Improve Deep Learning with Unsupervised Objective

Shufei Zhang[1], Kaizhu Huang[1](✉), Rui Zhang[2], and Amir Hussain[3]

[1] Department of EEE, Xi'an Jiaotong- Liverpool University,
111 Ren'ai Road, Suzhou, Jiangsu, People's Republic of China

[2] Department of MS, Xi'an Jiaotong- Liverpool University,
111 Ren'ai Road, Suzhou, Jiangsu, People's Republic of China

[3] Department of Computing Science and Maths, University of Stirling,
Stirling FK9 4LA, Scotland, UK

Abstract. We propose a novel approach capable of embedding the unsupervised objective into hidden layers of the deep neural network (DNN) for preserving important unsupervised information. To this end, we exploit a very simple yet effective unsupervised method, i.e. principal component analysis (PCA), to generate the unsupervised "label" for the latent layers of DNN. Each latent layer of DNN can then be supervised not just by the class label, but also by the unsupervised "label" so that the intrinsic structure information of data can be learned and embedded. Compared with traditional methods which combine supervised and unsupervised learning, our proposed model avoids the needs for layer-wise pre-training and complicated model learning e.g. in deep autoencoder. We show that the resulting model achieves state-of-the-art performance in both face and handwriting data simply with learning of unsupervised "labels".

Keywords: Deep learning; Multi-layer perceptron; Unsupervised learning; Recognition

1 Introduction

Image classification is a very challenging task, due to large amount of intra-class variability, arising from different lightings, misalignment, non-rigid deformations, occlusion, corruptions and different background. To solve such problems, deep neural networks (DNNs) have been considered as the state-of-the-art framework. Especially, convolutional neural network (CNN) achieves a surprising success in visual tasks [4]. The basic idea of DNNs is to exploit a deep structure to extract multiple levels of information from input images, hoping the higher-level information more abstract and more invariant for intra-class variability. However, main drawbacks exist in deep learning models. In particular, DNNs

Published in Liu D, Xie S, Li Y, Zhao D, El-Alfy E-SM (ed.) *Neural Information Processing*. ICONIP 2017, Cham, Switzerland: Springer. Part of the Lecture Notes in Computer Science book series (LNCS, volume 10634). The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-70087-8_74

usually require a large amount of training data so as to learn reliable invariant features. Moreover, the learning for most DNNs is merely supervised with the class label, while important intrinsic structure information of data is usually discarded.

To tackle this problem, auxiliary unsupervised learning can be engaged with the supervised learning in DNNs. For example, Suddarth et al. proposed a method to introduce an auxiliary task to help train a neural network [9]. Hence, the neural network can generalize better. Some other methods are also able to apply both unsupervised and supervised learning simultaneously. However, most of them just exploit unsupervised learning to pretrain models [3]. Recently, Antti et al. proposed a famous model named ladder network (LN) where the unsupervised auxiliary task is to denoise representations at every level of model [7]. The model takes an autoencoder structure with skip connections from the encoder to decoder and the learning task is similar to that in denoising autoencoders but applied to every layer. Although LN achieves state-of-the-art performance, it need train a large amount of parameters to denoise representations at every level of the model. By replacing the convolutional layers of CNN by PCA, binary hashing, and block-wise histograms, a PCANet model was proposed [1]. PCANet receives attention due to its easy and efficient implementation. However, its performance might be limited since this method can be considered as a prepossessing technique for input data.

We propose a novel method termed deep neural networks with unsupervised objective (DNNUO) to generate additional labels (called unsupervised labels). Such labels contain more information (intrinsic structure information of data) rather than class information. With the help of such unsupervised labels, while extracting invariant features for intra-class variability, the latent layers of DNN can better preserve intrinsic structure information; this enables the DNN model a better generalization ability for future data. Compared with existing methods, our proposed model avoids the needs for layer-wise pre-training and complex training for models to reconstruct data itself. In particular, we simply apply PCA on input data and treat the output of PCA as additional unsupervised labels so as to promote better the learning of DNNs. Expanding the label information, the proposed model can efficiently combine both supervised and unsupervised learning in DNNs.

