Deep Low-Rank Coding for Transfer Learning*

Zhengming Ding[1], **Ming Shao**[1] and **Yun Fu**[1,2]
Department of Electrical & Computer Engineering[1],
College of Computer & Information Science[2],
Northeastern University, Boston, MA, USA
{allanding,mingshao,yunfu}@ece.neu.edu

Abstract

Recent researches on transfer learning exploit deep structures for discriminative feature representation to tackle cross-domain disparity. However, few of them are able to joint feature learning and knowledge transfer in a unified deep framework. In this paper, we develop a novel approach, called Deep Low-Rank Coding (DLRC), for transfer learning. Specifically, discriminative low-rank coding is achieved in the guidance of an iterative supervised structure term for each single layer. In this way, both marginal and conditional distributions between two domains intend to be mitigated. In addition, a marginalized denoising feature transformation is employed to guarantee the learned single-layer low-rank coding to be robust despite of corruptions or noises. Finally, by stacking multiple layers of low-rank codings, we manage to learn robust cross-domain features from coarse to fine. Experimental results on several benchmarks have demonstrated the effectiveness of our proposed algorithm on facilitating the recognition performance for the target domain.

1 Introduction

In machine learning and pattern recognition fields, there is always a situation that we have plenty of unlabeled data while no or insufficient labeled data for training in the target domain. Transfer learning [Pan and Yang, 2010] has been demonstrated as a promising technique to address such difficulty by borrowing knowledge from other well-learned source domains, which might lie in different distributions with the target one. Many recent researches on transfer learning have witnessed appealing performance by seeking a common feature space where knowledge from source can be transferred to assist the recognition task of target domain [Chen *et al.*, 2012; Ding *et al.*, 2014; Shao *et al.*, 2012; Shekhar *et al.*, 2013; Long *et al.*, 2014b]. Therefore, it is the key to uncover the rich and discriminative information across source and target domains in transfer learning.

Recently, low-rank constraint [Liu *et al.*, 2013] has been widely studied in conventional transfer learning due to its locality aware reconstruction property, meaning that only appropriate knowledge is transferred from one local space in the source/target to another local space in the target/source. Two representative methods are LTSL [Shao *et al.*, 2014] and L^{2}TSL [Ding *et al.*, 2014], which explicitly impose low-rank constraint on the data reconstruction or latent factor in a learned common subspace. Those methods only employ a shallow architecture containing a single layer. However, knowledge transfer can be better learned from multiple layers with a deep structure.

Most recent researches on deep structure learning to capture a better feature representation attract increasing interest [Chen *et al.*, 2012; Nguyen *et al.*, 2013; Zhou *et al.*, 2014; Chen *et al.*, 2014], since discriminative information can be embedded in multiple levels of the features hierarchy. In fact, this is one of the major motivations to develop deep structure learning framework, so that more complex abstraction can be captured. However, current deep transfer learning methods failed to align different domains and learn deep structure features simultaneously. Without any knowledge about target domain, the feature extraction process performed on the source data would definitely ignore information important to the target domain.

In this paper, we propose a Deep Low-Rank Coding framework (DLRC) for transfer learning. The core idea of DLRC is to jointly learn a deep structure of feature representation and transfer knowledge via an iterative structured low-rank constraint, which aims to deal with the mismatch between source and target domains layer by layer (Figure 1). Our main contributions are summarized as:

- A deep structure is designed to capture the rich information across source and target domains. Specifically, the deep structure is stacked by multiple layer-wise low-rank codings. Therefore, it can refine features for source and target in a layer-wise fashion and preserve more essential information to the target domain.
- An iterative structure term is developed for each Single-layer Low-Rank Coding (SLRC), which works in a local-aware reconstruction manner. Through labeling

*This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

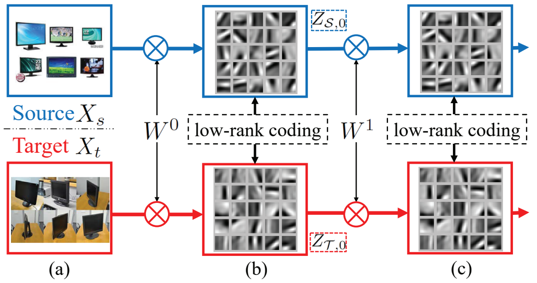


Figure 1: Illustration of our Deep Low-Rank Coding (DLRC). Input (a) is the original data of source (blue) and target (red) domains. (b) represents the first-layer low-rank coding guided by marginal denoising regularizer and iterative structure term. Marginal denoising regularizer aims to learn a transformation matrix W^0, whilst iterative structure is designed to guarantee the low-rank coding to have prior information, which is updated in a layer-wise manner. (c) denotes the second-layer low-rank coding, whose input is the low-rank coding produced from the first-layer (b) $Z_{\mathcal{S},0}$ for source and $Z_{\mathcal{T},0}$ for target, respectively. The whole framework stacks such multiple layers as (b) together to learn multi-level discriminative features across two domains.

most confident samples in target domain, the learned features become more discriminative, since the marginal and conditional disparities are both leveraged.

- Marginal denoising regularizer is incorporated to guide the low-rank coding by seeking a robust and discriminative transformation shared by two domains, which is jointly optimized with low-rank reconstruction by uncovering rich information from complex data across two domains.

