C2DNDA : A Deep Framework for Nonlinear Dimensionality Reduction

Qi Wang, *Member, IEEE,* Zequn Qin, Feiping Nie, and Xuelong Li*, *Fellow, IEEE*

***Abstract*—Dimensionality reduction has attracted many research interest in the past decades. Existing dimensionality reduction methods like LDA and PCA have achieved promising performance, but the single and linear projection properties limit the further improvements of performance. A novel convolutional two-dimensional nonlinear discriminant analysis (C2DNDA) method is proposed for dimensionality reduction in this study. In order to handle nonlinear data properly, we present a newly designed structure with Convolutional Neural Networks (CNN) to realize an equivalent objective function with classical 2DLDA and thus embed the original 2DLDA into an end-to-end network. In this way, the proposed dimensionality reduction network can utilize the nonlinearity of the CNN and benefit from the learning ability. The results of experiment on different image related applications demonstrate that our method outperforms other comparable approaches, and its effectiveness is proved.**

***Index Terms*—2DLDA, convolutional neural networks, classification, dimensionality reduction**

I. Introduction

LINEAR discriminant analysis (LDA) is a classical method for dimensionality reduction and classification, which is commonly used in pattern recognition and machine learning fields, and shows satisfactory results in face recognition [1]–[4]. The classical LDA aims to find the optimal projection vectors by maximizing the trace of between-class scatter matrices and minimizing the trace of within-class scatter matrices.

The idea of maximizing the ratio of trace is effective and intuitive. However, the effectiveness comes at a cost. The data format required for classical LDA must be the vector formation. This kind of constraint limits the application with existing data. The image is hard to process with the vector formation. In order to solve this problem, some methods utilize matrix-vector transformation [5]–[7] and combine it with classical LDA. However, this kind of transformation costs more computational resources. Another shortcoming of these methods is that the useful information hidden in the spatial relationship of image is wiped out. In fact, the reason why convolutional neural network gains huge success can be regarded as taking use of spatial relationship of the image data. This kind of spatial relation can be extended in some other machine learning fields such as nature language process [8], [9]. Another way to address this problem is to process 2D data directly without matrix-vector transformation [10], [11]. Typically, the two-dimensional linear discriminant analysis (2DLDA) [10], [12], [13] is the commonly used method. 2DLDA can utilize more related spatial information by using the original data without transformation, which leads to a better performance [14].

With the development of the machine learning, many classical approaches are showing its power once again when combined with deep learning. Recently, many methods derived from classical PCA [15], [16] and LDA [17] achieve better performance compared with its original version. The main idea of these approaches is to embed the classical problem into the deep neural networks. In this way, these deep version methods can utilize the nonlinear representation ability of networks and the more effective optimization methods, which are stochastic gradient descent (SGD) and many derivations. But the combination with the classical LDA and CNN always means complex neural network structure and difficult optimization.

In this paper, a convolutional two-dimensional nonlinear discriminant analysis (C2DNDA) method is proposed which aims to tackle these problems for complex nonlinear dimensionality reduction. In [17], the solution of deep LDA is derived from eigenvalue based optimization. In this paper, however, we use a specially designed CNN structure to optimize the classical LDA objective functions instead of maximizing the eigenvalues of scatter matrix. The main novelty of the proposed method is to utilize the novel CNN structure which makes the optimization of the classical LDA easier and gains better performance. Meanwhile, the whole networks can be seen as a nonlinear 2D dimensionality reduction framework that solves dimensionality reduction and classification tasks simultaneously. This paper is an extension of our previous work [18]. The major differences of this paper can be summarized in four parts. First, the effectiveness of F-loss function is analyzed and examined, which is the main novelty of our work. Second, the effects of different types of labels are analyzed. Both of these two factors form the theoretical analysis of the effectiveness of this paper. Third, the converge performance is compared with different experimental settings to prove the proposed algorithm. At last, the background information, the analysis of related work and the demonstration of our work are more comprehensive and clear.

The rest of the paper is organized as follows. In Section II, we review several related works and its merits and demerits.

This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, 61871470, and 61761130079, and Project of Special Zone for National Defense Science and Technology Innovation.

Q. Wang, Z. Qin, F. Nie and X. Li are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@nwpu.edu.cn; qinzequn@mail.nwpu.edu.cn; feipingnie@gmail.com; li@nwpu.edu.cn).

*Corresponding Author

Section III presents the C2DNDA in detail. The experiments on three datasets including traditional and deep learning based methods are shown in Section IV. Section V summarizes this work.

II. Related Work

In the past few years, dimensionality reduction has attracted a large number of researchers because of increasing demands on various data applications in the machine learning fields. There are two kinds of dimensionality reduction approaches which are the supervised and the unsupervised methods. Generally, the supervised methods gain a better performance. The LDA class methods are the representative supervised approaches. In this section we briefly review some LDA-related methods about this topic.

A. Classical LDA

The linear discriminant analysis tries to transform the input data into a low dimensional subspace and separate the data in the transformed subspace [19]. Two kinds of scatter matrices of transformed data are employed to measure the effectiveness of separation in the lower-dimensional subspace. Denote the input as $X \in R^{l\times n}$, which belongs to c classes $\pi = [\pi_1, \pi_2, \ldots, \pi_c]$. Suppose the projection is $W \in R^{l\times c}$. Then the transformation defined by the classical LDA is $y_i = W^T x_i$, where $x_i \in R^{l\times 1}$ is the input. In order to find the optimal transformation, we use the between-class S_b and within-class scatter matrix S_w which are defined as:

$$S_b = \sum_{i=1}^{c} n_i (M_i - M)(M_i - M)^T, \tag{1}$$

$$S_w = \sum_{i=1}^{c} \sum_{X_j \in \pi_i} (X_j - M_i)(X_j - M_i)^T, \tag{2}$$

where n_i is the whole number of samples in the $i-th$ class π_i, N is the number of input data, $M_i = \frac{1}{n_i}\sum_{X_j\in\pi_i} X_j$ is the mean value of $i-th$ class π_i and $M = \frac{1}{N}\sum_{i=1}^{c}\sum_{X_j\in\pi_i} X_j$ is the mean value of the entire input data.

In this way, the transformed class scatter matrices can be formulated as:

$$\widetilde{S_b} = W^T S_b W, \tag{3}$$

$$\widetilde{S_w} = W^T S_w W. \tag{4}$$

According to the above equations, the optimal solution to W can be:

$$\max_W \frac{\left\|\widetilde{S_b}\right\|}{\left\|\widetilde{S_w}\right\|}. \tag{5}$$

In [20], the objective function can be formulated as the ratio of trace:

$$\max_W Tr((\widetilde{S_w})^{-1}\widetilde{S_b}), \tag{6}$$

where $Tr(\cdot)$ denotes the trace of the matrix.

