^{n}\|\mathbf{D}_i\cdot\mathbf{s}_i\|^2. \tag{13}$$

To solve $\mathbf{S}$, by taking the derivative of Eq. (13) with respect to $\mathbf{S}$, and setting the derivative to zero, we have,

$$\mathbf{s}_i = \left(\mathbf{B}^\top\mathbf{B} + \alpha\mathbf{M}_{ii}\mathbf{I} + \beta\mathbf{D}_i^\top\mathbf{D}_i\right)^{-1}\mathbf{B}^\top\mathbf{x}_i. \tag{14}$$

Step 2. Optimize codebook B. Optimizing Eq. (11) with respect to $\mathbf{B}$ when $\mathbf{S}$ is fixed can be reformulated as optimizing the following problem:

$$\min_{\mathbf{B}} \|\mathbf{X}-\mathbf{BS}\|_F^2+\beta\sum_{i=1}^{n}\|\mathbf{d}_i\odot\mathbf{s}_i\|^2, \quad s.t.\ \|\mathbf{b}_j\|^2 \leq c, \quad \forall j. \tag{15}$$

The problem has been well investigated in previous work [32]. Limited by spaces, we do not present the details here. For clarity, we sketch out the main steps of our approach in Algorithm 1.

3.5 Complexity Analysis

Here we present the theoretical complexity of Algorithm 1 by big O notation. For the initialization part, the KMeans operation costs $O(n)$ and the construction of MMD matrix costs $O(n^2)$. For the iteratively updating part, solving the coding matrix $\mathbf{S}$ is some matrix multiplications, it generally costs $O(m^3)$, optimizing the codebook $\mathbf{B}$ costs $O(m+k^2)$. In sum, the overall time costs of Algorithm 1 is $O(n+n^2+T(m^3+m+k^2))$, where T is the number of iterations.

Algorithm 1. *HDA via locality-constrained transfer coding*

Input: Sample sets $\mathbf{X}_s$ and $\mathbf{X}_t$, parameters α, β, σ and c.
Output: Label information of $\mathbf{X}_t$.

Initialize
1. Initialize $\mathbf{d}_i := \mathbf{0}$, $\mathbf{s}_i := \mathbf{0}$, σ=1 and c=1.
2. Use KMeans clustering [32] to initialize $\mathbf{B}$.
3. Compute MMD matrix $\mathbf{M}$ by Eq. (9).
Repeat
4. Optimize the coding matrix $\mathbf{S}$.
5. Optimize the codebook $\mathbf{B}$.
6. Update the locality adaptor $\mathbf{d}$.
until *Convergence or max iteration*
7. Classify $\mathbf{S}_t$ with $\mathbf{S}_s$ used as reference.

4 Experiments

4.1 Data Preparation

Office + Caltech-256 dataset [4] consists of 4 sub-dataset from **Amazon** (A), **Webcam** (W), **DSLR** (D), and **Caltech-256** (C). Samples in Amazon are downloaded from amazon.com. Webcam consists of low-resolution images captured by a web camera. On the contrary, images in DSLR are high-resolution ones captured by a digital SLR camera. In our experiments, we follow the same settings with previous work [2]. Ten common classes shared by four dataset are selected. There are 8 to 151 samples per category per domain, and 2,533 images in total. Furthermore, 800-dimensional SURF features and 4,096-dimensional DeCAF_6 features [24] are extracted as our low-level input.

Multilingual Reuters Collection [33] is a cross-lingual text dataset. It consists of 11,000 articles from 6 categories in 5 languages, i.e., English, French, German, Italian, and Spanish. Here we follow the same settings in previous work [11, 10]. Specifically, all the articles are represented by BoW with TF-IDF. Then, the BoW features are processed by PCA with dimensionality of 1,131, 1,230, 1,417, 1,041, and 807 for different language categories English, French, German, Italian, and Spanish, respectively.

4.2 Experimental Protocols

To fully evaluate our model, we perform three experiments. For instance, image classification across features, image classification across features and datasets and heterogeneous text categorization. Office + Caltech-256 are used in the first two experiments and Multilingual Reuters Collection are used in the last.

For fair comparison, we follow the same settings with previous work. Specifically, for **image classification** tasks, the source domain consists of 20 samples per category for training, and 3 labeled target samples per category are randomly selected as reference for classification. For **text categorization** tasks, we have Spanish as the target domain and the others as the source domain. We

Table 2: Accuracy (%) of HDA across features.

