Traffic Sign Recognition with Transfer Learning

Xishuai Peng
School of Aeronautics
and Astronautics
Shanghai Jiao Tong University
Shanghai,China 200240
Email: xishuai_peng@sjtu.edu.cn

Yuanxiang Li
School of Aeronautics
and Astronautics
Shanghai Jiao Tong University
Shanghai,China 200240
Email: yuanxli@sjtu.edu.cn
Tel.: +86-21-34206160
fax: +86-21-34206320

Xian Wei
School of Aeronautics
and Astronautics
Shanghai Jiao Tong University
Shanghai,China 200240
Email: xian.wei@tum.de

Jianhua Luo
School of Aeronautics
and Astronautics
Shanghai Jiao Tong University
Shanghai,China 200240
Email: jhluo@sjtu.edu.cn

Yi Lu Murphey
Department of Electrical
and Computer Engineering
University of Michigan-Dearborn
MI,USA 48128
Email:yilu@umich.edu

***Abstract*—Traffic signs are characterized by a wide variability in their visual appearance in real-world environments. Supervised algorithms have achieved superior results on German Traffic Sign Recognition Bench-mark (GTSRB) database. However, these models cannot transfer knowledge across domains, e.g. transfer knowledge learned from Synthetic Signs database to recognize the traffic signs in GTSRB database. Through Synthetic Signs database shares exactly the same class label with GTSRB, the data distribution between them are divergent. Such task is called transfer learning, that is a basic ability for human being but a challenge problem for machines.**

In order to make these algorithms have ability to transfer knowledge between domains, we propose a variant of Generalized Auto-Encoder (GAE) in this paper. Traditional transfer learning algorithms, e.g.Stacked Autoencoder(SA), usually attempt to reconstruct target data from source data or man-made corrupted data. In contrast, we assume the source and target data are two different corrupted versions of a domain-invariant data. And there is a latent subspace that can reconstruct the domain-invariant data as well as preserve the local manifold of it. Therefore, the domain-invariant data can be obtained not only by de-noising from the nearest source and target data but also by reconstructing from the latent subspace. In order to make the statistical and geometric property preserved simultaneously, we additionally propose a Local Coordinate Coding (LCC)-based relational function to construct the deep nonlinear architecture. The experimental results on several benchmark datasets demonstrate the effectiveness of our proposed approach in comparison with several traditional methods.

***Index Terms*—transfer learning, traffic sign recognition, stacked auto-encoder, local coordinate coding, deep learning**

I. Introduction

Traffic Sign Recognition (TSR) is a crucial task with respect to many applications, such as self-driving, driver-less vehicles, traffic mapping and traffic surveillance. The public dataset German Traffic Sign Recognition Benchmark (GTSRB) [1] provides a number of difficult challenges such as viewpoint variations, poor illumination conditions, motion-blur, occlusions, colors fading and low resolutions. Convolutional Neural Networks (CNN) with supervised training has proven to be a superior method on this dataset. However, the test performance will significantly decay when they are trained on some related databases, e.g. Synthetic Signs database. Through Synthetic Signs database shares exactly the same class label with GTSRB, the data distribution between them are divergent. Such kind of task is called Transfer Learning (TL), it is a basic ability for human being but a challenge problem for machines.

To address such difficulty, TL methods [2], [3] transfer knowledge from related source domain. Domain adaptation (DA) is one category of TL, it focuses on dealing with data that share common class label but follow divergent data distribution. It has been widely studied in many real world applications, e.g., image classification [4], text categorization [5] and video event recognition [6].

The basic assumption of DA is there some common latent spaces are shared by involved domains. Under this subspace, the difference between the source and target domains is reduced as well as the original properties are preserved as much as possible. Therefore, the mechanism of most approaches is to explore the common latent factors and utilize them to mitigate both the marginal and conditional distribution, i.e., instance re-weighting [7] and feature extraction [8], [9]. Among these approaches, the feature-based TL methods have proven to be superior for the scenarios where original raw data between domains are very different [10].