B. 2DLDA

Because the 2DLDA method can handle the 2D-data directly, data format is the major difference between the classical LDA and 2DLDA. In order to make more use of spatial information and handle 2D-data, the 2DLDA method employs several projection matrices. Typically, We use U, V as the projection matrices in 2DLDA and $X = [X_1, X_2, \ldots, X_n]$ as the input data, where $X_i \in R^{m\times n}$. As a result, the two projected between-class and within-class scatter matrices are:

$$\widetilde{S_b} = \sum_{i=1}^{c} n_i U^T (M_i - M) V V^T (M_i - M)^T U, \tag{7}$$

$$\widetilde{S_w} = \sum_{i=1}^{c} \sum_{X_j\in\pi_i} U^T (X_j - M_i) V V^T (X_j - M_i)^T U. \tag{8}$$

Because the scatter matrices have the same meaning as the classical LDA, the similar objective function can be calculated. What makes the differences between the LDA and 2DLDA is the number of the projection matrices. In this way, the objective target is changed to two projection matrices:

$$\max_{U,V} \frac{\left\|\widetilde{S_b}\right\|}{\left\|\widetilde{S_w}\right\|}. \tag{9}$$

As mentioned in [20], the objective function is defined as follows:

$$\max_{U,V} Tr((\widetilde{S_w})^{-1}\widetilde{S_b}). \tag{10}$$

C. Regularized LDA

During the realization of the classical LDA problem, the calculation of the inverse matrix is important. However, the matrix $\widetilde{S_w}$ doesn't have the inverse counterpart when the dataset is small. This property makes the dimensionality reduction unable to optimize. In order to avoid meaningless inverse matrix, regularization terms are applied to the classical LDA [21], [22]. Besides, these regularization terms can overcome the over-fitting problem.

In [23], a method called MVLDA is proposed. The estimations of scatter matrices are used for the regularization terms. An iterative method is used to obtain the estimation of the within-class scatter matrix.

$$S_{WL} = \frac{1}{Nn}\sum_{i=1}^{c}\sum_{X_j\in\pi_i} (X_j - M_i) S_{WR}^{-1} (X_j - M_i)^T, \tag{11}$$

$$S_{WR} = \frac{1}{Nm}\sum_{i=1}^{c}\sum_{X_j\in\pi_i} (X_j - M_i)^T S_{WL}^{-1} (X_j - M_i). \tag{12}$$

The estimations of between-class scatter matrix are defined as:

$$S_{BL} = \sum_{i=1}^{c} n_i (M_i - M)(M_i - M)^T, \tag{13}$$

$$S_{BR} = \frac{1}{Tr(S_{BL})}\sum_{i=1}^{c} n_i (M_i - M)^T (M_i - M). \tag{14}$$

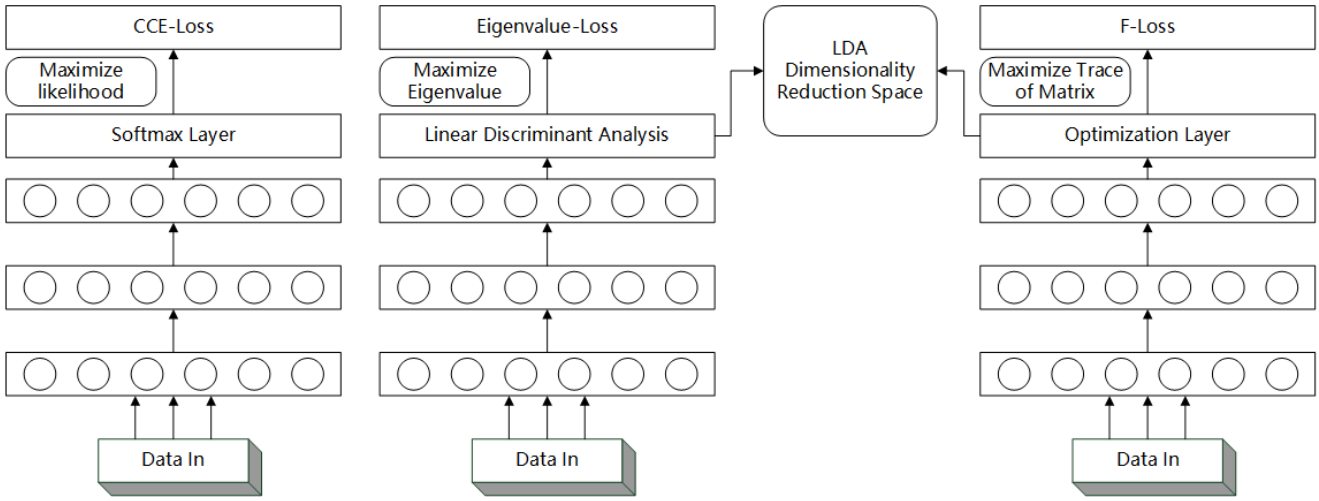


Fig. 1. Different formulations of loss function. The CCE loss can be regarded as maximization of classification accuracy by minimizing entropy. The DeepLDA method maximize the eigenvalues and our method maximize the trace of scatter matrix. Both methods convert the data into a low-dimensional subspace.

According to these estimations, the iterative regularization terms can be defined as:

$$S_w^r = (1-\lambda_w)S_w + \lambda_w S_w^s, \tag{15}$$

$$S_b^r = (1-\lambda_b)S_b + \lambda_b S_b^s, \tag{16}$$

where $S_w^s = S_{WR} \otimes S_{WL}$ and $S_b^s = S_{BR} \otimes S_{BL}$.

Different from the above approaches, there is one simple regularized method treat the identity matrix as the regularization term to avoid singularity. The original optimization of the classical LDA can be written in another unified way:

$$\max Tr((\widetilde{S_w})^{-1}\widetilde{S_b}). \tag{17}$$

Adding an identity matrix to the within-class scatter matrix, the term $\widetilde{S_w} + \lambda I$ is nonsingular matrix. Thus, the regularized version can be written as:

$$\max Tr((\widetilde{S_w} + \lambda I)^{-1}\widetilde{S_b}). \tag{18}$$

D. Deep LDA

In [24], a deep version of the Canonical Correlation Analysis (DCCA) method is proposed in order to handle acoustic and articulatory speech data. In [17], another version of deep linear discriminant analysis is proposed.

The basic idea of deep linear discriminant analysis is utilizing deep neural networks to solve the classical regularized LDA eigenvalue problem, which is formulated as:

$$S_b e_i = v_i(S_w + \lambda I)e_i, \tag{19}$$

where $e = [e_1, ..., e_{c-1}]$ are the resulting eigenvectors and $v = [v_1, ..., v_{c-1}]$ the corresponding eigenvalues.