We list the main contributions as follows. (1) To our best knowledge, our proposed method is the first method to generate the unsupervised labels for promoting supervised learning. (2) Our proposed approach is very simple yet effective. Unlike the previous methods introducing too many additional parameters, DNNUO utilises a rather limited number of new parameters. (3) The proposed framework could be readily extended to most supervised learning methods. (4) Our proposed method achieves the state-of-the-art performance in both face and handwriting dataset.

2 Notation and Convolutional Neural Network

Essentially, a convolutional neural network (CNN) is composed by one or more convolutional layers (often with a subsampling layer) followed by one or more fully connected layers as in a conventional multilayer neural network as shown in Fig. 1. The architecture of CNN is designed to take advantages of the $2D$ structure of an input image. This is achieved with local connections and tied weights followed by some forms of pooling which results in translation invariant features.

Suppose a L_2-layer CNN (with L_1 convolutional layers and $L_2 - L_1$ fully connected layers) is trained to perform prediction in a classification task. CNN maps the input matrix to the D-dimension label space. Figure 1 illustrates the structure of a typical CNN, whose optimization problem can be formulated as follows:

$$\begin{aligned} \min_{W,\mathbf{b},K} \quad & L(\mathbf{x}^{L_2},\mathbf{y}) \quad \text{s. t.} \\ X_j^l &= f\Big(\sum_{x_i\in M_j} X_i^{l-1} * K_{ij}^l + \mathbf{b}_j^l\Big), l=1,...,L_1 \\ \mathbf{x}^{L_2+r+1} &= f(\mathbf{x}^{L_2+r} W_{L_1+r+1} + \mathbf{b}_{L_1+r+1}), r=1,...,L_2-L_1-1 \\ \mathbf{x}^{L_2} &= \mathbf{x}^{L_2-1}W_{L_2}+\mathbf{b}_{L_2} \end{aligned} \tag{1}$$

where, the matrix X_i^0 represents the i^{th} input data matrix ($X_i^0 \in M_0$ where M_0 is the input set). X_i^{l-1} indicates the i^{th} feature map of the $(l-1)^{th}$ convolutional layer of CNN (where $l = 1, 2, \ldots, L_1$). $\mathbf{x}^{L_1+r}$ denotes the output of r^{th} fully connected layer and $\mathbf{x}^{L_2}$ represented the output of the CNN. y represents the class label with D dimensions. K_{ij}^l is the convolutional kernel with input feature map X_i^{l-1} and output feature map X_j^l. $f(.)$ is the activation function. In this paper, we engage element-wise sigmoid function $\sigma(\cdot)$. For each element x of matrix, the sigmoid function is defined as $\sigma(x) = \frac{1}{1+\exp(-x)}$. $L(\cdot)$ is the cross entropy loss.

In NN, the sigmoid function is usually exploited to perform the non-linear transformation and it can be also replaced by other functions such as $\max(0, x)$ and $\tanh(x)$.

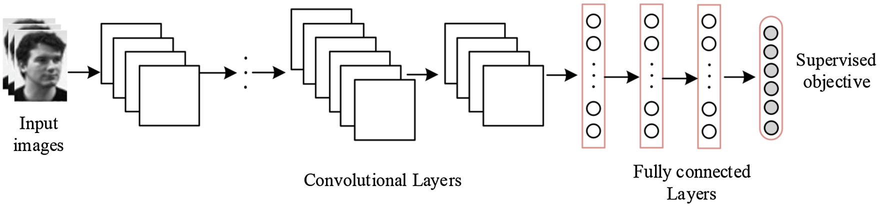

Fig. 1. The structure of conventional convolutional neural network

3 Deep Learning with Unsupervised Objective

This section will detail our proposed method deep neural network with unsupervised objective (DNNUO). We will first present the structure of the proposed model and then introduce the optimization algorithm.