Most recent researches on Stacked Autoencoders (SA) learn to capture a more domain-invariant feature representation. It attracts increasing interest, due to the superior power of its nonlinearity and deep architecture [5], [8]–[11]. [12] demonstrates that the Stacked Denosing Autoencoders (SDA) learned features can boost the performance on the sentiment analysis

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

across different product domain. [8] proposes marginalized SDA (mSDA), it addresses two limitations of SDA: highly computational cost and lack of stability with high-dimensional features. [5] jointly learns the low-rank coding and transfer knowledge from source to target in a deep mSDA structure. These methods attempt to learn domain-invariant feature representation by preserving the statistical property of data.

On the other hand, some methods endeavor to preserve the geometric structure of data. [13] proposes Generalized Auto-Encoder (GAE) to capture the structure of data space through minimizing the weighted distance between data and its local neighbors. [14] proposes Bi-shift Auto-Encoder (BAE), which attempts to minimize the distance of instances in the source and target manifolds. [15] proposes Geodesic Flow Kernel (GFK) to characterize the domain shift by integrating an infinite number of subspaces.

In fact, these two strategies, to preserve statistical and geometric property, are complementary with each other and jointly exploration could benefit from both sides. [16] supposes the two properties are two observations of data from different viewpoints. They propose a "shallow" framework of feature selection and geometric structure preservation for unsupervised domain adaptation.

Though their remarkable and promising results, these methods cannot employ a nonlinear model to preserve statistical and geometric property simultaneously. In this paper, we propose an unsupervised representation learning method for domain adaptation. It is based on following two assumptions: (1) Both the source and target data are corrupted versions of a domain-invariant data. And the domain-invariant data can be reconstructed not only from the nearest source and target data but also from the latent subspace. (2) The latent subspace has the property to reconstruct the domain-invariant data as well as to preserve the local manifold of it. The proposed model has a superior non-linear power that ensures the statistical and geometric property of both domains can be preserved simultaneously. Moreover, the proposed Local Coordinate Coding (LCC)-based relational function is able to find more reasonable cross-domain neighbors along the training. Our main contributions are summarized as follows:

- With the two assumptions, we propose a objective function for learning a reconstruct-able latent subspace. In this subspace, the statistical and geometric property of domain-invariant data can be preserved simultaneously.
- We propose a variant of GAE and expand its scope of application to DA. It results in exploring a more meaningful local manifold structure for the domain-invariant data.
- We propose a relational function to weight the reconstruction error for each instance, which is inspired by LCC for enhancing the locality.

The remainder of this paper is organized as follows. Section II presents some brief discussion with related works. Section III introduces the proposed method in detail. Experiments are reported in Section IV, and Section V is the conclusion.

II. Related Works

Most recently some researchers have realized the combination of geometric and statistic property can transcend the specific limitations of each perspective. [16] joints feature selection and geometric structure preservation for unsupervised DA. Their method try to find a linear projection matrix p, which is constrained by $l_{2,1} - norm$, that make relevant features selection and transformation learning simultaneous. However, the linear projection enforces the transformation consistency and it might be quite difficult to remove the discrepancy between the source and target domains completely. In contrast, we employ non-linear transformation to model and moreover enhance its non-linear power by stacking multiple auto-encoders to construct a deep structure.

Another class of very promising techniques that has recently gained superior performance is to learn a low-rank code for both source and target domains. [5] treats the transformed source domain as dictionary and employ it to reconstruct the transformed data from two domains. Compared to their works, our approach is a symmetric feature-based transfer learning method. We assume there exits domain-invariant data, which can be reconstructed by its similar local neighbors, including the nearest source and target data. Therefore, locality is our priori other than low-rank, which is inspired by the observations in several works [17]–[19]. They suggest a good first-order approximation to nonlinear function requires the codes to be local.

