Dict Layer: A Structured Dictionary Layer

Yefei Chen, Jianbo Su
Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
{marschen,jbsu}@sjtu.edu.cn

Abstract

Dictionary learning and deep learning both have played important roles in computer vision. Dictionary learning strives to learn best representative atoms to reconstruct the source images or signals. Alternating iterative methods are developed to solve such problems. On the other hand, deep learning performs in an end-to-end mode, where both feature extraction and classification are achieved simultaneously. Obviously, the scheme of deep learning is different from that of traditional dictionary learning which is shallow and focuses more on data reconstruction. However, studies on building a deep layer-stacked model imply that there could be a relationship between them. In this paper, the relationship between dictionary learning and deep learning is studied. Dictionary learning can be viewed as a special full connection layer (FC Layer) of deep learning. According to the relationship, we try to introduce those mature improvements from dictionary learning to deep learning. Hence, a new kind of layer named as Dict Layer is introduced in this paper, where the idea of structured dictionary is adopted. In Dict Layer, neural units (coefficients) are class specified, which means the activated neural units are encouraged to be the same class. The proposed method is evaluated on MNIST, CIFAR-10 and SVHN as an improvement of FC Layer. Experiments on AR and Extended YaleB are conducted where Dict Layer is viewed as a special form of dictionary learning method. Results show that the outputs of Dict Layer are more discriminative and class specific than that of the traditional FC Layer.

1. Introduction

Deep learning has achieved significant accomplishment in image classification and face recognition [11, 20]. It is considered to be an end-to-end learning method that both feature extraction and classification are intergraded. Stacking layer by layer is a classical skill to build a deep model [4, 7]. On the other hand, dictionary learning attempts to build representative dictionary to reconstruct images [1]. It seems that dictionary learning is a shallow model. However, recent studies [38–40] point out that dictionary learning could also be stacked into a deep model. Lu *et al*. [27] attempt to combine deep learning with dictionary learning to perform feature extractions. Hence, it is intuitive to consider whether there is any specific relationship between those two methods.

In the field of deep learning, researchers put emphasis on designing loss function [14, 36, 49], regularity [41], activation function [10, 31], architecture of the stacking layers[3, 11, 18, 20, 43] and so on, in order that the network could converge more faster, achieve less classification error and avoid over-fitting. As a result, the neural network is going deeper and deeper [44, 48]. In response to the requirements on accurate classification, the distribution of output layer is enforced to have certain properties such that features from the same class cluster closely [49] and the margin between positive and negative pairs should be greater than a constant value [36]. Studies [2, 12, 34] on mimicking a big model show that learning a distribution is more easier than training a model directly. It implies that knowing more class information is good for training a model and such information can be used as a hint not only on the last layer but also on the mid layer [35]. Therefore, we argue that class information could also be introduced into former layers such as FC Layer. Meanwhile, researchers have sought out a way to explain the network [6, 28, 37, 54]. Though the meaning of activation of neural units are studied [37], it still remains an opening problem.

Dictionary learning [1] is motivated by the study of sparse representation which encodes the original signal by a specific dictionary. At first, it aims to learn atoms of dictionary for signal reconstruction, which has a definite explanation of structure. Lately, the classification ability of coefficients is taken into consideration and the orthogonality constraints are released [52]. Many discriminative penalties are designed on the coefficients where label information is added. Then, methods focusing on supervised dictionary learning are proposed [17, 51, 52]. Discriminative cost function like softmax [29] is also applied in dictionary learning. Previous study shows that collaboration instead of sparsity can also achieve good results on face recognition [56]. In general, three aspects of improvements are proposed: feature extraction on raw data [8], structured information for atoms of dictionary [16] and classification loss function on coefficients [57]. Despite of the regularization on coefficients, dictionary learning differs from deep learning mainly in the aspect of structured information.

Structured dictionary information [42, 45, 51] can significantly enhance the classification performance of dictionary learning and it has two advantages. The first one is that the coefficients are class specified which is helpful for classification. When a signal is input, the related group of coefficients are expected to be large or activated. The second one is that it has a clear explanation for such phenomenon: class specified dictionary can promote the coefficients being class specified. Class information is introduced to dictionary and coefficients by enforcing class constraints on dictionary. The activated coefficient denotes the class that the input belongs to, while this property is not shared by traditional FC Layer. It is intuitive to think of introducing such information into deep learning and encourage all the activated neural units coming from the same group.

In order to use such structured dictionary information, one straightforward approach is training deep learning and dictionary learning as feature extraction and classification respectively. Linear feature extraction [8, 26, 55] and nonlinear feature extraction [27] is utilized in dictionary learning. Such training procedures need alternating iterative methods. Those approaches eliminate the drawback that dictionary learning highly depends on the feature extracted on the raw data. But the most significant characteristic of deep learning is lost, where feature extraction and classification is conducted in an unified model. As mentioned above, introducing class information into former layer is good for training. Therefore, we consider to combine structured dictionary information into deep learning structure. The new layer structure is named as Dict Layer and the desired framework is illustrated in Fig. 1. It inherits the characteristics from both of them.

In this paper, the relationship between dictionary learning and deep learning is discussed. Dictionary learning can be considered as a special layer of deep learning. Then, we explore a way to utilize those classical techniques of dictionary learning to improve deep learning. Thus, a new FC layer named Dict Layer is proposed, which inherits the characteristics of structured dictionary information. The proposed layer can be used to replace FC Layer on any existing deep learning methods under the condition that FC Layer is used as their last but one layer and the classification category is identical in training and testing. Experimental results show that Dict Layer introduces additional class specific information and generates a more discriminative distribution.