The maximization of eigenvalues $v = v_1, ..., v_{c-1}$ equals to the maximization of original trace ratio between scatter matrices. In this way, two kinds of objective functions can be written as:

$$\max_\theta \frac{1}{c-1}\sum_{i=1}^{c-1} v_i \tag{20}$$

and

$$\max_\theta \frac{1}{k}\sum_{i=1}^{k} v_i,$$

$$\{v_1, ..., v_k\} = \{v_j | v_j < \min\{v_1, ..., v_{c-1}\} + \epsilon\}, \tag{21}$$

in which the θ represents the model parameters of network.

The original Categorical Cross Entropy (CCE) is replaced with the eigenvalue loss, which is another form of the classical 2DLDA.

The reason why deep linear discriminant analysis method has two similar objective functions is that optimization of such eigenvalue based method can be difficult. Further more, the back propagation of summation of eigenvalues is complex. Different from deep linear discriminant analysis method, we intend to convert such optimization problem into an end-to-end network which can be easily constructed and trained. The comparison of our method, DeepLDA and CCE can be seen in Figure 1.

III. Convolutional 2DLDA

In this section, we demonstrate the detailed proof of C2DNDA and its corresponding CNN construction. The main idea of the proposed method is to find an embedded structure which has the equivalent optimal solution to the LDA objective function.

A. LDA Based on Nonlinear Projection

Utilizing nonlinear projection to perform dimensionality reduction is an obvious solution to the above problems. Our target is to find a nonlinear projection which is able to perform dimensionality reduction. We denotes $g(\cdot)$ as the projection and $D = [D_1, D_2, \ldots, D_n]$ as the input. When the projected data $d_i = g(D_i)$ meets the the classical LDA assumption, we can treat this nonlinear method as a variant of the classical LDA.

It is obvious that nonlinear projection can obtain better representative information compared with the classical LDA or 2DLDA. However, the most difficult part of nonlinear LDA is the effective optimization method. In order to tackle above problems, a specially designed network structure is utilized to optimize the nonlinear projection. The proposed optimization method is an important novelty of this paper.

B. LDA in a Network

As mentioned in the above section, using CNN to realize LDA-like nonlinear projection is our goal. Suppose the transformation of CNN is $x = f(A) \in R^{m\times 1}$, in which x is the results of dimensionality reduction, $f(\cdot)$ represents the networks and A denotes the input. Denote the output of the networks as $X = [f(A_1), f(A_2), \ldots, f(A_n)] \in R^{m\times n}$ and label as $Y \in R^{n\times c}$. In order to prevent the singularity and over-fitting problem, we use the regularized LDA as our target objective function. The objective function of the regularized LDA is defined as:

$$\max_f Tr((S_t + \lambda I)^{-1} S_b), \tag{22}$$

where

$$S_t = XHX^T, \tag{23}$$

$$S_b = XHY(Y^TY)^{-1}Y^THX^T, \tag{24}$$

$$H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T. \tag{25}$$

It is tough to optimize the regularized LDA objective function because of the complicated structure hiding in the CNN projection $f(\cdot)$. Here, we propose another way based on frobenius norm to optimize this function:

$$\max_f Tr((S_t + \lambda I)^{-1} S_b)$$

$$\Leftrightarrow \min_{f,W,b} \left\| X^TW + \mathbf{1}b^T - \hat{Y} \right\|_F^2 + \lambda \|W\|_F^2, \tag{26}$$

where

$$\hat{Y} = Y(Y^TY)^{-\frac{1}{2}}. \tag{27}$$

proof: We use Lagrange multiplier method to prove Eq.26. If the optimized g in Eq.26 has the same solution with Eq.22, we would be able to optimize the regularized LDA with Eq.26 in a network. Following Lagrange multiplier method, setting the derivative of Eq.26 with respect to b to zero, we have:

$$b = \frac{1}{n}(\hat{Y}^T\mathbf{1} - W^TX\mathbf{1}). \tag{28}$$

Then substitute the above b into Eq.26:

$$\min_{f,W,b} \left\| X^TW + \mathbf{1}b^T - \hat{Y} \right\|_F^2 + \lambda \|W\|_F^2$$

$$\Leftrightarrow \min_{f,W} \left\| HX^TW - H\hat{Y} \right\|_F^2 + \lambda \|W\|_F^2. \tag{29}$$

Setting the derivative of Eq.III-B with respect to W to zero:

$$W = (XHX^T + \lambda I)^{-1} XH\hat{Y}. \tag{30}$$

Substituting W into Eq.III-B:

$$\min_{f,W} \left\| HX^TW - H\hat{Y} \right\|_F^2 + \lambda \|W\|_F^2$$

$$\Leftrightarrow \min_f Tr(\hat{Y}^THY) - Tr(\hat{Y}^THX^T(XHX^T + \lambda I)^{-1}XH\hat{Y}) \tag{31}$$

$$\Leftrightarrow \max_f Tr((S_t + \lambda I)^{-1} S_b).$$

C. Singularity of Normalized Label Matrix

The normalized label used in the above proof could not be constantly invertible. In this section, we discuss the singularity of normalized label matrix.

proof: Suppose $Y \in R^{n\times c}$ is the one-hot label of n samples. c denotes the class number. Then $B = Y^TY$ is the number matrix of each class. To be more specific, B is a diagonal matrix and the i-th diagonal element B_{ii} is the sample number of i-th class.

It is easy to know that $B_{ij} = Y_{i\cdot}^T Y_{\cdot j}$. Suppose n_i denotes the sample number of i-th class. When $i = j$,

$$B_{ij} = \sum_{k=1}^{n} Y_{ki} \times Y_{kj} = n_i. \tag{32}$$

If $B_{ij} \neq 0$ and $i \neq j$, then there will be a k^* which makes $Y_{k^*i} \times Y_{k^*j} = 1$. This property breaks the one-hot constraint. Thus the number matrix B is a diagonal matrix and contains the sample number of each class.

When we use the mini-batch stochastic gradient descent algorithm to train our network, lack of certain class in the mini-batch will make number matrix contains zero diagonal elements. This phenomenon usually happens with a small batch size. In this case, the number matrix will be singular and the normalized label will not be solvable.

D. Network Construction

What we can see from the above theorem is that adding an optimization layer in the networks $f(\cdot)$ can be used to optimize the parameters W and b simultaneously, in which the added layer inputs the dimensionality reduction representation X and outputs the $X^TW + \mathbf{1}b^T$ results. Besides, the optimal network projection $f(\cdot)$ in Eq.22 can be obtained because Eq.22 and Eq.26 have the same solution to $f(\cdot)$. In this way, the dimensionality reduction network is obtained by the newly added optimization layer. The newly added layer which is composed of $X^TW + \mathbf{1}b^T$ can be seen as a kind of evaluation metric of dimensionality reduction.