The original GAE [13] is proposed to discover the underlying effective manifold structure by explicitly modeling the data relation. It weight the reconstruction error of each instance by a relational function. These relational functions can be constructed by various supervised or unsupervised dimensionality reduction ideas, e.g., PCA [20], ISOMAP [21], LLE [22], LDA [23] . We propose a variant of GAE and expand its scope of application to DA. The core part of the variant of GAE is its LCC-based relational function, our experiments demostrate that is effective in finding meaningful neighbors across domains.

III. The Proposed Approach

A. Problem Definition

A domain D is defined by a feature space $\mathcal{X}$ and its probability distribution $P(X)$, where $X \in \mathcal{X}$. For a specific domain, a classification task T consists of class information $\mathcal{Y}$ and a classifier $f(x)$, that is $T = \{\mathcal{Y}, f(x)\}$. We use subscript s, t, o to indicate the source domain, target domain and domain-invariant domain or uncorrupted domain respectively. This paper focuses on the following problem:

Given a labeled source domain D_s and an unlabeled target domain D_t, where $D_s \neq D_t$, $\mathcal{Y}_s = \mathcal{Y}_t$, $P(x_s) \neq P(x_t)$ and $P(y_s|X_s) \neq P(y_t|X_t)$. We attempt to find a common subspace $\mathcal{Z}$, in which the features $Z \in \mathcal{Z}$ are able to reconstruct the uncorrupted data as well as to preserve the local manifold of it.

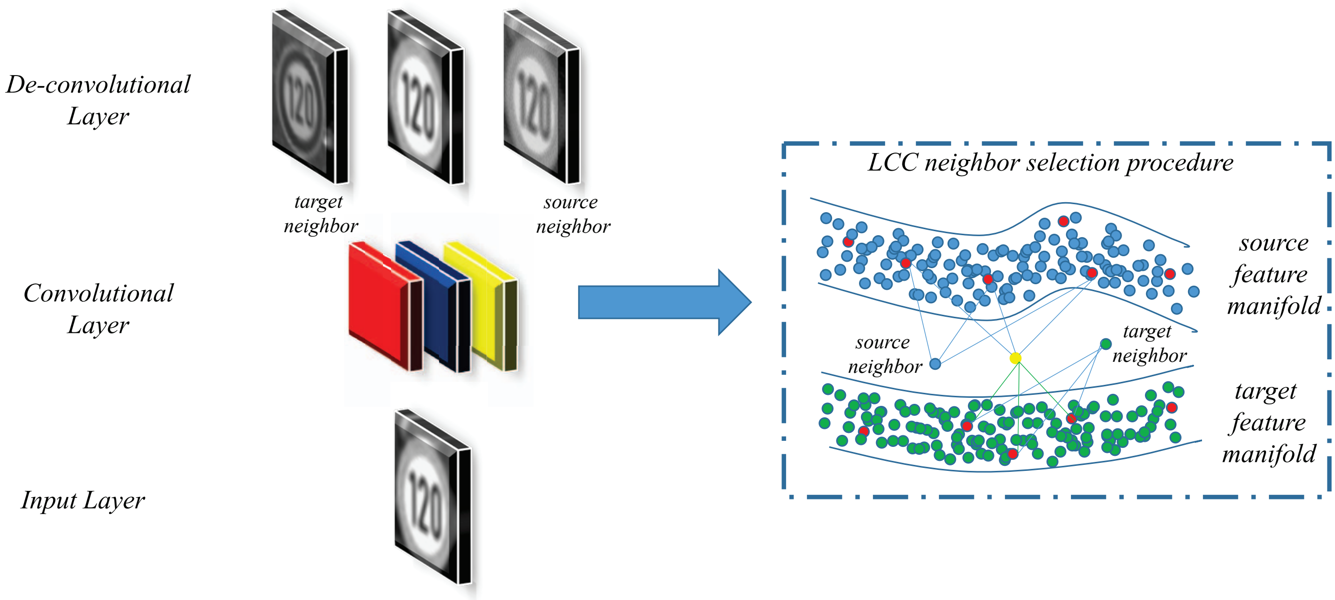


Fig. 1. the architecture of signal layer GAE

B. Problem Formulation

The basic assumption behind our approach is that the source X_s and target data X_t are two different corrupted versions of X_o. From this view point, X_o can be obtained by de-nosing from the nearest source and target data. Therefore, the problem can be simply formulated as follows, minimizing the weighted reconstruction error from the source and the target data:

$$X_o^* = \min_{X_o} E_{X_s,X_o} + E_{X_t,X_o} \tag{1}$$

where

$$E_{X_s,X_o} = \sum_{i=1,X_s^i \in \Omega_{X_o}}^{k_s} d(X_s^i, X_o)||X_s^i - X_o||_2 \tag{2}$$

and

$$E_{X_t,X_o} = \sum_{j=1,X_t^j \in \Omega_{X_o}}^{k_t} d(X_t^j, X_o)||X_t^j - X_o||_2 \tag{3}$$

where Ω_{X_o} is a set of the nearest neighbors of X_o, k_s and k_t are the number of source and target instances in Ω_{X_o}, d is a similar measurement of two instances and several measurements can be used here, e.g.,Euclidean distance, heat kernel method. We explore a measurement based on LCC which will introduce in next section.

On the other hand, X_o is assumed can be obtained by reconstructing from the common latent features Z, which can be formulated as follows:

$$X_o = W^{'}Z = W^{'}WX \tag{4}$$

where the parameter $W^{'}$ is used to reconstruct X_o from Z, W is used to learn the common latent feature Z from X, X denotes the data from either the source or target domain.

The general idea of domain adaptation is the involved domains share some common latent factors, these factors can be specific features or geometric structure. In our work, we additionally make the learned common feature Z preserve the local structure of X_o. In this way, the nearest neighbors of X_o and their weights can be estimated by the nearest neighbors of Z which can be formulated as follows:

$$d(X_s^i, X_o) = d(WX_s^i, WX) \tag{5}$$

and

$$d(X_t^j, X_o) = d(WX_t^j, WX) \tag{6}$$

Therefore, the problem of data reconstruction can be transformed into a common feature learning problem, which can be formulated as following objective:

$$W^* = \min_{W,W'} E_{X_s,W'WX} + E_{X_t,W'WX} \tag{7}$$

where

$$E_{X_s,W'WX} = \sum_{i=1,WX_s^i \in \Omega_{(WX)}}^{k_s} d(WX_s^i, WX)||X_s^i - W^{'}WX||_2 \tag{8}$$

and

$$E_{X_t,W'WX} = \sum_{j=1, WX_t^j \in \Omega_{(WX)}}^{k_t} d(WX_t^j, WX)||X_t^j - W'WX||_2 \tag{9}$$

Moreover, the selected features may be different in the case of DA and the problem has been discussed in [16]. We follow their solution to deploy $l_{2,1} - norm$ on project matrix W, which leads to select common features shared by the domains. As a result, we formulate our objective function as follows:

$$W^* = \min_{W,W'} E_{X_s,W'WX} + E_{X_t,W'WX} + ||W||_{2,1} \tag{10}$$

C. Problem Optimization and Implementation

Since GAE has superior ability of non-linear representation and local structure exploration, we propose a variant of it to optimize the Eq.(10). In the framework of GAE, the parameter W is corresponding to non-linear encode function $F(\bullet)$ and W' is corresponding to non-linear decode function $G(\bullet)$. As can be seen from [13], through defining different reconstruction sets and weights the GAE has various variants, i.e., LDA, PCA, LLE. However, these relational principle are not suitable for DA, due to the insufficient labeled target data or the divergence of involved domains. Therefore, a relational function based on LCC is proposed in our work to find and weight the meaningful neighbors across domains.