2. Related Work

Deep Stacking Networks. Deep Stacking Networks (DSNs) is composed of simplified neural network modules [23]. The scheme of DSN usually has two steps: formulating each layer separately and then stacking them together. Li *et al.* [24] take local dependencies among hidden units into consideration by utilizing both ℓ_1 and ℓ_2 regularization [42]. Thus, coefficients are split into separate groups related to each class. Inputs of former layers are concatenated with the output of current layer and then the concatenated feature is used as input for the next layer. Though DSN proposed in [24] can be stacked on CNN, it is trained separately and the output of CNN is only used as features extracted. The performance of DSN depends on the features extracted by the pretrained model which will not participate in training. It prevents the deep model from extracting features more feasible for classification.

Deep Dictionary Learning. Dictionary learning can be treated as a method of learning atoms of dictionary by matrix factorization [46]. The atoms of dictionary can be considered as features of images like convolutional filters in deep learning. Singhal *et al.* [40] introduce the idea of Stacked Auto-Encoder into dictionary learning, where dictionaries are learned layer by layer and stacked together. It is a good attempt to extend the shallow model into a deep one and combine those two paradigms together [38]. But each layer is trained separately without a final fine-tuning procedure and the connection between deep learning and dictionary learning is not revealed.

Simultaneous Feature Learning and Dictionary Learning. As mentioned above, feature extraction on raw image is vital for dictionary learning. Thus, lots of studies focus on optimizing feature extraction and dictionary learning jointly [5, 8, 25–27]. Previous works on linear feature extraction perform like a linear subspace learning, where a projection matrix is learned to project the high dimensional raw image into a low dimensional manifold [5, 8, 25]. Lu *et al.* [27] extend the linear feature extraction to nonlinear feature extraction and try to solve them jointly. However, the relationship between those two methods are not discussed and dictionary learning is not considered as a part of deep learning.

Multi-Task Convolution Neural Network. The proposed Dict Layer based on the idea of structured dictionary gives rise to the coefficients of output grouped class by class. This characteristic is also desired in Multi-Task Learning proposed by Yin and Liu [53]. They argue that features from the FC Layer are entangled by different tasks. Thus, weights matrix is directly separated for different tasks and the feature is shared by all tasks. However, we introduce the idea from dictionary learning and a regularity term is used to control the degree of class specific.

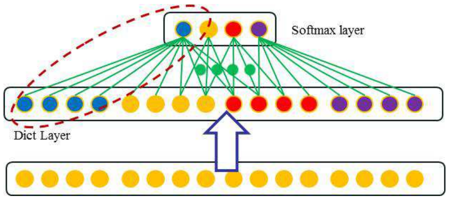

Figure 1. Framework of Dict Layer. The neural units in Dict Layer are grouped class by class and sensitive to class. Blue, yellow, red and purple stand for 4 classes. The neural units(ceofficients), related to specific color(class), in Dict Layer are encouraged to be activated(significant) when an image from the specific class is input.

3. Dict Layer

In this section, we firstly discuss the relationship between dictionary learning and deep learning. It can be proven that dictionary learning can be considered to be a special form of deep learning. Therefore, dictionary learning can be used to improve the traditional FC Layer of deep learning. Class specific information is introduced into FC Layer and this new layer, named as Dict Layer, is guided by structured dictionary information. Each activation of neural units has a clear meaning with respect to each class as illustrated in Fig. 1.

3.1. Annotation

Annotation used in this paper is given here. Let N be the number of training data. $Y = [Y_1, \dots, Y_c, \dots, Y_C]$ denotes input images for training with C classes. Each $Y_c = [y_{c_1}, \dots, y_{c_i}, \dots, y_{c_k}]$ is composed of c_k images. So that, $N = \sum_{i=1}^{k} n_{c_i}$. $D = [D_1, \dots, D_c, \dots, D_C]$ denotes atoms of dictionary separated by C classes. Each $D_c = [d_{c_1}, \dots, d_{c_i}, \dots, d_{c_p}]$ is composed of c_p atoms. $D_{co} = [0, \dots, D_c, \dots, 0]$ denotes dictionary with atoms only from class c. $X = [X_1, \dots, X_c, \dots, X_C]$ denotes coefficients related with D. Each $X_c = [x_{c_1}, \dots, x_{c_i}, \dots, x_{c_k}]$ is composed of c_k coefficients. Let X_c^c denotes coefficients related belong to class c and related with dictionary D_c. $L = [l_1, \dots, l_i, \dots, l_N]$ denotes labels for each data. $f(\cdot)$ denotes feature extraction, which has a layer by layer structure. $g(\cdot)$ denotes classification loss function on coefficients, which is 0-1 cross-entropy with softmax as last layer in deep learning or other classification loss functions in dictionary learning.