Because the optimization layer has no constraint to the backbone networks, we can utilize different kinds of network structure to test the effectiveness of our method. Considering LeNet-5 [25] which has 5 layers and ResNet [26] that contains more than 1000 layers, the depths of CNN networks are the essential parameters to the network performance. We choose two kinds of structures to construct our networks. The first one is a simplified dimensionality reduction network which has one fully connected (fc) layer and only two convolutional layers. The purpose of this network is to compare the proposed algorithm with the classical dimensionality reduction methods.

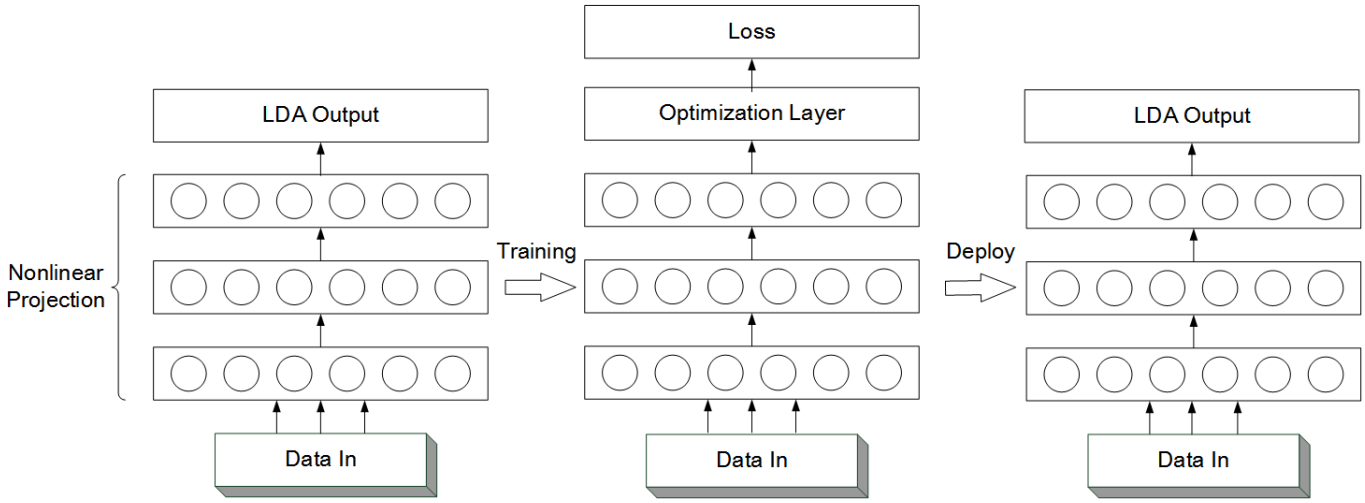


Fig. 2. Different strategy in training and deploy stage. Only in the training stage the optimization layer is added.

Another one is a complex full functional network which has the same structure with [17], including batch norm [27] and dropout [28] layers. Similarly, the purpose of this network is to compare the proposed algorithm with other deep methods.

As mentioned in Eq.26, the optimization layer is used for the evaluation of the dimensionality reduction results. The optimization layer converts the output results into the label space. In fact, such evaluation metric is the classification accuracy. It is noteworthy that the evaluation metric is derived from the dimensionality reduction task instead of classification. Meanwhile, the proposed optimization layer is in charge of the optimization. Figure 2 illustrates how the network works during the training and deploy stages. Only in the training stage, the optimization layer is added to achieve the same goal as the regularized LDA. After the network is converged, the dimensionality reduction part is free to the optimization layer.

In this way, we have proposed a method which combines the dimensionality reduction goal with the classification stage into an end-to-end network, which has the same optimization goal with the classical LDA. Because the proposed method transforms the input data directly, it is obvious that the proposed algorithm is a convolutional nonlinear 2DLDA method.

E. F-loss vs CCE

Different from other classification or regression networks, the proposed network can achieve dimensionality reduction goals. The reasons why such network structure can be used for dimensionality reduction is shown in the formatting of loss function. In order to achieve different goals, many different losses are proposed. In [29] [30], the smooth L1 loss is used for bounding box regression while traditional regression method is mean square loss. The definition of smooth L1 loss can be written as [31]:

$$L = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise. \end{cases} \tag{33}$$

In terms of classification, softmax function is the most widely used method [32]. The corresponding cost function is cross entropy function with softmax, which has many excellent properties, such as simple realization and fast convergence speed. The definition of cross entropy cost function is

$$L = -\frac{1}{N}\sum_{n=1}^{N}(y_n \log \hat{y_n} + (1 - y_n)\log(1 - \hat{y_n})). \tag{34}$$

The loss function we used in C2DNDA is formed with Frobenius norm, which can be written as:

$$L = \left\| X^T W + \mathbf{1}b^T - \hat{Y} \right\|_F^2 + \lambda \left\| W \right\|_F^2 \tag{35}$$

and

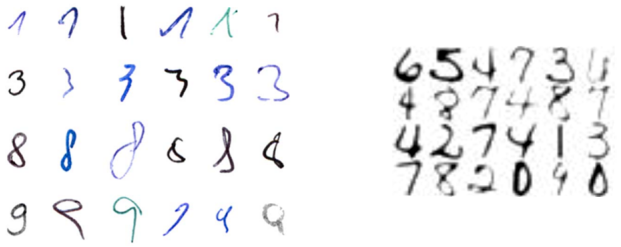
$$L = -Tr((S_t + \lambda I)^{-1} S_b). \tag{36}$$

From the formulation of Frobenius norm loss and cross entropy loss we can see that the CCE method tends to reduce the entropy of information difference between prediction and real label, while the F-loss method tends to reduce the difference in every position between the data after dimensionality reduction and normalized label. These two kinds of formulation reflect the key difference of classification and dimensionality reduction.

IV. Experiment

In this section, our method is named as C2DNDA. We compare the proposed C2DNDA with eight classical dimensionality reduction algorithms, including LDA [1], 2DPCA [11], 2DLDA [12], S2DLDA [33], P2DLDA [33], Tensor LPP (T-LPP) [34], [35], Bilinear SVM (B-SVM) [36] and CRP [37]. Due to the limited learning capacity of the classical methods, we choose two relatively small handwritten digit datasets to conduct our experiments. As for the deep learning algorithms, several representative dimensionality reduction networks are used for comparison, including NIN [38], Maxout [39], DeepCNet [40] and DeepLDA [17]. The detailed network structure and hyper parameters are shown in this section.