An overview of the proposed algorithm is shown in Figure. 1. The neighbors and their weights are determined in the LCC neighbor selection procedure. Firstly, the hidden layer or convolutional layer is used to map data into a feature space. Then the heat kernel method is used to get their similarity metric for each pair of data which can be formulated as follows:

$$d(X^i, X^j) = \begin{cases} exp(-\frac{||L(WX^i) - L(WX^j)||_2}{2\sigma^2}) & , X^i \in \Omega_{X^j} \\ 0 & , otherwise \end{cases} \tag{11}$$

where the $L(\bullet)$ represent the LCC code and σ is a empirical coefficient. After this procedure, the GAE is employed to minimize the regularization reconstruction error, which described in Eq.(10).

The LCC is proposed by Yu [17], [18], its principle is to encourage the code to be local by minimizing the following objective function:

$$\min_{\gamma,C} \sum_x \frac{1}{2}||x - \gamma(x)||^2 + \mu \sum_{v \in C} |\gamma_v(x)|||v - x||^2 \tag{12}$$

where $\gamma(x) = \sum_{v \in C} \gamma_v(x)v$, C is a set of anchor points, v is the anchor point, $v \in C$, $\gamma(x)$ is the physical approximation of x. The first term in Eq. (12) measures the similarity between x and its physical approximation $\gamma(x)$, the second term measures the locality of coding.

Through the LCC-based relational function, we assume the cross-domain neighbors are pairs that not only can be reconstructed by similar anchor points but also have a relatively small distance. Compared to the original LE [24]-based approach, the distance of two instances is measured by their LCC codes instead of directly by their latent features. In this way, two far away instances e.g. the same label samples in different domain. If they can be reconstructed by similar anchor points, then they have the chance to be selected as neighbors. However it is impossible if their relationship is only measured by their feature, e.g. original LE-based measurement. The specific detail of neighbor selection procedure is described as follows:

Algorithm 1 LCC-based neighbor selection procedure

Input: $\{z^i\}_1^n$, the hidden representation;
Output: Ω_{x^i}, the reconstruction set for x^i;
S_i, the set of reconstruction weight for x^i;
Parameters: C, the number of anchor points;
W, the parameter of the encoder;
Notation: k_s, the number of neighbors from the source domain;
k_t, the number of neighbors from the target domain;

1. Computer the LCC code from hidden representation $\{z^i\}_1^n$ with parameter C;
2. Select the nearest k_s neighbors from source domain and k_t neighbors from target domain for $\{z^i\}_1^n$ and construct the reconstruction set Ω_{x^i};
3. Computer the weight for each pair through Eq.(11) , construct the reconstruction weight set S_i;

As the neighbors are selected and their weights are determined, our problem can be re-described as to reconstruct both its source and target neighbors from the data either from the source or target domain. Then the GAE can be used to minimize the Eq.(10) using following Algorithm:

Algorithm 2 Iterative learning procedure for Generalized Auto-encoder

Input: $\{x^i\}_1^n$, the training set;
Parameters: W, the parameter of the encoder;
W', the parameter of the decoder;
Notation: Ω_{x^i}, the reconstruction set for x^i;
S_i, the set of reconstruction weight for x^i;
$\{z^i\}_1^n$, the hidden representation;

1. Computer the reconstruction set Ω_{x^i} and the reconstruction weight set S_i from $\{z^i\}_1^n$ through Algorithm 1 ;
2. Minimize the objective in Eq.(10) using the stochastic gradient to update W and W' for t step ;
3. Computer the $\{z^i\}_1^n$ and update Ω_{x^i} and S_i through Algorithm 1;
4. Repeat step 2 and step 3 until convergence;

D. Deep Structure Construction

In order to enhance the non-linear power of model, we stack the LCC-based GAE into a deep architecture. It contains a multilayer encoder network to transform the high-dimensional input data to a low-dimension feature and a multiple decoder network to reconstruct the feature to the uncorrupted data. In order to obtain good initial weights of network, the layer-wise pre-training procedure [25] is introduced.