3.2. Dictionary Learning and Deep Learning

Dictionary learning [1] is firstly derived from the study of sparse representation [50] where the sparsity of coefficients is desired. Lately, Zhang *et al.* [56] point out that collaboration can also help in dictionary learning, which is easier for computation than solving a Lasso problem. For simplicity, we focus on the aspect that only collaborative representation is considered. Dictionary learning focuses on minimizing the reconstruction error. Penalty constraints on coefficients also take part in the final loss function. Thus, the objective function is constructed as Eq. (1).

$$\min_{D,X,f} \{\|f(Y) - DX\|_F^2 + \lambda\|X\|_F^2 + g(X, L)\}, \tag{1}$$

where the first term is used for reconstruction and representation. The second term can be viewed as a regularizer. The third term is the classification loss function, where softmax [29], pairwise distance [5] or affinity matrix [27] can be applied.

Three aspects of improvements can significantly enhance the performance of dictionary learning: designing a more representative feature extraction function $f(\cdot)$, introducing a more discriminative classification loss function $g(\cdot)$ and adding structured dictionary information. Thus, the classical objective function with class specific dictionary is constructed as Eq. (2).

$$\begin{aligned}\min_{D,X,f} \{ &\sum_{c=1}^{C} \|f(Y_c) - DX_c\|_F^2 + \lambda\|X\|_F^2 \\ &+ \sum_{c=1}^{C} \|f(Y_c) - D_c X_c^c\|_F^2 + g(X, L)\},\end{aligned} \tag{2}$$

where the third term enforces the dictionary to be class specific. Meanwhile, the coefficients X are encouraged to be grouped by class. Specifically, each x_i in X is a column vector, where each row is corresponding with the column of dictionary D. Therefore, the value of x_i related to specific class will be encouraged to be large or significant when inputting an image. It is a good property which is achieved by introducing class information to dictionary. Below, we will discuss the relationship between deep learning and dictionary learning.

Deep learning is usually trained by backpropagation. Recent studies try to find a way to get rid of backpropagation, because the poor conditioning and vanishing gradients will slow down gradient-based methods. Jaderberg *et al.* [6, 15] introduce the idea of decoupling backpropagation, where derivation is estimated in forward procedure. The idea of target propagation introduced by Lee *et al.* [22] also aims to solve the problem mentioned above. Taylor *et al.* [47] explore an unconventional training method that uses alternating direction methods and Bregman iteration to train networks without gradient descent steps. This method decomposes deep neural network training into a sequence of substeps. It is similar to the optimization method used in dictionary learning. We consider that similarity comes from the close relationship between those two optimization problem.

The optimization problem of deep learning is formed as:

$$\begin{aligned} \min_{x_m, a_m, W_m} \quad & \{g(x_M, l)\} \\ s.t. \quad & x_m = W_m a_{m-1}, for\ m = 1, 2, \ldots, M \\ & a_m = h_m(x_m),\ for\ m = 1, 2, \ldots, M \end{aligned} \tag{3}$$

where $h(\cdot)$ is the nonlinear activation function, *e.g.* ReLU, sigmoid and tanh. The network is stacked by M layers. If we absorb W_m, h_m, a_m and x_m into one entire function $f(\cdot)$, the constraints can be written as $x_M = f(a_1) = f(y)$.

$$\begin{aligned} \min_{f, x_M} \quad & \{g(x_M, l)\} \\ s.t. \quad & x_M = f(y) \end{aligned} \tag{4}$$

Let's consider the traditional form(Eq. (1)) of dictionary learning without structured dictionary information. If we rigorously restrict the Frobenius-norm constraints to be equality constraints, Eq. (5) can be derived from Eq. (1).

$$\begin{aligned} \min_{D, X, f} \quad & \{\sum_{c=1}^{N}(\lambda\|x_c\|_2^2 + g(x_c, l_c))\} \\ s.t. \quad & f(y_c) = Dx_c \end{aligned} \tag{5}$$

If D is absorbed in $f(\cdot)$ and the regularizer is absorbed in $g(\cdot)$, the traditional dictionary learning has the same mathematical expression as deep learning(Eq. (4)). Therefore, dictionary learning is considered to be a special layer of deep learning. The dictionary learning layer is different from traditional FC Layer in the aspects that the dictionary D is assumed to be orthogonal or ℓ_2-norm 1 and no nonlinear activation function is needed. Though the dictionary learning focuses more on data reconstruction, they can be written in the same mathematical expression. Here, we construct a relationship between dictionary learning and deep learning. The effort of learning a feature extraction $f(\cdot)$ and designing loss function $g(\cdot)$ consist with deep learning. Hence, in this paper coefficients and neural units are the same thing viewed in different ways. The dimensionality of FC Layers can also be viewed as the number of atoms in dictionary. Next, we try to introduce structured dictionary information into deep learning.

3.3. Dict Layer: A Structured Dictionary Layer

In this section, a new full connection layer namely Dict Layer is introduced.

As mentioned above, dictionary learning can be viewed as a special form of deep learning. Also, the structured dictionary information helps for classification and gives a clear meaning for the activation of coefficients. It encourages the coefficients grouped by class and sensitive to specific class. It is a good property which traditional FC Layer does not have. It can help to understand the meaning of activation happened in layer, where each neural unit is related with each class and only neural units related with specific class are encouraged to be activated or significant as illustrated in Fig. 1.