Another part of experiments are carried out to examine the effectiveness of proposed network structure. The experiments



(a) (b)

Fig. 3. Example handwritten images in CVL and USPS dataset.

are divided into two parts which are experiment about different label formats and losses. The experiment setups are identical to the classification experiments.

A. Experiment Setup and Datasets

Four datasets are used in our experiments.

MNIST: The MNIST dataset is a handwritten digits recognition which contains 60,000 examples. Typically, 50,000 examples are used for training and the rest is used for testing.

CIFAR-10: The CIFAR-10 dataset consists of 60000 32×32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

CVL: The CVL dataset [41] is generated for the ICDAR2013 Handwritten Digit Recognition Competition. Like the MNIST dataset, the image size is 32×32. The number of images in the CVL dataset is 21,780.

USPS: This dataset is a small handwritten dataset which contains 9,298 images with a size of 16×16.

As mentioned in previous sections, two networks using different number of convolutional layers are employed. Both of them have the same structure of optimization layer. In Table I we show the composition of the proposed network in detail. We use momentum [42] and Adam [43] SGD optimizer in TensorFlow [44] to train our network. In some rare cases, small batch size can lead to loss of some classes in a batch, which makes the normalized label $\hat{Y}$ can't be solved. In order to give enough labels to the network, the batch size is set to 100. The regularization item coefficient λ is set to 10^{-4} empirically.

B. Classification Accuracy

1) Comparison with Traditional Methods: In this experiment, we choose classification accuracy as our evaluation metric as described in the above section. The eight algorithms used for comparison are LDA, 2DLDA, 2DPCA, B-SVM, S2DLDA, P2DLDA, T-LPP and CRP. SVM and one nearest neighbor (1NN) are used in these eight methods. The proposed method uses the proposed layer to obtain classification results. Because the B-SVM is a classification method, so we conduct our experiment with it without extra classifier. Three kinds of settings of training sets are employed to test the effect of different training data sizes. Note that although different numbers of training data are utilized, the testing set is the same.

TABLE I
THE STRUCTURE COMPARISON OF THE PROPOSED C2DNDA NETWORKS, IN WHICH C REPRESENTS THE CHANNEL NUMBER, B REPRESENTS THE BATCH NORMALIZATION LAYER AND R REPRESENTS THE RELU LAYER.

stage	simplified	complex
conv1	3x3, 32C, R 2x2 Pooling	3x3, 64C, B, R 3x3,64C, B, R 2x2 Polling
conv2	3x3, 64C, R 2x2 Pooling	3x3, 96C, B, R 3x3, 96C, B, R 2x2 Polling
conv3		3x3, 256C, B, R 1x1, 256C, B, R 1x1, 64C, B, R
fc1	64	64, Dropout
classification	10	10

In Table II, we show the results of experiment with 80% training samples in every dataset. We can see that our method outperforms all of the other methods, which demonstrates that the proposed method benefits from the proposed network. As a result, the effectiveness of the proposed optimization approach can be verified. When compares with CVL dataset and SVM, the C2DNDA outperforms the classical 2DLDA by 8.9%.

The second experimental results which use only 20 samples are shown in Table III. No matter using which kind of classifier, the CRP gains the best performance in the CVL dataset. The proposed method achieves the second best results. When compares in the USPS dataset, our method still gets the best results. There is one thing we should notice is that the proposed network structure starts to perform poorly because of insufficient data.

The last experimental results which use only 10 samples are shown in Table IV. As expected, all of the methods show a relatively worse result with the decreasing number of samples. The proposed method gains the 3rd place in the experimental setting of CVL dataset and 1NN classifier. When compared with SVM, we can say that the proposed approach obtains good results. In the USPS dataset, the proposed method gains the 2nd place. When the sample number is extremely small, the proposed method still shows better results than the classical 2DLDA by 10.8%.

From these experiments we can see that our method could achieve satisfying results with sufficient data. When dealing with insufficient data like 20 samples, which is far from the requirement of deep networks, our method could still rank 1st and 2nd places on different datasets. In this way, the effectiveness of our method under various conditions can be proved.

2) Comparison with Deep Learning Methods: In this part of experiment, we still use the same evaluation metric to conduct our experiment with deep learning methods. There are four methods, including NIN, Maxout, DeepCNet, DeepLDA. The experimental setting of DeepLDA is "DeepLDA-60k" [17]. Both of our two kinds of structures are compared in this experiment. On the CIFAR-10 dataset, we keep the same

TABLE II
CLASSIFICATION ACCURACY OF EIGHT CLASSICAL ALGORITHMS AND THE PROPOSED SIMPLIFIED C2DNDA NETWORK ON TESTING DATA. 80 PERCENT OF THE TRAINING DATA ARE USED FOR TRAINING.

Dataset	2DLDA	LDA	P2DLDA	S2DLDA	2DPCA	B-SVM	CRP	T-LPP	C2DNDA (simplified)
CVL(SVM)	88.2±1.4	87.3±1.5	89.9±1.1	89.3±1.1	87.7±1.7	93.4±1.5	91.5±1.3	90.3±1.5	**96.6**
CVL(1NN)	91.1±1.4	90.2±1.6	92.8±1.3	92.4±1.5	90.6±1.5	93.4±1.5	94.7±1.3	93.1±1.3	**96.6**
USPS(SVM)	93.8±1.1	93.1±1.6	94.3±1.8	94.1±1.2	93.5±1.5	96.2±1.4	95.6±1.3	94.7±1.4	**97.9**
USPS(1NN)	95.6±1.2	94.8±1.3	94.5±1.2	96.4±1.4	95.2±1.1	96.2±1.4	96.8±0.9	95.1±1.1	**97.9**

TABLE III
CLASSIFICATION ACCURACY OF EIGHT CLASSICAL ALGORITHMS AND THE PROPOSED SIMPLIFIED C2DNDA NETWORK ON TESTING DATA. ONLY 20 SAMPLES OF THE TRAINING DATA ARE USED FOR TRAINING.

Dataset	2DLDA	LDA	P2DLDA	S2DLDA	2DPCA	B-SVM	CRP	T-LPP	C2DNDA (simplified)
CVL(SVM)	69.2±1.6	67.9±1.3	58.3±1.7	68.6±1.7	68.3±1.4	70.9±1.6	**79.2±1.5**	69.1±1.4	72.1
CVL(1NN)	66.7±1.2	63.7±1.3	60.2±1.7	67.3±1.3	64.1±1.8	70.9±1.6	**74.2±1.1**	65.2±1.9	72.1
USPS(SVM)	85.8±1.5	83.8±1.8	80.8±1.3	86.8±1.9	84.2±1.9	86.6±1.8	88.4±1.4	81.8±1.6	**90.2**
USPS(1NN)	84.5±1.8	83.1±1.3	74.7±1.5	85.6±1.2	83.6±1.4	86.6±1.8	89.2±1.4	79.5±1.4	**90.2**

TABLE IV
CLASSIFICATION ACCURACY OF EIGHT CLASSICAL ALGORITHMS AND THE PROPOSED SIMPLIFIED NETWORK ON TESTING DATA. ONLY 10 SAMPLES OF THE TRAINING DATA ARE USED FOR TRAINING.