3.1 Network Structure

The structure of DNNUO is plotted in Fig. 2. As can be seen, the structure of DNNUO is composed of two parts, supervised part (traditional CNN) and unsupervised part (the unsupervised label generator with linear transformation). We will detail these two parts in turn.

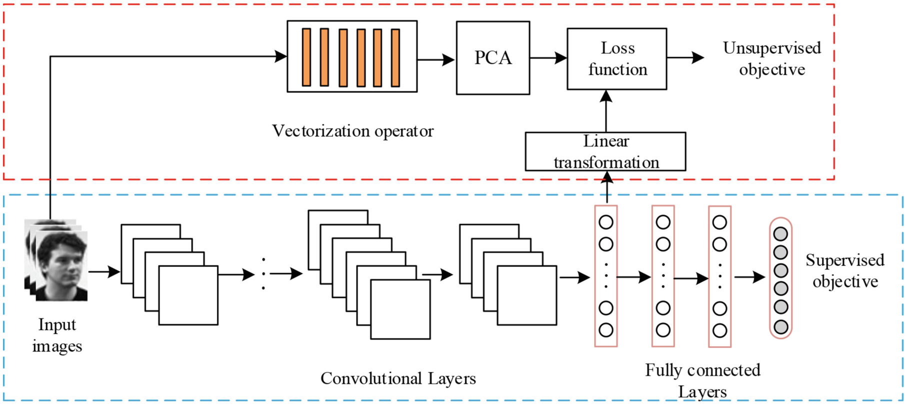

Fig. 2. Structure of deep neural network with unsupervised objective

Supervised Part: We take the supervised learning model as a starting point and improve it further by adding the unsupervised objective. The supervised models could be any DNNs, however we use the CNN and the fully connected MLP network in the paper. We will mainly describe the extension of CNN while omitting details of MLP, since they are much similar. In Fig. 2, our proposed model is illustrated on a CNN with 6 convolutional layers and 3 fully connect layers in blue dashed rectangular with dropout. We also build the fully connected MLP network with 6 layers. Both of these feedforward frames are to extract more invariant features for the same class.

Unsupervised Part: The whole structure of unsupervised part is shown in the red dashed rectangular of Fig. 2. It can be divided into three steps: vectorization operator, PCA operator, and linear transformation. First, the input images are vectorized by vectorization operator. Then, the vectorized data are fed to PCA operator (PCA model is trained by the whole dataset including training and test set) to extract the principle components of data. Finally, the latent features of CNN pass through the linear transformation and a loss is calculated with respect to the outputs of PCA operator (unsupervised labels). In this paper, the loss for unsupervised objective is given by square error. Our aim is to minimize this loss while implementing the supervised learning.

The purpose of using the unsupervised label generator is to extract the intrinsic information of data; this would encourage the DNN to learn both the information about data itself (unsupervised information) as well as the discriminative information (supervised information).

3.2 Model Formulation

After introducing the structure of our proposed model, we now describe the model formulation. The main optimization problem can be formulated as:

$$\begin{aligned}
\min_{W,\mathbf{b},k} L =&\ L_1(\mathbf{x}^{L_2}, \mathbf{y}) + \frac{1}{2}\lambda \|T(\mathbf{x}^{L_1+1}) - G(\mathbf{x}^0))\|^2 \quad \text{s. t.} \\
X_j^l =&\ f(\sum_{\mathbf{x}_i \in M_j} X_i^{l-1} * K_{ij}^l + \mathbf{b}_j^l), l = 1, ..., L_1 \\
\mathbf{x}^{L_2+r+1} =&\ f(\mathbf{x}^{L_2+r} W_{L1+r+1} + \mathbf{b}_{L1+r+1}), r = 1, ..., L_2 - L_1 - 1 \\
\mathbf{x}^{L_2} =&\ \mathbf{x}^{L_2-1} W_{L_2} + \mathbf{b}_{L_2}
\end{aligned} \tag{2}$$