Let's consider the reconstruction term in Eq. (2), which can be written as Eq. (6).

$$\|f(Y_c) - DX_c\|_F^2 = \sum_{c=1}^{c_k} \|f(y_c) - Dx_c\|_F^2. \tag{6}$$

If we rigorously restrict Frobenius-norm constraints to be equality constraints, Eq. (7) can be derived.

$$\begin{aligned} f(y_c) \quad & = Dx_c \\ \Rightarrow x_c \quad & = (D^T D)^{-1} D^T f(y_c) \\ W \quad & \triangleq (D^T D)^{-1} D^T \\ \Rightarrow x_c \quad & = W f(y_c) \end{aligned} \tag{7}$$

Similarly, the structured dictionary term in Eq. (2) can be written as Eq. (8).

$$\|f(Y_c) - D_c X_c^c\|_F^2 = \sum_{c=1}^{c_k} \|f(y_c) - D_{co} x_c\|_F^2. \tag{8}$$

With Frobenius-norm constraints being restricted to be equality constraints, Eq. (9) can be derived. Here, D is considered to be orthogonal between different classes which means the product of atoms from different classes is encouraged to be zeros.

$$\begin{aligned} f(y_c) \quad & = D_{co} x_c \\ \Rightarrow x_c \quad & = (D_{co}^T D_{co})^{-1} D_{co}^T f(y_c) \\ x_c \quad & = [0; (D_c^T D_c)^{-1} D_c^T; 0] f(y_c) \\ W_{co} \quad & \triangleq [0; (D_c^T D_c)^{-1} D_c^T; 0] \\ \Rightarrow x_c \quad & = W_{co} y_c = [0, I, 0] W f(y_c) \end{aligned} \tag{9}$$

It is obvious that x_c derived from Eq. (7) and Eq. (9) are identical. So a Frobenius-norm constraints written as Eq. (10) can be applied to enforce the x_c being class specific.

$$\begin{aligned} & \|W_{co} f(y_c) - W f(y_c)\|_F^2 = \|[I, 0, I] W f(y_c)\|_F^2 \\ & \triangleq \|W_{oc} f(y_c)\|_F^2. \end{aligned} \tag{10}$$

The constraint is easy to understand that it encourages the units related with other classes being small. W_{oc}, where rows related with class c are set to be zeros, is complementary with W_{co}. Such a constraint can be added into the final loss function of deep learning and trained in traditional backpropagation manner. It is easy to derive the derivation from such constraint. The derivation and gradient of parameter W are given in Eq. (11). It can be applied to any conventional stochastic gradient method by simply adding a multiplier λ to control regularization. Therefore, the final optimization problem for deep learning with Dict Layer at m-th layer can be written as

$$\begin{aligned} \delta^m_\star &= \lambda W_{oc} f_m(Y_c) \circ h'_m(X^m_c), \\ \delta^m &= W^T_{m+1} \delta^{m+1} \circ h'_m(X^m_c) + \delta^m_\star, \\ \frac{\partial E}{\partial W_m} &= \frac{(f_{m-1}(Y_c))^T \delta^m}{n_c}. \end{aligned} \tag{11}$$

$$\min_f \{ \sum_{c=1}^{N} (g(f(y_c), l_c) + \lambda \| W_{co} f(y_c) \|^2_F) \}. \tag{12}$$

Here, a new full connection layer, named as Dict Layer, is introduced, where a structured dictionary constraint is taken into consideration. Dict Layer inherits the property of structured dictionary that the output of this layer is grouped by class and sensitive to class. Neural units in this layer are class specific. Only neural units related to a specific class are encouraged to be activated or significant and others are forced to be near zero. This property gives a clear meaning of units being activated and class information is used as a guidance during the mid-layer of training.

4. Experiments

In this section, experimental results of Dict Layer and FC Layer are evaluated on MNIST, SVHN, CIFAR-10, AR and Extended YaleB databases. The network framework used to compare Dict Layer and FC Layer is identical in each experiment except the last but one layer, which is Dict Layer and FC Layer respectively. Also, Dict Layer can be viewed as a special training method for dictionary learning. The output of the Dict Layer can also be viewed as coefficients of dictionary learning. Thus, in the following experiments the dimensionality of Dict Layer has the same meaning as the number of dictionary atoms which is the product of class number and atoms number for each class. Experiments are conducted in comparison with other state-of-the-art dictionary learning methods on AR database.

4.1. MNIST

MNIST [21] database consists of 28×28 images of handwritten digits ranging from 0 to 9, with $60,000$ images for training and $10,000$ images for testing. CNN with FC Layer is used as baseline. The network is composed of 2 convolution layers each followed by a max-pooling layer, 1 FC Layer and softmax as last layer. Convolution layer uses 32 and 64 feature maps respectively. Filter size of each layer is set to be 5×5 with padding 2. The pooling scale is set to be 3 with stride 2. ReLU is applied as activation function and dropout for FC Layer is 0.2. In our network, all layers are identical except the FC Layer which is replaced by Dict Layer. The dimensionality of Dict Layer is same as that of FC Layer. Different numbers of atoms are evaluated. Both models are trained by 20 epochs.

Dim	$\lambda = 0.01$	baseline
	Dict Layer	FC Layer
100	**0.50%**	0.85%
200	**0.60%**	0.71%

Table 1. Classification errors on MNIST database with different number of dictionary atoms.