Dataset	2DLDA	LDA	P2DLDA	S2DLDA	2DPCA	B-SVM	CRP	T-LPP	C2DNDA (simplified)
CVL(SVM)	57.9±1.9	56.8±1.5	48.4±1.3	51.9±1.4	57.1±1.6	64.2±1.9	**68.3±1.5**	66.9±1.6	55.2
CVL(1NN)	50.3±1.8	47.3±1.4	46.9±1.4	51.9±1.5	47.9±1.5	64.2±1.9	**67.3±1.3**	55.1±1.6	55.2
USPS(SVM)	77.3±1.3	73.4±1.7	73.9±1.4	79.2±1.8	74.4±1.9	79.4±1.9	**84.3±1.3**	78.2±1.5	82.6
USPS(1NN)	71.8±1.4	68.5±1.7	64.8±1.9	71.2±1.6	69.2±1.5	79.4±1.9	**84.4±1.5**	76.7±1.6	82.6

network as DeepLDA [17]. So, the original versions of our method are not tested.

From Table V, we can see that the proposed method gains a satisfactory result. Our method gains better results on both datasets when compared with DeepLDA. Although the simplified version shows a relatively bad performance compared with other deep methods, we can say that our method performs not that bad in consideration of its simple structure. The reason is that the simplified version has only two convolutional layers, while other methods have much more convolutional layers.

What we can see from these experiments is that our method shows fairly good performance with enough data. The proposed algorithm gains better performance compared with the classical 2DLDA in every experimental setting. When training with insufficient samples, our method gains a satisfactory result. Taking into account all these experiments, we can see that the proposed method is effective.

TABLE V
CLASSIFICATION ACCURACY OF DEEP METHODS.C2DNDA(DEEPLDA) KEEPS THE SAME STRUCTURE AS DEEPLDA USED IN CIFAR-10.

Method	MNIST	CIFAR-10
C2DNDA(simplified)	99.20	-
Maxout	99.55	90.62
DeepLDA	99.68	92.71
NIN	99.53	89.59
DeepLDACCE	99.66	92.81
DeepCNet	**99.69**	**93.72**
C2DNDA(complex)	**99.69**	92.60
C2DNDA(DeepLDA)	-	92.88

C. The Effect of Normalized Label

There are two kinds of label format we can use in dimensionality reduction network. One kind of label is one-hot encoding label. Another one is normalized label. According to Eq. 26, the final optimization layer projects the original data into a normalized one-hot label $\hat{Y}$ instead of normal one-hot label, which has the formation of:

$$\hat{Y} = Y(Y^T Y)^{-\frac{1}{2}}. \tag{37}$$

Such kind of normalized label leads to two problems.

The first one is that the number matrix $Y^T Y$ might be not invertible when a data batch is small. Another problem is that the normalized label might affects the classification accuracy. The normalized label has the same form as one-hot label but it is weighted by the sample number of its class. The detailed proof of singularity of the number matrix $Y^T Y$ is presented in section III-C.

In order to compare two kinds of label, we use the simplified network structure and MNIST dataset to test our method. All of the other hyper parameters are identical.

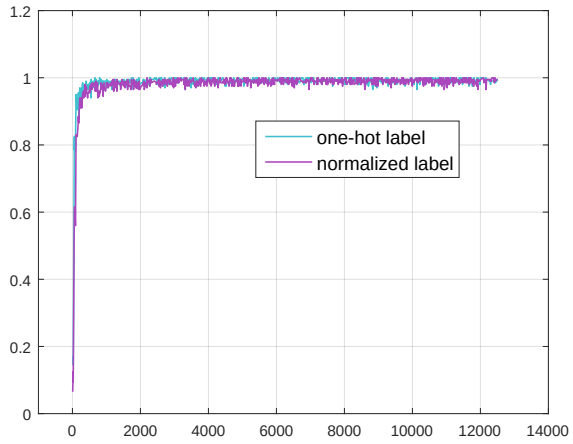


Fig. 4. Training accuracy of one-hot label and normalized label.

The training accuracy is shown in Figure 4. It is difficult to say which one is better. For both labels, two kinds of training accuracy are very close to 100%. The training loss is shown in Figure 5. The training loss of one-hot label is much lower than normalized one. Classification accuracy in test set of one-hot label form is 99.37%, while the accuracy of normalized form is 99.20%. The reason why one-hot label outperforms normalized one is that both of them are one-hot format while the normalized is smaller. This property can be a drawback when training a network because the difference between predict and label is smaller. In this case, we use one-hot label in our experiment. In fact, many LDA methods have such property that the optimal solution to projection matrices are not unique. If W is the optimal projection matrix, then WQ would be the optimal projection matrix, in which Q is a

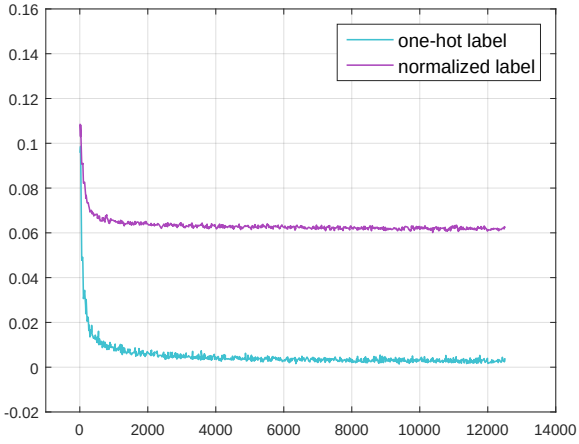


Fig. 5. Training loss of one-hot label and normalized label.

diagonal matrix. Without consideration of regularization item and suppose $Q = (Y^T Y)^{\frac{1}{2}}$, then

$$\begin{aligned} &\min_{f,W,b} \left\| X^T W + \mathbf{1} b^T - \hat{Y} \right\|_F^2 \\ \Leftrightarrow &\min_{f,W,b} \left\| (X^T W + \mathbf{1} b^T - \hat{Y}) Q \right\|_F^2, \end{aligned} \tag{38}$$

because Q is constant. Besides, we have $\hat{Y} Q = Y$. In this way, the optimization can be:

$$\min_{f,W,b} \left\| X^T W Q + \mathbf{1} b^T Q - Y \right\|_F^2, \tag{39}$$

which means when we use normal one-hot label Y, it is the same as optimize WQ and $b^T Q$ from the corresponding W and b^T with normalized label. Although the optimization of W and b is changed, the optimal solution to f, i.e. the network, is the same.