2 Related Work

In this section, we briefly discuss some related works, and highlight the differences between them and our method.

Transfer learning has been widely discussed recently and for the survey of state-of-the-art methods, please refer to [Pan and Yang, 2010]. Recently, low-rank transfer learning has been well-studied to ensure that accurate data alignment is achieved after data adaptation [Shao *et al.*, 2014; Ding *et al.*, 2014; Ding and Fu, 2014]. The low-rank constraint enforced on the reconstruction coefficients matrix between domains is able to reveal underlying data structure, especially when the data lie in multiple subspaces, which can guide the conventional transfer subspace learning. Different from existing methods in this line, we introduce iterative structure learning to recover the low-rank structure of the coefficient matrix in a supervised way. Furthermore, we employ the low-rank constraint on the data transformed by a mapping learned from marginal denoising regularizer, and therefore our method is more robust to corrupted data.

Most recently, the thought of deep structure is incorporated into transfer learning to uncover the rich information across domains. Chen et al. developed marginalized Stacked denoising Autoencoder (mSDA) to learn a better representation by reconstruction, recovering original features from data that are artificially corrupted with noise [Chen *et al.*, 2012]. Zhou et al. managed to learn a feature mapping between cross-domain heterogeneous features as well as a better feature representation for mapped data to reduce the bias issue caused by the cross-domain correspondences [Zhou *et al.*, 2014]. In this paper, we also adopt the thought of deep transfer learning, however, our method jointly learns the low-rank codings and transfers knowledge from source to target in a unified deep structure framework. By stacking multiple layers' low-rank coding, we build a deep structure to capture more discriminative features across two domains.

3 Deep Low-Rank Coding

In this section, we first briefly discuss our motivation, then propose our single-layer low-rank coding with its solution. Finally, we introduce our deep low-rank coding framework by stacking single-layer low-rank coding to multiple layers.

3.1 Motivation

Recently, mSDA [Chen *et al.*, 2012] and its variants [Zhou *et al.*, 2014], achieve exciting recognition results for transfer learning by extracting layer-wise features across different domains. These works stack marginalized denoising Autoencoder (mDA) layer by layer to capture the rich and discriminative features. mDA has shown the effectiveness in transfer learning and proven to be much more efficient [Chen *et al.*, 2012], due to its linear property.

Considering previous work only learn deep structure feature [Chen *et al.*, 2012], or separately learn feature and transfer knowledge [Zhou *et al.*, 2014], we propose to refine layer-wise features and align different domains in a unified framework. In such way, knowledge from source domain can be transferred to the target one layer by layer, which guides low-rank coding to produce more discriminative and important feature to the target domain. In the following sections, we will present our Deep Low-Rank Coding (DLRC) based on Single-layer Low-Rank Coding (SLRC).

3.2 Single-layer Low-Rank Coding

Given a set of target domain $X_{\mathcal{T}} = \{x_{\mathcal{T},1}, \cdots, x_{\mathcal{T},n_{\mathcal{T}}}\}$ with $n_{\mathcal{T}}$ unlabeled data points and a set of source domain $\{X_{\mathcal{S}}, Y_{\mathcal{S}}\} = \{(x_{\mathcal{S},1}, y_{\mathcal{S},1}), \cdots, (x_{\mathcal{S},n_{\mathcal{S}}}, y_{\mathcal{S},n_{\mathcal{S}}})\}$ with $n_{\mathcal{S}}$ labeled data points and $Y_{\mathcal{S}}$ is the label vector. Assume $X = [X_{\mathcal{S}}, X_{\mathcal{T}}] \in \mathbb{R}^{d\times n}$, where d is the original dimension of two domains and $n = n_{\mathcal{S}} + n_{\mathcal{T}}$ is the total size of two domains.

Our Single-layer Low-Rank Coding (SLRC) adopts the thought of conventional low-rank transfer learning [Shao *et al.*, 2014; Ding *et al.*, 2014] to seek discriminative low-rank codings. With its locality-aware reconstruction property, marginal distribution divergence across source and target domains would be reduced so that well-established source knowledge can be passed to target domain. Therefore, we develop the following objective function as:

$$\min_{Z,W} \ \mathrm{rank}(Z) + \lambda\Omega(W), \ \ \text{s.t.} \ \ WX = WX_{\mathcal{S}}Z, \qquad (1)$$

where $\mathrm{rank}(Z)$ is the operator to calculate the rank of low-rank coding matrix $Z \in \mathbb{R}^{n_{\mathcal{S}} \times n}$, which can be solved with nuclear norm [Liu *et al.*, 2013]. $W \in \mathbb{R}^{d \times d}$ is the transformation matrix (or rotation) on original data shared by two domains. $\Omega(W)$ is the loss function concerned W and λ is the trade-off parameter.