Here, 2 sizes of dictionary are evaluated in the experiment, of which 10 and 20 atoms for each class are used. The classification errors are illustrated in Table 1. Results show that model with Dict Layer is slightly better than baseline. Fig. 2 demonstrates the output distributions of Dict Layer compared with the that of groundtruth. 10 training images are selected randomly from the database and their corresponding coefficients of Dict Layer are calculated. Each coefficient vector is divided into 10 classes and its absolute value is summed up class by class and illustrated in Fig. 2. Fig. 3 demonstrates the result from testing database. These results show that the structured dictionary information is persevered by Dict Layer and the coefficients of Dict Layer are strongly class specific which is helpful for classification. In order to illustrate the relationship between coefficients and class for all the testing samples. The magnitude of coefficients from the same class are summed up together and averaged. The first 10 significant atoms for each class are highlighted in yellow. Fig. 4 and Fig. 5 demonstrate the results of Dict Layer and FC Layer respectively. Most of the blocks highlighted are clustered along the diagonal in Fig. 4, while the blocks looks randomly in Fig. 5. This phenomenon show that the coefficients from Dict Layer are class specific. The results show that activation of neural units from Dict Layer is highly related with classification problem. The value of neural units related to specific class are encouraged to be significant when an image from the specific class is input.

4.2. SVHN

Street View House Numbers(SVHN) [32] database is composed of $604,388$ training images and $26,032$ test images. Each image is cropped 32×32 color image with candidate digit located in the center. In our experiment, only $73,257$ images from the difficult training set are used. C-NN with FC Layer proposed by Hinton *et al.* [13] is used as baseline, which has 3 convolution layers with 5×5 filters and 64 feature maps per layer. Each convolution layer is followed by a max-pooling layer with 3×3 filters and stride 2. ReLU is applied as activation function and dropout for FC Layer is 0.2. In our network, all layers are identical except the FC Layer, which is replaced by Dict Layer. The

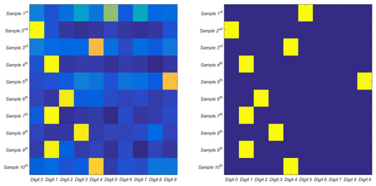

Figure 2. The magnitude of coefficient distributions of Dict Layer(left) from 10 random training samples on MNIST and its groundtruth(right). Each row stands for a sample. The column stands for the magnitude of coefficients and is divided into 10 classes according to the database. The magnitude of coefficients is then summed up according to the division.

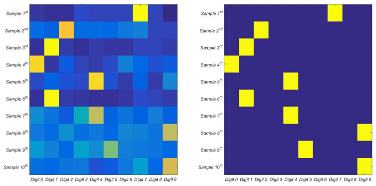

Figure 3. The magnitude of coefficient distributions of Dict Layer(left) from 10 random testing samples on MNIST and its groundtruth(right).

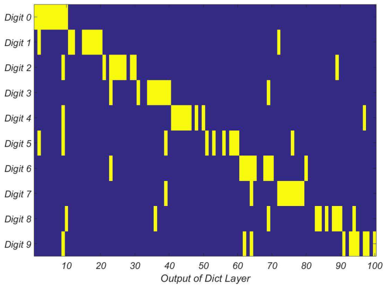

Figure 4. Rank-10 coefficients distributions of Dict Layer with respect to class on all testing samples from MNIST. Rows stand for 10 classes. Columns stand for 100 atoms of Dict Layer. For each class, 10 most significant magnitude of coefficients are highlighted in yellow.

dimensionality of Dict Layer is same as that of FC Layer. Different numbers of dictionary atoms are evaluated. Both models are trained by 60 epochs.

Here, 3 sizes of dictionary are evaluated, of which 30, 40 and 50 atoms for each class is used. The classification errors are illustrated in Table 2. Results show that model with Dict Layer is slightly better than baseline.

4.3. CIFAR-10

CIFAR-10 [19] database is composed of 10 classes of natural images with 50,000 training images and 10,000 testing images. Each image is cropped 32×32 color image. In our experiment, only 73,257 images from the training set are used, where only difficult training set is evaluated. CNN framework conducted in this experiment is identical to that used in SVHN database. Different numbers of dictionary atoms are evaluated. Here, 2 sizes of dictionary are evaluated, of which 40 and 50 atoms for each class are used. Both models are trained by 80 epochs. Experimental results show that classification accuracy of Dict Layer is about 1% better than that of baseline model.

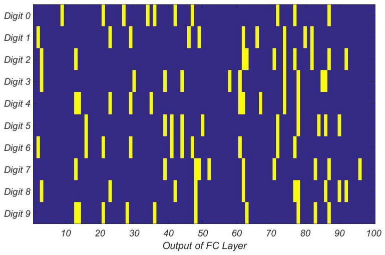

Figure 5. Rank-10 coefficients distributions of FC Layer with respect to class on all testing samples from MNIST.

Dim	$\lambda = 0.01$	$\lambda = 0.1$	baseline
	Dict Layer	Dict Layer	FC Layer
300	8.48%	**8.05%**	8.94%
400	8.68%	**7.92%**	8.69%
500	8.41%	**7.59%**	8.62%

Table 2. Classification errors on SVHN database with different number of dictionary atoms.

Dim	$\lambda = 0.01$	baseline
	Dict Layer	FC Layer
400	**22.78%**	22.83%
500	**22.98%**	23.12%

Table 3. Classification errors on CIFAR-10 database.