We apply the deep LCC-based GAE to domain adaptation by first learning features in an unsupervised fashion on the union of the source and target data sets. After all the parameters are learned, a linear Support Vector Machine (SVM) [26] is then trained on the deep feature of labeled source data and tested on the target data.

IV. Experiments

In this section, we first present the benchmarks and experimental sitting. Then comparison results are presented followed by some properly analysis.

A. Datasets and Experimental Setting

MNIST, USPS, COIL20, Synthetic Signs, GTSRB are 5 image benchmark datasets widely adopted, the examples of images from the 5 datasets are demonstrated as Figure.2 and their specific details are shown in Tab. I.

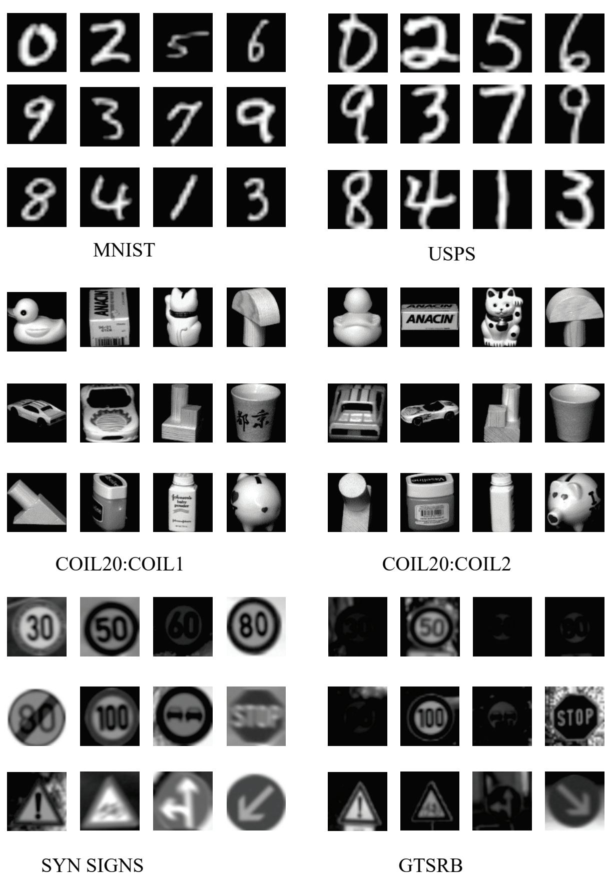


Fig. 2. Image Samples from MNIST, USPS, COIL20 object,respectively

USPS + MNIST share ten common digital categories from two datasets: (1) USPS dataset consist of 7,291 training images and 2,007 test images; (2) MNIST dataset has a training set of 60,000 images and a testing set of 10,000 images. All the images are rescaled to 16x16, and the raw pixel values are adopted.

COIL20 object dataset includes 20 objects with 1440 images. Each object has 72 images and all the images are 32x32. In the experiments, we follow the setting of [27] to split the dataset into 2 subsets: COIL1 and COIL2. COIL1 includes samples of $[0°, 85°] \cup [180°, 265°]$ and COIL2 includes samples of $[90°, 175°] \cup [270°, 355°]$. We select one subset as one domain, then build two transfer learning tasks.

Synthetic Signs + GTSRB share 43 common traffic signs from two datasets: (1) Synthetic Signs was provided by the authors of [28] and consists of 100,000 images that were generated by taking street signs from Wikipedia and applying various artificial transformations. (2) GTSRB provides 39,209 (training set) and 12,630 (test set) cropped images of German traffic signs. The images vary in size and were resized with bilinear interpolation to match the Synthetic Signs images size of 40x40 pixels. We follow the experimental setting described in [29] to use the Synthetic Signs dataset for the source dataset and the GTSRB dataset for the target dataset.