To seek a better transformation matrix W in low-rank constraint, we incorporate recent popular mDA [Chen *et al.*, 2012], which is designed to seek a mapping W from the original data to the corrupted one so that the learned W is robust to corrupted data. mDA has an advantage on efficient performance and small computational cost, whose objective function is formulated as follows:

$$\Omega(W) = \mathrm{tr}\big[(\bar{X} - W\tilde{X})^{\mathrm{T}}(\bar{X} - W\tilde{X})\big], \tag{2}$$

where $\bar{X}$ is the composition of X by repeating m times, and $\tilde{X}$ is the corrupted version of $\bar{X}$ with different ratios of corruption. And $\mathrm{tr}(\cdot)$ is the operator to calculate the trace of a matrix. Eq. (2) manages to minimize the original data with its transformed corrupted version so that the learned transformation is robust to noise and captures more shared discriminative information across domains. In this way, the learned transformation matrix would well leverage the disparity of two domains.

It should be noted that the single-layer low-rank coding we discussed only relies on data distributions. However, we are always accessible to labels of source domain in transfer learning. Therefore, we could pre-load these label information into model (1) where whole data with certain labels are only reconstructed by source data with the corresponding labels. Similar thought has been discussed in [Zhang *et al.*, 2013] where image codings are guided through structured low-rank constraint. Then, we propose the final objective function:

$$\begin{aligned} \min_{W,Z} \, & \|Z\|_* + \lambda \mathrm{tr}(\mathcal{E}^{\mathrm{T}}\mathcal{E}) + \alpha\|Z_l - H\|_{\mathrm{F}}^2, \\ & \text{s.t. } WX = WX_{\mathcal{S}}Z, \end{aligned} \tag{3}$$

where α is the balancing parameter and $\mathcal{E} = \bar{X} - W\tilde{X}$. $\|\cdot\|_*$ is the nuclear norm, which is a surrogate of $rank()$ to seek a low-rank representation, whilst $\|\cdot\|_{\mathrm{F}}$ is the Frobenius Norm, which aims to make the labeled representation Z_l approximate to the structure matrix H. This structure term is optimized layer by layer, since most confident samples will be labeled in the target domain (refer the detail to Section 3.4). Z_l is the labeled partial columns out of Z, which includes all source samples and partial target samples. We define $Z = [Z_l, Z_u]$, where each column of Z_u is correlated to unlabeled sample in target domain after each layer's optimization.

Discussion: Different from previous low-rank transfer learning methods [Shao *et al.*, 2014; Ding *et al.*, 2014], which employ the target domain to reconstruct the source one or opposite direction, we treat the transformed source domain as the dictionary and employ it to reconstruct the transformed whole data from two domains. Such constraint would optimize W, coupling source with target and also itself. Furthermore, previous ones deploy low-rank constraint on the data lying in the common subspace projection. However, our low-rank coding reconstructs the transformed data with a linear mapping learned from mDA, which would capture more discriminative and robust information shared by two domains.

Our single-layer low-rank coding (3) is developed to seek discriminative codings Z, which is guided with an iterative structured term and optimized under the transformed data via mDA [Chen *et al.*, 2012]. In this way, single-layer low-rank coding can mitigate both the marginal and conditional distributions across two domains, and therefore, it potentially transfers knowledge from source to target and boosts the recognition performance to the target domain. Furthermore, we can stack the single-layer low-rank coding into a deep structure, where the output coding $Z = [Z_{\mathcal{S}}, Z_{\mathcal{T}}]$ from the previous layer would be the input of the next layer. $Z_{\mathcal{S}}$ is the low-rank coding for source, while $Z_{\mathcal{T}}$ is for target.

3.3 Optimization Solution

To solve Eq. (3), we first introduce a relaxing variable J and convert it to the following equivalent problem as:

$$\begin{aligned} \min_{W,Z,J} \, & \|J\|_* + \lambda \mathrm{tr}(\mathcal{E}^{\mathrm{T}}\mathcal{E}) + \alpha\|Z_l - H\|_{\mathrm{F}}^2, \\ & \text{s.t. } WX = WX_{\mathcal{S}}Z, \;\; Z = J, \end{aligned} \tag{4}$$

which can be solved via the Augmented Lagrange Multiplier (ALM) method [Lin *et al.*, 2010]. Since $Z = [Z_l, Z_u]$,we introduce an auxiliary matrix $\mathcal{H} = [H, Z_u]$. We have the augmented Lagrangian function of Eq. (4) as:

$$\begin{aligned} & \|J\|_* + \lambda \mathrm{tr}(\mathcal{E}^{\mathrm{T}}\mathcal{E}) + \alpha\|Z - \mathcal{H}\|_{\mathrm{F}}^2 \\ & + \mathrm{tr}(Y_1^{\mathrm{T}}(WX - WX_{\mathcal{S}}Z)) + \mathrm{tr}(Y_2^{\mathrm{T}}(Z - J)) \\ & + \tfrac{\mu}{2}(\|WX - WX_{\mathcal{S}}Z\|_{\mathrm{F}}^2 + \|Z - J\|_{\mathrm{F}}^2), \end{aligned} \tag{5}$$

where Y_1 and Y_2 are the two Lagrange multipliers and $\mu > 0$ is the penalty parameter. Each variable in optimization (5) can be addressed in an iterative manner by updating J, Z, W one by one. Then, those variables are optimized in the $t + 1$ iteration as follows:

Update J:

$$\begin{aligned} & J_{t+1} \\ = \; & \arg\min_J \|J\|_* + \mathrm{tr}(Y_{2,t}^{\mathrm{T}}(Z_t - J)) + \tfrac{\mu_t}{2}\|Z_t - J\|_{\mathrm{F}}^2 \\ = \; & \arg\min_J \tfrac{1}{\mu_t}\|J\|_* + \tfrac{1}{2}\|J - (Z_t + \tfrac{Y_{2,t}}{\mu_t})\|_{\mathrm{F}}^2, \end{aligned} \tag{6}$$

which can be solved by Singular Value Thresholding (SVT) [Cai *et al.*, 2010].