4.4. AR

AR [30] database is a face database which consists of over 4,000 images of 126 subjects, which varies in illumination, expression and accessory like scarfs and sunglasses. The subset containing 1,400 images of 100 subjects with 50 males and 50 females separately is chosen for evaluation. For each subject, 7 images from Session 1 are used for training and the rest 7 images from Session 2 are used for testing. All images are resized into 60×43. Models are trained by 200 epochs.

Because there is a small amount of training samples, training a deep model on this dataset is not feasible. Firstly, we construct a baseline model with 2 FC Layers and softmax as last layer and compare with our Dict Layer by

Dim	Randomly initial		WPCA initial	
	$\lambda = 0.001$	baseline	$\lambda = 1$	baseline
	Dict Layer	FC Layer	Dict Layer	FC Layer
500	**24.29%**	29.43%	**11%**	28.29%
600	**23.43%**	27.43%	**10%**	27.14%
700	**23.29%**	26.86%	**9%**	26.29%
800	**23.14%**	28.29%	**9.86%**	24.43%
900	**24.29%**	26.71%	**9.43%**	25.57%
1000	**25.29%**	28.00%	**10.57%**	24.14%

Table 4. Classification errors on AR database using softmax.

Dim	Randomly initial		WPCA initial	
	$\lambda = 0.001$	baseline	$\lambda = 1$	baseline
	Dict Layer	FC Layer	Dict Layer	FC Layer
500	**26.14%**	33.86%	**7.29%**	35.14%
600	**23.00%**	33.43%	**7.43%**	35.43%
700	**22.57%**	33.00%	**7.14%**	35.29%
800	**25.43%**	32.71%	**7.57%**	32.14%
900	**23.14%**	34.00%	**6.86%**	32.57%
1000	**26.27%**	34.00%	**6.71%**	33.43%

Table 5. Classification errors on AR database using cosine distance.

Dim	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	baseline
	Dict Layer	FC Layer	Dict Layer	FC Layer
500	11.43%	6.57%	**5.29%**	20.57%
600	9.00%	5.71%	**5.00%**	19.29%
700	8.71%	5.43%	**4.86%**	19.14%
800	9.86%	5.71%	**5.29%**	19.43%
900	8.43%	5.29%	**5.00%**	18.43%
1000	7.71%	5.29%	**4.57%**	16.71%

Table 6. Classification errors on AR database using softmax based on the pretrained model: VGGNet(pool4).

Dim	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	baseline
	Dict Layer	Dict Layer	Dict Layer	FC Layer
500	12.86%	6.57%	**5.57%**	27.43%
600	11.71%	5.71%	**4.86%**	26.86%
700	10.14%	5.00%	**4.57%**	26.00%
800	10.71%	5.00%	**4.71%**	25.14%
900	10.00%	5.00%	**4.57%**	24.57%
1000	9.71%	4.71%	**4.86%**	23.86%

Table 7. Classification errors on AR database using cosine distance based on the pretrained model: VGGNet(pool4).

replacing the last but one layer with Dict Layer. The number of hidden neural units in first FC Layer is set to be 300. The dimensionality of the second layer is identical to that of FC Layer and Dict Layer and changes with the number of atoms for each class. It is obvious that such structure is similar with those dictionary learning method that learns a linear subspace projection and dictionary jointly. Hence, we initialize the first FC Layer with projection matrix by WPCA and remove the nonlinear activation function. In such a way, the model acts like a dictionary learning manner that learns linear subspace projection and dictionary simultaneously. Experiments are conducted on these two conditions with different numbers of atoms for each class ranging from 50 to 100.

Results illustrated in Table 4 use the last softmax layer as classifier. Results show that under both conditions Dict Layer performs better than baseline. Meanwhile, cosine distance is used to calculate the distance of the outputs from the last but one layer. Table 5 shows that classification accuracy is enhanced by 3% at most with WPCA initialization. The results are also competitive with state-of-the-art dictionary learning method. Comparing Table 4 and Table 5, the difference between Dict Layer and FC Layer is enlarged when using cosine distance. The reason is that outputs from Dict Layer are class specific and have structured dictionary information. Therefore, coefficients related to specific class are encouraged to be significant, which improve the classification ability of coefficients. The visualization figures from MNIST and SVHN also support such conclusion.

Secondly, we construct the baseline model with pretrained layers from VGGNet [33]. In this experiment, images are resized into 64×64 and input into VGGNet without resizing. The former part of VGGNet, ranging from input layer to pool4 layer, followed by 2 FC Layers and softmax as last layer is used as baseline model. The dimensionality of first FC Layer is still 300. The second FC Layer is replaced by Dict Layer in our model. Experimental results are illustrated in Table 6 and Table 7 corresponding to different classifiers. The classification error using the output of pool4 layer from VGGNet directly is 13.86% which is worse than that of our model with Dict Layer. Results show that Dict Layer adding on a pretrained model achieves better accuracy than that of the FC Layer.

At last, Dict Layer, viewed as an improvement for dictionary learning, is compared with other state-of-the-art dictionary methods. The number of atoms for each class is set to be 7, so that the dimensionality of Dict Layer is 700. Results are illustrated in Table 8. Dict Layer^1 denotes the model with WPCA initialization, while Dict Layer^2 denotes the model with VGGNet(pool4) pretrained. The result of Dict Layer is competitive with that of the state-of-the-art dictionary learning methods.