Method	A→A	C→C	W→W	Avg.
SVM_t	44.2 ± 1.1	30.1 ± 0.9	58.3 ± 1.2	44.2 ± 1.1
DAMA	39.5 ± 0.7	19.5 ± 0.8	47.5 ± 1.6	35.5 ± 1.1
MMDT	40.7 ± 1.0	31.5 ± 1.1	60.3 ± 0.8	44.2 ± 1.0
SHFA	43.4 ± 0.9	29.8 ± 1.3	62.4 ± 0.9	45.2 ± 1.0
SHFR	44.5 ± 1.1	33.4 ± 1.0	54.3 ± 0.9	44.1 ± 1.0
LCTC	$\mathbf{45.5 \pm 1.4}$	$\mathbf{34.5 \pm 1.2}$	$\mathbf{64.3 \pm 1.1}$	$\mathbf{48.1 \pm 1.2}$

randomly select 100 articles per category for the source domain and 500 articles for the target domain. Then, we select 10 labeled target samples as reference. For simplicity and without loss of generality, we set the hyper-parameters $\alpha = 1$ and $\beta = 1$ in our experiments.

Table 3: Accuracy (%) of HDA across features and datasets. The source domain and the target domain are represented by $DeCAF_6$ features and SURF features respectively.

Method	SVM_t	MMDT	SHFA	SHFR	LCTC
A→C	30.1 ± 0.9	28.5 ± 1.4	29.6 ± 1.5	27.4 ± 1.7	$\mathbf{34.7 \pm 1.5}$
A→W	58.3 ± 1.2	58.1 ± 1.3	58.8 ± 1.3	52.5 ± 1.4	$\mathbf{60.4 \pm 1.4}$
C→A	44.2 ± 1.1	44.7 ± 1.0	45.9 ± 1.3	43.6 ± 1.3	$\mathbf{50.1 \pm 1.2}$
C→W	58.3 ± 1.2	57.8 ± 1.1	59.1 ± 1.2	52.7 ± 1.5	$\mathbf{61.9 \pm 2.0}$
W→A	44.2 ± 1.1	45.3 ± 1.5	46.6 ± 1.3	43.1 ± 1.6	$\mathbf{52.3 \pm 1.4}$
W→C	30.1 ± 0.9	29.9 ± 1.3	29.7 ± 1.6	27.1 ± 1.3	$\mathbf{32.9 \pm 1.6}$
Avg.	44.2 ± 1.1	44.1 ± 1.3	45.0 ± 1.4	41.1 ± 1.5	$\mathbf{48.7 \pm 1.5}$

We compare our method with several state-of-the-art HDA approaches, e.g., domain adaptation using manifold alignment (DAMA) [9], maximum margin domain transform (MMDT) [12], semi-supervised heterogeneous feature augmentation (SHFA) [11], sparse heterogeneous feature representation (SHFR) [14] and cross-domain landmark selection (CDLS) [10]. SVM trained on the labeled reference samples (SVM_t) is used as baseline. SVM is also used as the final classifier for the tested approaches. We report the accuracy rate [2, 6] on the target domain, i.e., the ratio between the number of correctly predicted samples and the number of total samples in the target domain. Since the evaluated instances are randomly selected, each of the reported results of our algorithm is the average of 10 runs.

4.3 Experimental Results and Discussions

The image classification results of **HDA across features** on Office+Caltech-256 are shown in Table 2. The two domains are sampled from different feature

Table 4: Accuracy (%) of HDA on text categorization.

Method	English	French	German	Italian
SVM_t	67.1 ± 0.8			
DAMA	67.8 ± 0.7	68.3 ± 0.8	67.7 ± 1.0	66.5 ± 1.1
MMDT	68.9 ± 0.6	69.1 ± 0.7	68.3 ± 0.6	67.5 ± 0.5
SHFA	68.2 ± 0.9	68.7 ± 0.4	68.9 ± 0.5	68.5 ± 0.7
SHFR	67.7 ± 0.4	68.5 ± 0.7	68.1 ± 0.8	67.2 ± 1.0
CDLS	71.1 ± 0.7	71.2 ± 0.9	70.9 ± 0.7	$\mathbf{71.5 \pm 0.6}$
LCTC	$\mathbf{73.7 \pm 0.5}$	$\mathbf{74.0 \pm 0.6}$	$\mathbf{72.5 \pm 0.4}$	71.3 ± 0.5

representations but from the same dataset. DSLR is not tested for the limited number of samples. It can be seen from Table 2 that HDA approaches generally perform better than baseline SVM. However, DAMA not always can outperform the baseline. A possible explanation is that DAMA only considers the manifold matching and topology structure preservation between two domains. The further knowledge transfer after domain alignment is ignored in DAMA. Our approach considers not only the topology structure (by locality constraint) and the distribution alignment (by minimizing MMD), but also the knowledge transfer by a shared codebook. As a result, our approach performs better than state-of-the-art methods.

Table 3 shows the results of **HDA across features and datasets** on Office+Caltech-256. The source domain and the target domain are drawn not only from different feature representations but also different datasets. We can see that our model stays ahead of the evaluations. It further verifies the effectiveness of our model. It is worth noting that DAMA and MMDT are approaches that emphasize on distribution matching, whilst SHFA and SHFR mainly consider learning new feature representations and new classifiers. Both strategies are important and effective in some ways. However, the two obstacles of HDA are inevitable when the domain difference is substantially large. Thus, jointly optimizing both of them, as our approach does, can further improve the performance.