Update Z:

$$\begin{aligned} Z_{t+1} = \; & \arg\min_Z \alpha\|Z - \mathcal{H}\|_{\mathrm{F}}^2 + \mathrm{tr}(Y_{1,t}^{\mathrm{T}}W_t(X - X_{\mathcal{S}}Z)) \\ & + \mathrm{tr}(Y_{2,t}^{\mathrm{T}}(Z - J_{t+1})) + \tfrac{\mu_t}{2}(\|W_t(X - X_{\mathcal{S}}Z)\|_{\mathrm{F}}^2 \\ & + \|Z - J_{t+1}\|_{\mathrm{F}}^2), \end{aligned}$$

which is convex and has closed form solution as follows:

$$\begin{aligned} Z_{t+1} = \; & \big((2\alpha + \mu_t)\mathrm{I}_z + \mu_t\Psi_t^{\mathrm{T}}\Psi_t\big)^{-1}\big(\Psi_t^{\mathrm{T}}Y_{1,t} \\ & - Y_{2,t} + \mu_t\Psi^{\mathrm{T}}W_tX + \mu_tJ_{t+1} + 2\alpha\mathcal{H}\big), \end{aligned} \tag{7}$$

where I_z is the identity matrix of size $n_{\mathcal{S}} \times n_{\mathcal{S}}$ and $\Psi_t = W_tX_{\mathcal{S}}$.

Update W:

$$\begin{array}{rl} W_{t+1} = & \underset{W}{\arg\min}\, \lambda \mathrm{tr}\big[(\bar{X} - W\tilde{X})^{\mathrm{T}}(\bar{X} - W\tilde{X})\big] + \\ & \mathrm{tr}(Y_{1,t}^{\mathrm{T}} W R_t) + \frac{\mu_t}{2}\|WR_t\|_{\mathrm{F}}^2, \end{array} \tag{8}$$

where $R_t = X - X_{\mathcal{S}}Z_{t+1}$. Eq. (8) is convex and we can achieve its closed form solution by defining $P = \bar{X}\tilde{X}^{\mathrm{T}}$ and $Q = \tilde{X}\tilde{X}^{\mathrm{T}}$:

$$W_{t+1} = (Y_{1,t}R_t^{\mathrm{T}} + \lambda P)(\lambda Q - \mu_t R_t R_t^{\mathrm{T}})^{-1} = \hat{P}_t\hat{Q}_t^{\ -1},$$

where the repeated number m for $\bar{X}$ is expected to be ∞, giving rise to a robust denoising transformation W_{t+1} learned from infinitely many copies of noisy data. Fortunately, the matrices $\hat{P}_t$ and $\hat{Q}_t$ converge to their expectations when m becomes very large with the weak law of large numbers. In this way, we can derive the expected values of $\hat{P}_t$ and $\hat{Q}_t$, and calculate the corresponding mapping W_{t+1} as:

$$\begin{array}{rl} & W_{t+1} \\ = & \mathbb{E}[\hat{P}_t]\mathbb{E}[\hat{Q}_t]^{-1} \\ = & \mathbb{E}[\lambda P + Y_{1,t}R_t^{\mathrm{T}}]\mathbb{E}[\lambda Q - \mu_t R_t R_t^{\mathrm{T}}]^{-1} \\ = & \big(\lambda\mathbb{E}[P] + \mathbb{E}[Y_{1,t}R_t^{\mathrm{T}}]\big)\big(\lambda\mathbb{E}[Q] - \mathbb{E}[\mu_t R_t R_t^{\mathrm{T}}]\big)^{-1} \\ = & \big(\lambda\mathbb{E}[P] + Y_{1,t}R_t^{\mathrm{T}}\big)\big(\lambda\mathbb{E}[Q] - \mu_t R_t R_t^{\mathrm{T}}\big)^{-1} \end{array} \tag{9}$$

where $Y_{1,t}R_t^{\mathrm{T}}$ and $\mu_t R_t R_t^{\mathrm{T}}$ are treated as constant values when optimizing W_{t+1}. The expectations $\mathbb{E}[P]$ and $\mathbb{E}[Q]$ can be derived in a similar way as in mDA [Chen *et al.*, 2012]. The detailed optimization is outlined in **Algorithm 1**.

```
Algorithm 1 Solving Problem (3) by ALM
Input: $X = [X_{\mathcal{S}}, X_{\mathcal{T}}], \lambda, \alpha, H$,
Initialize: $W_0 = Z_0 = J_0 = Y_{1,0} = Y_{2,0} = 0, \mu_0 = 10^{-6}$,
    $\mu_{\max} = 10^6, \rho = 1.1, \varepsilon = 10^{-6}, t = 0$.
while not converged do
  1. Fix others and update $J_{t+1}$ by Eq. (6);
  2. Fix others and update $Z_{t+1}$ by Eq. (7);
  3. Fix others and update $W_{t+1}$ by Eq. (9);
  4. Update two multipliers via
     $Y_{1,t+1} = Y_{1,t} + \mu_t W_{t+1}(X - X_{\mathcal{S}}Z_{t+1})$;
     $Y_{2,t+1} = Y_{2,t} + \mu_t(Z_{t+1} - J_{t+1})$;
  5. Update $\mu$ via $\mu_{t+1} = \min(\rho\mu_t, \mu_{\max})$;
  6. Check the convergence conditions:
     $\|W_{t+1}(X - X_{\mathcal{S}}Z_{t+1})\|_\infty < \varepsilon$, $\|Z_{t+1} - J_{t+1}\|_\infty < \varepsilon$.
  7. $t = t + 1$.
end while
output: $Z, J, W$
```