4.5. Extended YaleB

Extended YaleB [9] database consists of $2,414$ images of 38 individuals captured under various lighting conditions. 20 images from each person are randomly selected for

Method	Accuracy
FDDL [51]	92.2%
JDDRDL [8]	94.0%
BDDL [25]	93.6%
SEDL [5]	94.2%
VGG(pool4) [33]	86.14%
Dict Layer[1]	92.86%
Dict Layer[2]	**95.43%**

Table 8. Classification accuracy comparison with state-of-the-art dictionary learning method on AR database.

Dim	Randomly initial		WPCA initial	
	$\lambda = 0.001$	baseline	$\lambda = 0.5$	baseline
	Dict Layer	FC Layer	Dict Layer	FC Layer
190	**10.53%**	11.84%	**8.55%**	11.97%
380	**11.58%**	12.37%	**8.82%**	13.82%
570	**11.05%**	13.82%	**7.76%**	14.74%
760	**11.32%**	13.68%	**9.08%**	14.61%
950	**12.24%**	13.68%	**9.42%**	14.34%
1140	**13.03%**	13.95%	**8.68%**	14.61%

Table 9. Classification errors on Extended YaleB database using softmax with 20 training samples for each person.

training, while the rest for testing. All images are normalized to 54 × 48. Models are trained by 200 epochs.

Because there is a small amount of training samples, training a deep model on this dataset is not feasible. Thus, experiments are conducted the same as that on AR database. The baseline model is composed of 2 FC Layers and softmax as last layer and compare with our Dict Layer by replacing the last but one layer with Dict Layer. The number of hidden neural units in the first FC Layer is set to be 300. The number of atoms for each class ranges from 5 to 30 with step 5.

Comparing Table 9 and Table 10, the difference between Dict Layer and FC Layer is enlarged when using cosine distance. The reason is that outputs from Dict Layer are class specific and have structured dictionary information. Therefore, coefficients related to specific class are encouraged to be significant, which is useful to classification.

At last, Dict Layer is also compared with other state-of-the-art dictionary methods. Dict Layer is used with pretrained layers from VGGNet [33]. All the setting are identical to that AR database, except that images are resized into 80 × 80. The number of atoms for each class is set to be 20 equal to the number of training samples per person when comparing with other methods. Thus, the dimensionality of Dict Layer is 760. These results are illustrated in Table 11. Dict Layer denotes the model with VGGNet(pool4) pretrained. The result of Dict Layer the best of these dictionary learning methods.

Dim	Randomly initial		WPCA initial	
	$\lambda = 0.001$	baseline	$\lambda = 0.5$	baseline
	Dict Layer	FC Layer	Dict Layer	FC Layer
190	**10.00%**	11.84%	**7.24%**	20.79%
380	**10.66%**	14.47%	**7.11%**	23.80%
570	**11.58%**	16.32%	**6.97%**	27.50%
760	**11.05%**	18.55%	**6.97%**	29.08%
950	**12.37%**	19.21%	**6.71%**	30.92%
1140	**11.97%**	17.63%	**7.11%**	31.18%

Table 10. Classification errors on Extended YaleB database using cosine distance with 20 training samples for each person.

Method	Accuracy
FDDL [51]	94.4%
MFL [52]	91.3%
SEDL [5]	95.5%
FC Layer	95.75%
Dict Layer	**97.87%**

Table 11. Classification accuracy comparison with state-of-the-art dictionary learning method on Extended YaleB database.

5. Conclusion

In this paper, the relationship between dictionary learning and deep learning is discussed. Dictionary learning can be viewed as a special layer of deep learning. According to that discovery, we explore a way of improving deep learning by introducing those mature techniques from dictionary learning. A new kind of layer, named as Dict Layer, is proposed, of which structured dictionary information is taken into consideration. Dict Layer is realized with constraints on outputs which enforce the neural units to be grouped by class and sensitive to different class. Coefficients or neural units related to a specific class are encouraged to be significant or activated when an image from the specific class is input. Experimental results show that Dict Layer can enhance the classification ability of model and visualization figures can clearly explain the meaning of activation of the neural units. However, such Dict Layer method is restricted to the case that classification category of training set is identical to that of the testing set. The expansion of dimensionality with the growth of category still remains a problem to be investigated.

Acknowledgement

This work is supported by the program of National Nature Science Foundation of China under grants 61533012, 91748120 and 61521063.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, Nov 2006. 1, 3

[2] L. J. Ba and R. Caruana. Do deep nets really need to be deep? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2654–2662, Cambridge, MA, USA, 2014. MIT Press. 1

[3] I. Bello, B. Zoph, V. Vasudevan, and Q. V. Le. Neural optimizer search with reinforcement learning. *CoRR*, abs/1709.07417, 2017. 1

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 153–160, Cambridge, MA, USA, 2006. MIT Press. 1

[5] Y. Chen and J. Su. Sparse embedded dictionary learning on face recognition. *Pattern Recognition*, 64:51–59, 2017. 2, 3, 8

[6] W. M. Czarnecki, G. Swirszcz, M. Jaderberg, S. Osindero, O. Vinyals, and K. Kavukcuoglu. Understanding synthetic gradients and decoupled neural interfaces. *CoRR*, abs/1703.00522, 2017. 1, 3

[7] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, Mar. 2010. 1

[8] Z. Feng, M. Yang, L. Zhang, Y. Liu, and D. Zhang. Joint discriminative dimensionality reduction and dictionary learning for face recognition. *Pattern Recognition*, 46(8):2134–2143, 2013. 2, 8