TABLE I
Statistics of the 5 benchmark datasets

Dataset	Type	#Example	#Feature	#Class
COIL20	Object	1,440	1,024	20
USPS	Digit	1800	256	10
MNIST	Digit	2,000	256	10
Synthetic Signs	Traffic Signs	100,000	1600	43
GTSRB	Traffic Signs	51,839	1600	43

Note that the arrow ”⟶” is the direction from source to target. For example, ”USPS⟶MNIST” means USPS is the source domain and MNIST is the target domain.

B. Comparison Methods and Implementation Details

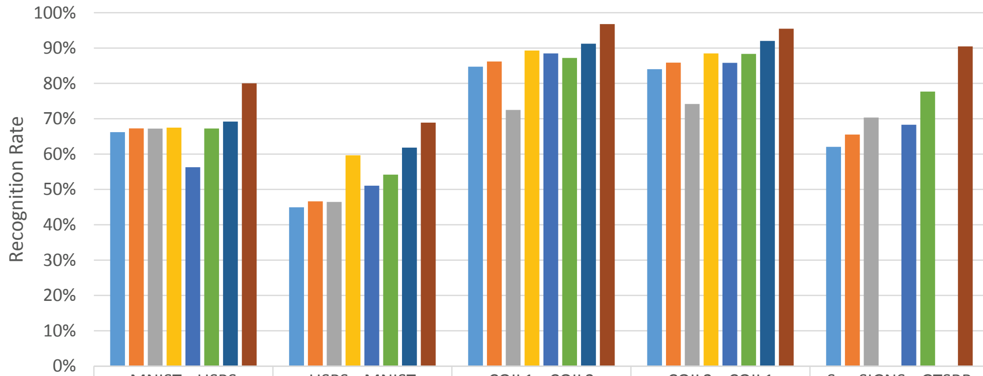
We mainly compare with 7 traditional methods to show the effectiveness of our algorithm as follows: Principle Component Analysis (PCA), Information-theoretic Metric Learning (ITML), Geodesic Flow Kernel (GFK), Joint Domain Adaptation (JDA), Transfer Component Analysis (TCA), Marginalized Donosing Auto-encoder (mSDA) and Robust Transfer Metric Learning (RTML).

The first two are the traditional metric learning algorithms, in which we train the metric on labeled source data while reuse for target learning. The last six are the transfer learning algorithms, we train on labeled source and unlabeled target to transfer knowledge during the model training. In transfer learning, it is hard to tune the optimal parameters through cross validation. Therefore, we empirically search the optimal parameter, and report each method's best results.

C. Experimental Results

In this section, we present the comparison results on different datasets, to show the effectiveness of our proposed algorithm. For these datasets, each one has two subsets, so we select one as the source domain and the other as the target domain, then we switch them. In all, we have two cases for each database, and the comparison results of 8 algorithms are shown in Figure. 3.

PCA and ITML are two traditional metric learning algorithms, they simply reuse the metric trained on source data to target data. Due to the discrepancy of involved domains, it results in relatively poor performance. GFK employs a



	MNIST->USPS	USPS->MNIST	COIL1->COIL2	COIL2->COIL1	Syn SIGNS->GTSRB
PCA	66.22%	44.95%	84.72%	84.03%	62.06%
ITML	67.26%	46.63%	86.20%	85.86%	65.54%
GFK	67.22%	46.45%	72.50%	74.17%	70.34%
JDA	67.48%	59.65%	89.31%	88.47%	-
TCA	56.28%	51.05%	88.47%	85.83%	68.30%
mSDA	67.24%	54.20%	87.20%	88.34%	77.68%
RTML	69.20%	61.82%	91.23%	92.03%	-
OURS	80.02%	68.89%	96.77%	95.47%	90.47%