3.4 Deep Low-Rank Coding

So far, model (3) works in a single-layer way to capture the shared information between two domains and meanwhile couple them in an iterative structure low-rank constraint.

As illustrated in our framework (Figure 1), we design a deep structure to learn more discriminative and richer information from source and target domains in a layer-wise manner. That is, we stack single-layer model (3) into multi-layer structure. Each single layer produces iteratively structured low-rank coding for both domains $Z_{\mathcal{S}}$ and $Z_{\mathcal{T}}$, which would be the input of next layer. Specifically, the output from the k-1th layer $Z_{\mathcal{S},k-1}$ and $Z_{\mathcal{T},k-1}$ would be the input of the kth layer, which produces $Z_{\mathcal{S},k}$ and $Z_{\mathcal{T},k}$. In such a layer-wise scheme, DLRC would generate multi-level features for both domains and refine them from coarse to fine. The details of DLRC are shown in **Algorithm 2**. In the experiments, we employ five-layer features and combine them together to evaluate the final performance of our DLRC.

```
Algorithm 2 Algorithm of Deep Low-Rank Coding (DLRC)
Input: $X_{\mathcal{S}}, X_{\mathcal{T}}, L$ is the number of layers,
for $k$ = 1 to $L$ do
  1. Use Algorithm 1 to learn coding $Z_{\mathcal{S},k}$ and $Z_{\mathcal{T},k}$;
  2. Set $X_{\mathcal{S},k+1} = Z_{\mathcal{S},k}$ and $X_{\mathcal{T},k+1} = Z_{\mathcal{T},k}$;
  3. Update $H_k$ via Eq. (10);
end for
output: Low-rank codings $\{Z_{\mathcal{S},k}, Z_{\mathcal{T},k}\}, (k = 1, \cdots, L)$.
```

For each layer, we need to update the iterative structure matrix H by introducing the pseudo labels of most confident samples in target domains. Suppose we label $n_{\mathcal{T}}^k$ samples from the target domain in the kth layer, and therefore, H_k in the k^{th} should be an $n_{\mathcal{S}} \times (n_{\mathcal{S}} + n_{\mathcal{T}}^k)$ matrix. $H_k^{i,j}$ denotes the element of i-th row and j-th column in H_k. We seek $H_k^{i,j}$ through:

$$H_k^{i,j} = \frac{s(W^k x_i, W^k x_j)}{\sum_{y_i = y_j} s(W^k x_i, W^k x_j)}, \tag{10}$$

where y_i denotes the label of x_i from the labeled source and pseudo-labeled target domains. W^k is the transformation matrix in the kth layer. And $s(W^k x_i, W^k x_j) = \exp(-\|W^k x_i - W^k x_j\|^2/2\sigma^2)$ is Gaussian kernel function with σ as bandwidth (we set $\sigma = 1$ in our experiment). In this way, we can achieve the structure matrix H_k, which guides the low-rank reconstruction to minimize the conditional distribution between source and target domains. Since it is optimized layer by layer, we define it as *iterative structure learning*. In the experiments, we first employ the nearest neighbour classifier to predict the labels of target data using source data. Then, we label 50% target samples, which are most closest to the labeled source data according to the Euclidean distances.

3.5 Complexity Analysis

The time-cost parts of our DLRC are (1) Trace norm computation in Eq. (6); (2) Matrix multiplication and inverse in Eqs. (7) and (9).

First, Eq. (6) solved by SVD computation would cost $\mathcal{O}(n_{\mathcal{S}}^2 n)$ for $J \in \mathbb{R}^{n_{\mathcal{S}} \times n}$. Generally, $n_{\mathcal{S}}$ is the same order of magnitude with n. When n is very large, this step would be computationally expensive. But Eq. (6) can be improved to $\mathcal{O}(rn^2)$ by accelerations of SVD, where $r \ll n$ is the rank of J. Second, Eqs. (7) and (9) both include a few matrix multiplications and a matrix inverse operation. Therefore, Eq. (7) takes $(l_1 + 1)\mathcal{O}(n^3)$ and Eq. (9) would take $(l_2 + 1)\mathcal{O}(d^3)$, where l_1 and l_2 are the number of multiplications for Eq. (7) and Eq. (9), respectively. In sum, the total cost of each single-layer low-rank coding is: $T_{\mathrm{SLRC}} = \mathcal{O}(t(rn^2 + (l_1 + 1)n^3 + (l_2 + 1)d^3))$, where t is the iteration of **Algorithm 1**. Finally, the total cost of DLRC is LT_{SLRC}, where L is the number of layers.