[9] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, Jun 2001. 7

[10] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, Dec 2015. 1

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 1

[12] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *ArXiv e-prints*, Mar. 2015. 1

[13] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. 5

[14] J. Hu, J. Lu, and Y. P. Tan. Discriminative deep metric learning for face verification in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, June 2014. 1

[15] M. Jaderberg, W. M. Czarnecki, S. Osindero, O. Vinyals, A. Graves, and K. Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. *CoRR*, abs/1608.05343, 2016. 3

[16] X. Jiang and J. Lai. Sparse and dense hybrid representation via dictionary decomposition for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1067–1079, May 2015. 2

[17] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1697–1704, June 2011. 1

[18] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. PVANET: Deep but Lightweight Neural Networks for Real-time Object Detection. *ArXiv e-prints*, Aug. 2016. 1

[19] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 6

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., Lake Tahoe, Nevada, USA, 2012. 1

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. 5

[22] D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio. *Difference Target Propagation*, pages 498–515. Springer International Publishing, Cham, 2015. 3

[23] J. Li, H. Chang, and J. Yang. Sparse deep stacking network for image classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3804–3810, 2015. 2

[24] J. Li, H. Chang, J. Yang, W. Luo, and Y. Fu. Visual representation and classification by learning group sparse deep stacking network. *IEEE Transactions on Image Processing*, 27(1):464–476, Jan 2018. 2

[25] H. Liu, M. Yang, Y. Gao, Y. Yin, and L. Chen. Bilinear discriminative dictionary learning for face recognition. *Pattern Recognition*, 47(5):1835–1845, 2014. 2, 8

[26] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *Computer Vision–ECCV 2014*, pages 265–280. Springer Berlin Heidelberg, 2014. 2

[27] J. Lu, G. Wang, and J. Zhou. Simultaneous feature and dictionary learning for image set based face recognition. *IEEE Transactions on Image Processing*, 26(8):4042–4054, Aug 2017. 1, 2, 3

[28] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196, Boston, MA, USA, June 2015. IEEE Computer Society. 1

[29] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, pages 1033–1040. Curran Associates, Inc., 2009. 1, 3

[30] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998. 6

[31] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Om-

nipress. 1

[32] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 5, 2011. 5

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. 7, 8

[34] G. Pereyra, G. Tucker, J. Chorowski, L. Kaiser, and G. E. Hinton. Regularizing neural networks by penalizing confident output distributions. *CoRR*, abs/1701.06548, 2017. 1

[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2014. 1

[36] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. 1

[37] W. Shang, K. Sohn, D. Almeida, and H. Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2217–2225, New York, New York, USA, 20–22 Jun 2016. PMLR. 1

[38] V. Singhal, H. K. Aggarwal, S. Tariyal, and A. Majumdar. Discriminative robust deep dictionary learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5274–5283, Sept 2017. 1, 2

[39] V. Singhal, A. Gogna, and A. Majumdar. Deep dictionary learning vs deep belief network vs stacked autoencoder: An empirical analysis. In *International conference on neural information processing*, pages 337–344. Springer, 2016.

[40] V. Singhal, S. Singh, and A. Majumdar. How to train your neural network with dictionary learning. In *2017 Data Compression Conference (DCC)*, pages 460–460, April 2017. 1, 2

[41] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. 1

[42] Y. Sun, Q. Liu, J. Tang, and D. Tao. Learning discriminative dictionary for group sparse representation. *IEEE Transactions on Image Processing*, 23(9):3816–3828, Sept 2014. 2

[43] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1891–1898, Washington, DC, USA, June 2014. IEEE Computer Society. 1

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA, June 2015. IEEE Computer Society. 1

[45] A. Szlam, K. Gregor, and Y. LeCun. *Fast Approximations to Structured Sparse Coding and Applications to Object Classification*, pages 200–213. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 2

[46] S. Tariyal, A. Majumdar, R. Singh, and M. Vatsa. Deep dictionary learning. *IEEE Access*, 4:10096–10109, 2016. 2

[47] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein. Training neural networks without gradients: A scalable admm approach. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2722–2731. JMLR.org, 2016. 3

[48] G. Urban, K. J. Geras, S. Ebrahimi Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, and M. Richardson. Do Deep Convolutional Nets Really Need to be Deep and Convolutional? *ArXiv e-prints*, Mar. 2016. 1

[49] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. *A Discriminative Feature Learning Approach for Deep Face Recognition*, pages 499–515. Springer International Publishing, Cham, 2016. 1

[50] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, June 2010. 3

[51] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 543–550, Barcelona, Spain, Nov 2011. IEEE Computer Society. 1, 2, 8

[52] M. Yang, L. Zhang, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1601–1604, Sept 2010. 1, 8

[53] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, Feb 2018. 2

[54] M. D. Zeiler and R. Fergus. *Visualizing and Understanding Convolutional Networks*, pages 818–833. Springer International Publishing, Cham, 2014. 1

[55] H. Zhang, Y. Zhang, and T. S. Huang. Simultaneous discriminative projection and dictionary learning for sparse representation based classification. *Pattern Recognition*, 46(1):346–354, 2013. 2

[56] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang. Collaborative representation based classification for face recognition. *CoRR*, abs/1204.2358, 2012. 2, 3

[57] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698, San Francisco, CA, USA, June 2010. IEEE Computer Society. 2