D. Differences between F-loss and CCE

As mentioned in section about F-loss and CCE, the formulation of different loss leads to diverse network application. Because we use classification accuracy as our evaluation criteria, it is necessary to compare our F-loss method with CCE method.

We use the simplified network structure and MNIST dataset. Learning rate is set to 0.001 and batch size is set to 200.

The compare results of training accuracy and loss are presented in Figure 6 and Figure 7. We can observe that our method has a faster convergence speed and even higher accuracy using the same learning rate. The classification accuracy in test set of softmax is 99.27%. The classification accuracy in test set of our method using one-hot label is 99.37%. Due to the regularized term in the loss function, the generalization ability of our network can be promoted. In some way, the generalization ability makes the loss more smooth.

The classification accuracy for different losses and labels is shown in Table VI. From Table VI we can observe that the proposed method achieves a comparable performance with Softmax method in terms of classification accuracy.

TABLE VI
CLASSIFICATION ACCURACY WITH DIFFERENT LOSSES AND LABELS.

Method	Classification accuracy
F-loss with normalized label	99.20
F-loss with one-hot label	**99.37**
Softmax	99.27

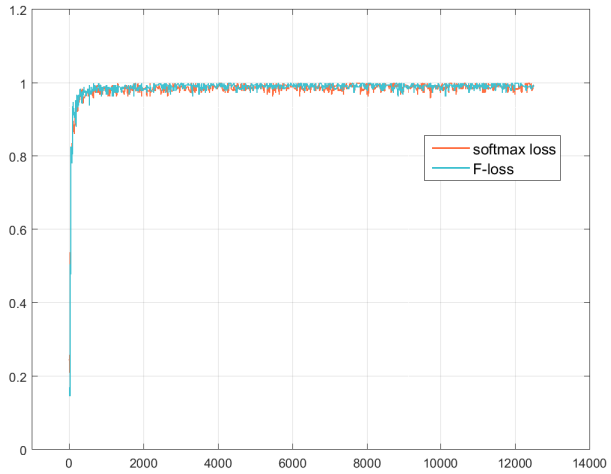


Fig. 6. Training accuracy of softmax method and F-loss method.

TABLE VII
CLASSIFICATION ACCURACY WITH DIFFERENT PARAMETERS WITH 20 TRAINING SAMPLES. THE BEST RESULTS OF OTHER EIGHT TRADITIONAL METHODS ARE SHOWN.

Dataset	C2DNDA (simplified)	Best	C2DNDA $^+$ (simplified)
CVL(SVM)	72.2	**79.2 ± 1.5**	75.3
CVL(1NN)	72.2	74.2 ± 1.1	**75.3**
USPS(SVM)	90.2	88.4± 1.4	**91.4**
USPS(1NN)	90.2	89.2 ± 1.4	**91.4**

TABLE VIII
CLASSIFICATION ACCURACY WITH DIFFERENT PARAMETERS WITH 10 TRAINING SAMPLES. THE BEST RESULTS OF OTHER EIGHT TRADITIONAL METHODS ARE SHOWN.

Dataset	C2DNDA (simplified)	Best	C2DNDA $^+$ (simplified)
CVL(SVM)	55.1	**68.3 ± 1.5**	62.9
CVL(1NN)	55.1	**67.3 ± 1.3**	62.9
USPS(SVM)	82.6	84.3± 1.3	**86.2**
USPS(1NN)	82.6	84.4 ± 1.5	**86.2**

E. Discussion about the Classification Accuracy

What we can see from the above experiments is that the proposed method shows better performance with more data. This phenomenon can be explained in three aspects.

First, the insufficient data will make the optimization of the objective function defined in Eq.26 converge into a bad point. With insufficient samples, the proposed dimensionality reduction network is optimized with non-optimal W and b.

Second, the over-fitting problem when training with small sample size and deep networks would be serious. From Table IV we can see that the overall training set contains 100 samples. The trained CNN would have bad generalization ability.

The last aspect is about a training trick. All of the experiments in different datasets use the same experimental settings. Some important hyper parameters like learning rate can lead to different performance especially with a small train set. If the learning rates can be adjusted for insufficient data or different datasets, the results would be better. In Table VII and Table VIII, some hyper parameters are adjusted for better performance. The experimental setting of the learning rate is changed to overcome the over-fitting problem. The proposed method with $^+$ in Table VII and VIII uses a higher learning rate.

V. CONCLUSION

In this paper, we propose an extension of the C2DNDA method for nonlinear dimensionality reduction. The difficult problem of embedding the classical LDA into a network is solved by an F-loss function which is equal to the classical LDA objective function. The proposed C2DNDA method utilizes a two-stage networks to realize dimensionality reduction. Effectiveness of this kind of structure is proved with various backbone networks, losses and labels. Our C2DNDA method outperforms the classical LDA and 2DLDA in every experimental settings. Meanwhile, the proposed method outperforms the DeepLDA due to an easier optimization approach. With enough data, the C2DNDA method gains a state-of-the-art classification and dimensionality reduction performance.

In the proposed C2DNDA framework, a more efficient backbone network compared with plain networks we use can be added. However, most of the proposed network structures are aimed at classification or detection. Thus, we will focus on more powerful backbone networks designed for dimensionality reduction in the future.

REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[2] W. Zhao, R. Chellappa, and P. J. Phillips, *Subspace linear discriminant analysis for face recognition*. Computer Vision Laboratory, Center for Automation Research, University of Maryland, 1999.

[3] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[4] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2599–2606, 2009.