4 Experimental Results

In this section, we evaluate our proposed method on several benchmarks. We will first introduce the datasets and experimental setting. Then comparison results will be presented followed by some properties analysis and discussion.

4.1 Datasets & Experimental Setting

MRSC+VOC includes two datasets: (1) MSRC dataset[1] is provided by Microsoft Research Cambridge, which contains 4,323 images labeled by 18 classes; (2) VOC2007 dataset[2] contains 5,011 images annotated with 20 concepts. They share the following 6 semantic classes: *aeroplane*, *bicycle*, *bird*, *car*, *cow*, *sheep*. We construct MSRC+VOC by selecting all 1,269 images in MSRC and all 1,530 images in VOC2007 following [Long *et al.*, 2013]. We uniformly rescale all images to be 256 pixels in length, and extract 128-dimensional dense SIFT (DSIFT) features.

USPS+MNIST[3] includes 10 common classes of digits from two datasets: (1) USPS dataset consists of 7,291 training images and 2,007 test images; (2) MNIST dataset has a training set of 60,000 examples and a test set of 10,000 examples. To speed up experiments, we randomly sample 1,800 images in USPS as one domain, and randomly select 2,000 images in MNIST as the other domain. We uniformly resize all images to 16×16, and represent each one by a feature vector encoding the gray-scale pixel values.

Reuters-215782[4] is a difficult text dataset with many top and subcategories. The three largest top categories are *orgs*, *people*, and *place*, each of which is comprised of many subcategories. For fair comparison, we adopt the preprocessed version of Reuters-21578 studied in [Gao *et al.*, 2008].

Office+Caltech-256[5] select 10 common categories from Office dataset and Caltech-256. Office dataset has been widely adopted as the benchmark for visual domain adaptation. It has three distinct domains: Amazon, Webcam, and DSLR, including 4652 images, and 31 common categories. Caltech-256 is a standard database for object recognition, including 30,607 images and 256 categories. We apply the 800-dim features by SURF+BagOfWords.

Note that the arrow "→" is the direction from "source" to "target". For example, "Webcam → DSLR" means Webcam is the source domain whilst DSLR is the target one. In the experiments, we learn five-layer features and combine them together to evaluate the final recognition performance through the nearest neighbor classifier.

4.2 Comparison Results

For **MRSC+VOC** and **USPS+MNIST**, we evaluate our algorithm by comparing with four baselines: TSC [Long *et al.*, 2013], TCA [Pan *et al.*, 2011], GFK [Gong *et al.*, 2012], TJM [Long *et al.*, 2014b]. Both two groups of datasets have two domains, therefore, we switch source and target to achieve two results for each group. The results are shown in Figure 2.

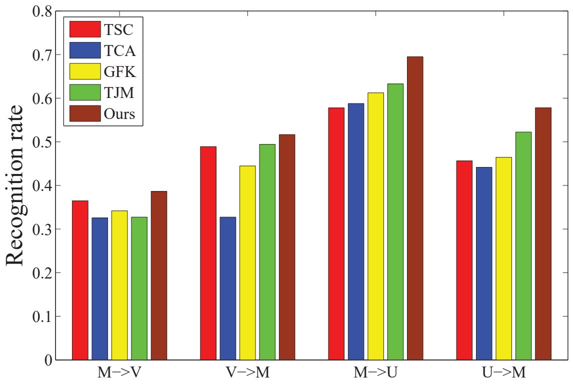


Figure 2: Recognition results of 5 algorithms on four cases from two groups of datasets: MSRC+VOC and USPS+MNIST. For MSRC+VOC, we have two cases, M→V and V→M, where M is short for MSRC and V for VOC. For USPS+MNIST, we also have two scenarios, M→U and U→M, where M represents MNIST and U denotes USPS.

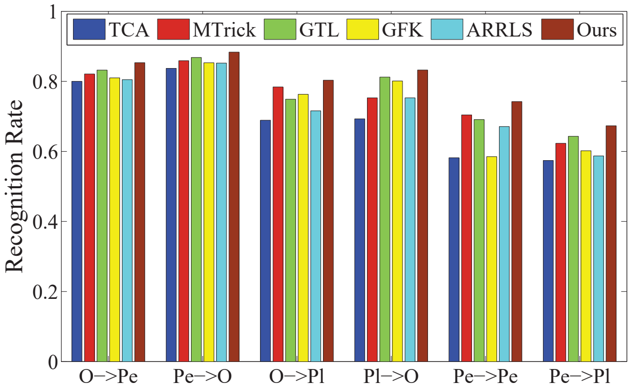


Figure 3: Recognition results of 6 algorithms on six different cases from three domains in Reuters-215782 text dataset, where Pe is short for *people*, O for *orgs*, and Pl for *place*, respectively.

For **Reuters-215782**, these five baselines: TCA [Pan *et al.*, 2011], MTrick [Zhuang *et al.*, 2011], GTL [Long *et al.*, 2014b], GFK [Gong *et al.*, 2012] and ARRLS [Long *et al.*, 2014a] are compared on six cases from three domains. The recognition results are listed in Figure 3.