From Table 2 and Table 3, it is clear that our approach performs well on image classification tasks. Now, we further test it on text categorization tasks. Specifically, we evaluate it on Multilingual Reuters Collection dataset. Following the previous work [11, 10], we use 'Spanish' as the target domain, 'English', 'French', 'German' and 'Italian' as the source domain respectively. The experimental results are reported in Table 4.

It can be seen from Table 4 that our algorithm also performs well on **text categorization** tasks. Limited by space, we only report the results of 10 labeled samples as reference. It is worth noting that although HDA methods outperform the baseline, the performance superiority between HDA methods and SVM would get smaller with the increasing number of labeled target samples. It means transfer learning is especially suitable for tasks where the target domain has just a few or even no labeled data. Besides, CDLS proposes to minimize the distribution gaps and re-weight samples for better adaptation. However, it does

not preserve the local structure of data samples. That is the reason that our method outperforms CDLS.

5 Conclusion

This paper proposes a novel approach for HDA, which takes both alleviating the feature discrepancy and mitigating the distribution divergence into consideration. By sharing a dictionary, the source domain and the target domain are coded in shared new feature representations. The probability distributions of the two domains are further aligned on the new feature space. A locality constraint is deployed to preserve the local structure and to reduce the computational costs. Extensive experiments on image classification and text categorization tasks verify the superiority of our approach.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61371183, ARC under Grant FT130101530 and DP170103954, the Applied Basic Research Program of Sichuan Province under Grant 2015JY0124, and the China Scholarship Council.

References

1. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE TKDE*, vol. 22, no. 10, pp. 1345–1359, 2010.
2. B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*. IEEE, 2012, pp. 2066–2073.
3. S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE TKDE*, vol. 22, no. 7, pp. 929–942, 2010.
4. K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *ECCV*, 2010, pp. 213–226.
5. Z. Ding, M. Shao, and Y. Fu, "Deep low-rank coding for transfer learning," in *AAAI*. AAAI Press, 2015, pp. 3453–3459.
6. J. Li, J. Zhao, and K. Lu, "Joint feature selection and structure preservation for domain adaptation," in *IJCAI*, 2016, pp. 1697–1703.
7. B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013, pp. 2960–2967.
8. S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE TNN*, vol. 22, no. 2, pp. 199–210, 2011.
9. C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *IJCAI*, vol. 22, no. 1, 2011, p. 1541.
10. Y.-H. Hubert Tsai, Y.-R. Yeh, and Y.-C. Frank Wang, "Learning cross-domain landmarks for heterogeneous domain adaptation," in *CVPR*, 2016, pp. 5081–5090.
11. W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE TPAMI*, vol. 36, no. 6, pp. 1134–1148, 2014.
12. J. Hoffman, E. Rodner, J. Donahue, B. Kulis, and K. Saenko, "Asymmetric and category invariant feature transformations for domain adaptation," *IJCV*, vol. 109, no. 1-2, pp. 28–41, 2014.

13. R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *ICML*, 2007, pp. 759–766.
14. J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, "Heterogeneous domain adaptation for multiple classes." in *AISTATS*, 2014, pp. 1095–1103.
15. A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *NIPS*, 2006, pp. 513–520.
16. M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *CVPR*. IEEE, 2014, pp. 1410–1417.
17. C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
18. M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, "Transfer learning with graph co-regularization," *IEEE TKDE*, vol. 26, no. 7, pp. 1805–1818, 2014.
19. L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *IEEE TPAMI*, vol. 32, no. 5, pp. 770–787, 2010.
20. R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *ICCV*. IEEE, 2011, pp. 999–1006.
21. Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification." in *AAAI*, 2011.
22. X. Shi, Q. Liu, W. Fan, S. Y. Philip, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *ICDM*, 2010, pp. 1049–1054.
23. L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, pp. 2579–2605, 2008.
24. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.
25. D. Cai, X. He, and J. Han, "Spectral regression: A unified approach for sparse subspace learning," in *ICDM*. IEEE, 2007, pp. 73–82.
26. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE TPAMI*, vol. 29, no. 1, pp. 40–51, 2007.
27. H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *NIPS*, vol. 19, p. 801, 2007.
28. Y. Censor and S. A. Zenios, *Parallel optimization: Theory, algorithms, and applications*. Oxford University Press on Demand, 1997.
29. M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu, "Transfer sparse coding for robust image representation," in *CVPR*, 2013, pp. 407–414.
30. M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE TIP*, vol. 20, no. 5, pp. 1327–1336, 2011.
31. Y.-H. H. Tsai, C.-A. Hou, W.-Y. Chen, Y.-R. Yeh, and Y.-C. F. Wang, "Domain-constraint transfer coding for imbalanced unsupervised domain adaptation," in *AAAI*. AAAI Press, 2016, pp. 3597–3603.
32. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*. IEEE, 2010, pp. 3360–3367.
33. M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views-an application to multilingual text categorization," in *NIPS*, 2009, pp. 28–36.