Fig. 3. Comparison of recognition performance on benchmarks datasets

kernel metric to minimize the divergence of source and target domain. It performs better than PCA but worse than ITML on MNIST + USPS datasets, which may results from the discriminative information used in ITML is more crucial. TCA is a unified framework to learn a projection by matching feature representation. In contrast, JDA is a unified dimensionality reduction algorithm to mitigate both the marginal and the conditional distribution gap. In all these 'shallow' methods, JDA and RTML can achieve better results in most cases. This may be caused by they take advantage of the pseudo labels of target data. mSDA adopts denoising strategy, however, it only reconstructs the data from its corrupted version in the same domain. Hence, the domain shift can not be well mitigated. From the results, we can notice that the our proposed algorithm outperform others, which shows the effectiveness of the proposed method in knowledge transfer.

V. Conclusion

Due to a wide variability in Traffic Signs' visual appearance, Transfer Learning is a important ability in Traffic Sign Recognition algorithms. In this paper, we assume both the source and target data are corrupted versions of a domain-invariant domain. Under the assumption that the latent subspace is able to reconstruct the domain-invariant data as well as to preserve the local manifold of it, we construct a general objective function for domain adaptation. Then, a variant of GAE is proposed to optimize this objective function. The LCC-based GAE consists of multiple non-linear convolutional layers that can find more meaningful cross-domain neighbors as well as preserve statistical and geometric property simultaneously. Experimental results on several benchmark datasets demonstrate the effectiveness of our proposed approach in comparison with several traditional methods. In our further work, we will research on the combination of the global and local strategy and use it to find more meaningful neighbors across domains.

Acknowledgment

This work was partially supported by China Scholarship Council under Grant No. 201706230171 .

References

[1] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, no. 0, pp. –, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608012000457

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40537-016-0043-6

[4] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1410–1417.

[5] Z. Ding, M. Shao, and Y. Fu, "Deep low-rank coding for transfer learning." in *IJCAI*, 2015, pp. 3453–3459.

[6] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1667–1680, 2012.

[7] R. Chattopadhyay, Q. Sun, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Multisource domain adaptation and its application to early detection of fatigue," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, p. 18, 2012.

[8] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," *arXiv preprint arXiv:1206.4683*, 2012.

[9] Z. Ding, M. Shao, and Y. Fu, "Deep robust encoder through locality preserving low-rank dictionary," in *European Conference on Computer Vision*, 2016, pp. 567–582.

[10] F. Zhuang, X. Cheng, P. Luo, S. J. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders." in *IJCAI*, 2015, pp. 4119–4125.

[11] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa, "Dash-n: Joint hierarchical domain adaptation and feature learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5479–5491, Dec 2015.

[12] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.

[13] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized autoencoder: A neural network framework for dimensionality reduction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 490–497.

[14] M. Kan, S. Shan, and X. Chen, "Bi-shifting auto-encoder for unsupervised domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3846–3854.

[15] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2066–2073.

[16] J. Li, J. Zhao, and K. Lu, "Joint feature selection and structure preservation for domain adaptation," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 1697–1703.

[17] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223–2231.

[18] K. Yu and T. Zhang, "Improved local coordinate coding using local tangents," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 1215–1222.

[19] Y. Lin, T. Zhang, S. Zhu, and K. Yu, "Deep coding network," in *Advances in Neural Information Processing Systems*, 2010, pp. 1405–1413.

[20] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[21] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[22] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley, New York, 1973.

[24] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[27] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[28] B. Moiseev, A. Konev, A. Chigorin, and A. Konushin, "Evaluation of traffic sign recognition methods trained on synthetically generated data," in *15th International Conference on Advanced Concepts for Intelligent Vision Systems - Volume 8192*, ser. ACIVS 2013. New York, NY, USA: Springer-Verlag New York, Inc., 2013, pp. 576–583.

[29] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds. Lille, France: PMLR, 07–09 Jul 2015, vol. 37, pp. 1180–1189. [Online]. Available: http://proceedings.mlr.press/v37/ganin15.html