For **Office+Caltech-256**, we compare the following baselines: SGF [Gopalan *et al.*, 2011], LTSL [Shao *et al.*, 2014], GFK [Gong *et al.*, 2012], TJM [Long *et al.*, 2014b], DASA [Fernando *et al.*, 2013], TCA [Pan *et al.*, 2011], mSDA [Chen *et al.*, 2012] and GUMA [Cui *et al.*, 2014]. We strictly follow the configuration of [Gong *et al.*, 2012] where 20 images per category from Amazon, Caltech-256, and Webcam. Since DSLR has a small number of samples, we do not use it as source domain. Finally, we conduct 3 × 3

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007

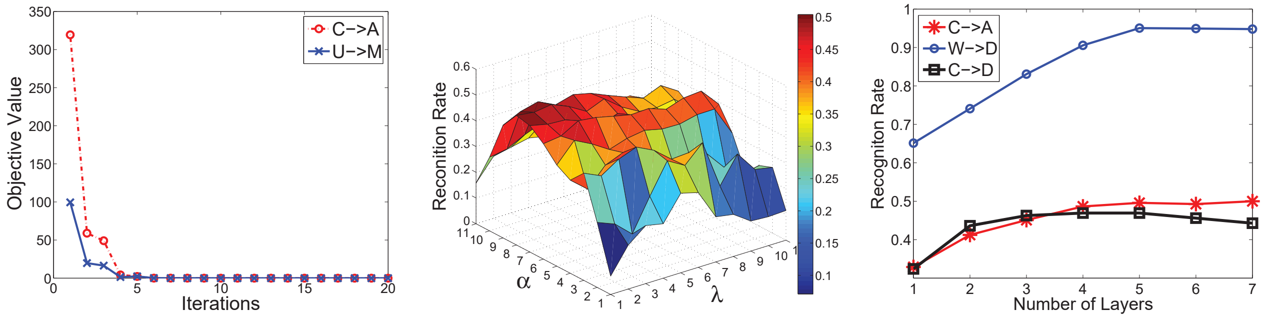
[3] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

[4] http://learn.tsinghua.edu.cn:8080/2011310560/mlong.html

[5] http://www-scf.usc.edu/∼boqinggo/domainadaptation.html

Table 1: Average recognition rate (%)± standard variation of 9 algorithms on Office+Caltech-256, where A = Amazon, D = DSLR, C = Caltech-256 and W = Webcam. Red color denotes the best recognition rates. Blue color denotes the second best recognition rates.

Config\Methods	SGF	DASA	GFK	LTSL	TJM	TCA	mSDA	GUMA	Ours
C→W	33.9±0.5	36.8±0.9	40.7±0.3	39.3±0.6	39.0±0.4	30.5±0.5	38.6±0.8	42.3±0.3	41.7±0.5
C→D	35.2±0.8	39.6±0.7	38.9±0.9	44.5±0.7	44.6±0.8	35.7±0.5	44.5±0.4	44.7±0.4	47.5±0.6
C→A	36.9±0.7	39.0±0.5	41.1±0.6	46.9±0.6	46.7±0.7	41.0±0.6	47.7±0.6	46.7±0.6	49.7±0.4
W→C	27.3±0.7	32.3±0.4	30.7±0.1	29.9±0.5	30.2±0.4	29.9±0.3	33.6±0.4	34.2±0.5	33.8±0.5
W→A	31.3±0.6	33.4±0.5	29.8±0.6	32.4±0.9	30.0±0.6	28.8±0.6	35.4±0.5	36.2±0.5	38.5±0.7
W→D	70.7±0.5	80.3±0.8	80.9±0.4	79.8±0.7	89.2±0.9	86.0±1.0	87.9±0.9	73.5±0.4	94.3±1.1
A→C	35.6±0.5	35.3±0.8	40.3±0.4	38.6±0.4	39.5±0.5	40.1±0.7	40.7±0.6	36.1±0.4	42.7±0.5
A→W	34.4±0.7	38.6±0.6	39.0±0.9	38.8±0.5	37.8±0.3	35.3±0.8	37.3±0.7	35.9±0.3	42.8±0.9
A→D	34.9±0.6	37.6±0.7	36.2±0.7	38.3±0.4	39.5±0.7	34.4±0.6	36.3±0.5	38.2±0.8	41.8±0.6



(a)

(b)

(c)

Figure 4: (a) Convergence curves of setting $C \rightarrow A$ on Office+Caltech and $U \rightarrow M$ on USPS+MNIST, where we only show 20 iterations. (b) Parameters analysis on λ and α of setting $C \rightarrow A$ on Office+Caltech, where the **x**-range and **y**-range from 1 to 11 means $[10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 1, 10, 50, 100, 500, 10^3]$, respectively. (c) represents the influence of different layers. Here we show three experiments on 7 layers to testify the recognition results with more layers' coding.

different groups of domain adaptation experiments. The recognition results are shown in Table 1.

Discussion: We experiment on such transfer learning scenarios, where we are only accessible to the labels of source domain. However, there are two lines. The first line, e.g. SGF, DASA, TCA, mSDA, trains in a totally unsupervised way, that is, the source label is not used in the training stage. The other line employs the source labels into training, e.g. GFK, LTSL and TSC, even introduces the pseudo labels of the target domains, e.g. TJM, ARRLS and Ours. From the results shown in Figures 2 & 3, and Table 1, we observe that our DLRC outperforms the compared baselines in most of cases under different scenarios on four benchmarks.