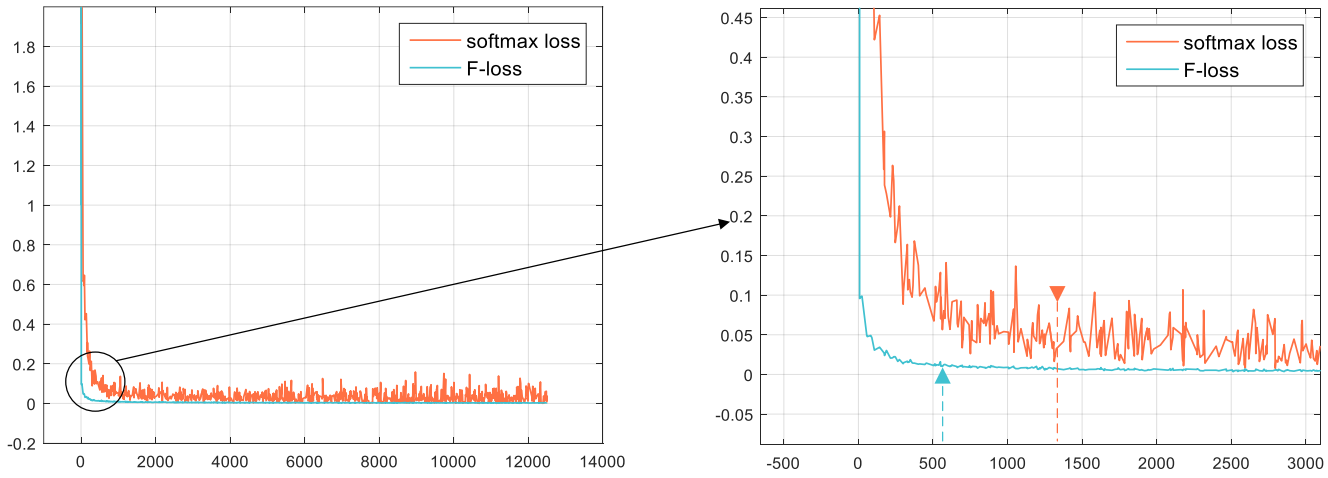


Fig. 7. Training loss of softmax method and F-loss method. The convergence point of two methods is shown by triangles respectively.

[5] Z. Ma, Y. Yang, F. Nie, and N. Sebe, "Thinking of images as what they are: compound matrix regression for image classification," in *International Joint Conference on Artificial Intelligence*, pp. 1530–1536, 2013.

[6] C.-X. Ren, D.-Q. Dai, and H. Yan, "Coupled kernel embedding for low-resolution face image recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3770–3783, 2012.

[7] B. Zou, L. Li, Z. Xu, T. Luo, and Y. Y. Tang, "Generalization performance of fisher linear discriminant based on markov sampling," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 2, pp. 288–300, 2013.

[8] M. M. Lopez and J. Kalita, "Deep learning applied to nlp," *arXiv preprint arXiv:1703.03091*, 2017.

[9] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[10] M. Li and B. Yuan, "2d-lda: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[11] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[12] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in neural information processing systems*, pp. 1569–1576, 2004.

[13] C.-X. Ren, D.-Q. Dai, X. He, and H. Yan, "Sample weighting: An inherent approach for outlier suppressing discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3070–3083, 2015.

[14] W.-S. Zheng, J.-H. Lai, and S. Z. Li, "1d-lda vs. 2d-lda: When is vector-based linear discriminant analysis better than matrix-based?" *Pattern Recognition*, vol. 41, no. 7, pp. 2156–2172, 2008.

[15] T. H. Chan, K. Jia, S. Gao, and J. Lu, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2014.

[16] X.-m. Wang, C. Huang, X.-y. Fang, and J.-g. Liu, "2dpca vs. 2dlda: Face recognition using two-dimensional method," in *International Conference on Artificial Intelligence and Computational Intelligence*, vol. 2, pp. 357–360, 2009.

[17] M. Dorfer, R. Kelz, and G. Widmer, "Deep linear discriminant analysis," *arXiv preprint arXiv:1511.04707*, 2015.

[18] Q. Wang, Z. Qin, F. Nie, and Y. Yuan, "Convolutional 2d lda for nonlinear dimensionality reduction," in *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2929–2935, 2017.

[19] J. Ye, R. Janardan, Q. Li, and H. Park, "Feature reduction via generalized uncorrelated linear discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1312–1322, 2006.

[20] Y. Hou, I. Song, H. K. Min, and C. H. Park, "Complexity-reduced scheme for feature extraction with linear discriminant analysis." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 6, pp. 1003–1009, 2012.

[21] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognition Letters*, vol. 26, no. 2, pp. 181–191, 2005.

[22] Y. Pang, S. Wang, and Y. Yuan, "Learning regularized lda by clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2191–2201, 2014.

[23] M. S. Mahanta, A. S. Aghaei, and K. N. Plataniotis, "Regularized lda based on separable scatter matrices for classification of spatio-spectral eeg patterns," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1237–1241, 2013.

[24] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, pp. 1247–1255, 2013.

[25] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *Computer Science*, vol. 3, no. 4, pp. 212–223, 2012.

[29] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[31] R. Girshick, "Fast r-cnn," in *IEEE international conference on computer vision*, pp. 1440–1448, 2015.

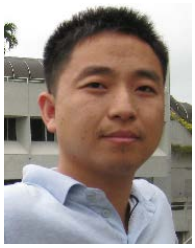
[32] J. Gao, Q. Wang, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," in *IEEE International Conference on Robotics and Automation*, pp. 219–224, 2017.

[33] K. Inoue and K. Urahama, "Non-iterative two-dimensional linear discriminant analysis," in *International Conference on Pattern Recognition*, pp. 540–543, 2006.

[34] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Advances in neural information processing systems*, pp. 499–506, 2005.
[35] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal laplacianfaces for face recognition," *IEEE transactions on image processing*, vol. 15, no. 11, pp. 3608–3614, 2006.
[36] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Bilinear classifiers for visual recognition," in *Advances in neural information processing systems*, pp. 1482–1490, 2009.
[37] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, "Compound rank-k projections for bilinear analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 7, pp. 1502–1513, 2016.
[38] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
[39] I. J. Goodfellow, D. Wardefarley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *Computer Science*, pp. 1319–1327, 2013.
[40] B. Graham, "Spatially-sparse convolutional neural networks," *arXiv preprint arXiv:1409.6070*, 2014.
[41] M. Diem, S. Fiel, A. Garz, M. Keglevic, F. Kleber, and R. Sablatnig, "Icdar 2013 competition on handwritten digit recognition (hdrc 2013)," in *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1422–1427, 2013.
[42] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, pp. 1139–1147, 2013.
[43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[44] M. Abadi, P. Barham, J. Chen, Z. Chen *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Zequn Qin received the B.E degree in computer science technology from Northwestern Polytechnical University, Xi'an, China, in 2016. He is currently pursuing the M.E degree with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His current research interests include computer version and machine learning.



Feiping Nie Feiping Nie received the Ph.D. degree in Computer Science from Tsinghua University, China in 2009, and currently is full professor in Northwestern Polytechnical University, China. His research interests are machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing and information retrieval. He has published more than 100 papers in the following journals and conferences: TPAMI, IJCV, TIP, TNNLS, TKDE, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, ACM MM. His papers have been cited more than 15000 times and the H-index is 67. He is now serving as Associate Editor or PC member for several prestigious journals and conferences in the related fields.

Xuelong Li (M'02-SM'07-F'12) is a full professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.