Compared with SGF and DASA, GFK, LTSL and TSC can achieve better results in most cases, since they incorporate the source label in order to transfer more useful knowledge to target domain. Based on this, TJM, ARRLS and Ours introduce the pseudo label of target domain into the training stage, therefore, more discriminative information can be learned in the training stage. However, mSDA in some cases performs better than other compared algorithms, which indicates that deep structure in feature learning could uncover more discriminative information across two domains. Our deep low-rank coding not only introduces the pseudo labels of the target domain, but also builds a deep feature learning framework. Therefore, our method could find plenty of rich information inside two domains and learn more helpful features for the target domains.

4.3 Properties Analysis

In this section, we evaluate on several properties of our DLRC. First, we analyze the convergence and influence of two parameters. Then, we testify the recognition performance of our DLRC with different layers. We show the evaluation results in Figure 4.

From Figure 4(a), we can observe our single-layer coding converges very fast, usually within 10-round iterations. The influence of parameters presents the recognition results on different values of two parameters in Figure 4(b). As we can see, α generates more important influence compared with λ. That means, our iterative structure term does play an important role in seeking more discriminative features for two domains. However, the larger value produces worse results. It results from the iterative structure term, which incorporates pseudo labels of target and they are not all accurate. Therefore, the larger α is, the more inaccurate information is introduced. In the experiments, we usually choose $\alpha = 10$

and $\lambda = 1$. From Figure 4(c), we witness that DLRC generally achieves better performance when the layer goes deeply. That is, more discriminative information shared by two domains can be uncovered with our deep low-rank coding. In other words, features would be refined from coarse to fine in a layer-wise fashion. However, we also observe that much deeper structure would bring negative transfer and decrease the recognition performance (see case $C \rightarrow D$ in Figure 4(c)). In the experiments, we achieve five-layer features and combine them together to do the final evaluation.

5 Conclusion

In this paper, we developed a Deep Low-Rank Coding (DLRC) framework for transfer learning. First, single-layer low-rank coding guided by iterative structure learning is incorporated to align two domains, by minimizing the marginal and conditional distributions across two domains. Meanwhile, marginal denoising regularizer aims to guide the low-rank reconstruction by seeking a better transformation matrix. Finally, by stacking several single-layer low-rank transfer codings, we obtain multi-layer features with more discrimination to target domain. Experimental results on several benchmarks have demonstrated the superior of our proposed algorithm, compared with the state-of-the-art transfer learning methods.

References

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIOPT*, 20(4):1956–1982, 2010.

[Chen *et al.*, 2012] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *ICML*, pages 767–774, 2012.

[Chen *et al.*, 2014] Minmin Chen, Kilian Q Weinberger, Fei Sha, and Yoshua Bengio. Marginalized denoising auto-encoders for nonlinear representations. In *ICML*, pages 1476–1484, 2014.

[Cui *et al.*, 2014] Zhen Cui, Hong Chang, Shiguang Shan, and Xilin Chen. Generalized unsupervised manifold alignment. In *NIPS*, pages 2429–2437, 2014.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119, 2014.

[Ding *et al.*, 2014] Zhengming Ding, Ming Shao, and Yun Fu. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, pages 1192–1198, 2014.

[Fernando *et al.*, 2013] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, pages 2960–2967, 2013.

[Gao *et al.*, 2008] Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *KDD*, pages 283–291, 2008.

[Gong *et al.*, 2012] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.

[Gopalan *et al.*, 2011] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006, 2011.

[Lin *et al.*, 2010] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE TPAMI*, 35(1):171–184, 2013.

[Long *et al.*, 2013] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S Yu. Transfer sparse coding for robust image representation. In *CVPR*, pages 407–414, 2013.

[Long *et al.*, 2014a] Mingsheng Long, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, et al. Adaptation regularization: A general framework for transfer learning. *TKDE*, 26(5):1076–1089, 2014.

[Long *et al.*, 2014b] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *IEEE TKDE*, 26(7):1805–1818, 2014.

[Nguyen *et al.*, 2013] Hien V Nguyen, Huy Tho Ho, Vishal M Patel, and Rama Chellappa. Joint hierarchical domain adaptation and feature learning. *IEEE TPAMI*, 2013.

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

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE TNN*, 22(2):199–210, 2011.

[Shao *et al.*, 2012] Ming Shao, Carlos Castillo, Zhenghong Gu, and Yun Fu. Low-rank transfer subspace learning. In *ICDM*, pages 1104–1109. IEEE, 2012.

[Shao *et al.*, 2014] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *IJCV*, pages 1–20, 2014.

[Shekhar *et al.*, 2013] Sumit Shekhar, Vishal M Patel, Hien V Nguyen, and Rama Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, pages 361–368, 2013.

[Zhang *et al.*, 2013] Yangmuzi Zhang, Zhuolin Jiang, and Larry S Davis. Learning structured low-rank representations for image classification. In *CVPR*, pages 676–683, 2013.

[Zhou *et al.*, 2014] 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.

[Zhuang *et al.*, 2011] Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, and Zhongzhi Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization? *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):100–114, 2011.