where X^0 is the input image. X^{L_1+r} indicates the i^{th} feature map of the $(l-1)^{th}$ convolutional layer of CNN (where $l = 1, 2, \ldots, L_1$); $\mathbf{x}^{L_1+r}$ denotes the output of r^{th} fully connected layer and $\mathbf{x}^{L_2}$ represents the output of the CNN; y represents the class label; k_{ij}^l denotes the convolutional kernel with input feature map X_i^{l-1} and output feature map X_j^l, and $f(.)$ is the activation function. In addition, $L_1(\cdot)$ is the cross entropy loss. In this paper, we exploit the element-wise sigmoid function $\sigma(\cdot)$.

The objective function L of this optimization problem consists of two terms. The first term is the supervised objective for CNN with the purpose of encouraging learning the invariant features; the second term is the unsupervised objective that encourages learning intrinsic information from data. The hyper-parameter λ is to balance such the two objectives. For the unsupervised objective, $T(\cdot)$ is the linear transformation for the latent feature of CNN and $G(\cdot)$ denotes the unsupervised label generator.

3.3 Optimization

For solving the above optimization problem, we can still rely on the traditional Back Propagation (BP) algorithm, since the gradients with respect to the parameters can be easily computed. The gradients can be seen as a combination of the gradients from the conventional CNN model and the gradients from the unsupervised objective. The gradients for the output of hidden layer $L_1 + 1$ can be given as:

$$\frac{\mathrm{d}L}{\mathrm{d}\mathbf{x}^{L1+1}} = \frac{\mathrm{d}L_1(\mathbf{x}_2^L, y)}{\mathrm{d}\mathbf{x}^{L1+1}} + \lambda \frac{\mathrm{d}T(\mathbf{x}^{L_1+1})}{\mathrm{d}\mathbf{x}^{L1+1}} (T(\mathbf{x}^{L1+1}) - G(\mathbf{x}^0)) \tag{3}$$

4 Experiments

In this section, we conduct a series of experiments on our proposed model including face and handwriting data.

4.1 Experimental Setup

The face dataset contains totally 195 images for 15 different persons [2]. Each person has 13 horizontal poses from −90 to 90° with interval 15°. It is noted that the test set shares very different pose from the training set which makes the problem very challenging. We follow the experimental setup exactly in [11] so as to compare with different algorithms fairly. We evaluate our proposed model against other state-of-art models including the traditional CNN, Field Bayesian Model (FBM) [11], conventional MLP network, and SVM.

The handwriting dataset is a subset of the CASIA-OLHWDB dataset [5]. 100 writers are chosen from number 1101 to 1200 and the first 30 classes are selected. The purpose of using a subset is to speed up the training for various models. Nonetheless, the training set contains still 2997 characters, with each writer writing about 30 isolated characters. On the other hand, about 288 characters were extracted from another set containing handwritten texts of these writers, which we adopted as the test set. We implement the training set on training models and then report the performance on the test set. We have also implemented the conventional MLP, Linear and nonlinear Support Vector Machine with the rbf kernel function (in short, linear-SVM, and rbf-SVM), conventional CNN, nearest class mean and Modified Quadratic Discriminant Function (MQDF) on this handwriting data.

The same base framework Alexnet [4] is adopted with six convolutional layers and three fully connect layers in both the datasets. The hyper-parameter λ is adjusted using cross validation.

4.2 Face Recognition with Different Pose

We have done a series of preprocessing including resizing the images to 48×36 and then reducing the dimension to 100 with Principal Component Analysis (PCA). For CNN, we resize the images to 224×224. Table 1 reports the performance (recognition rate) of different models.

As observed, our novel CNN-DNNUO achieves the best performance with 94.67%. More specifically, the proposed DNNUO significantly improves the performance of CNN from 90.67 to 94.67. On the other hand, Fisher Discriminant Analysis (FDA) is the state-of-the-art algorithm for face recognition, which only achieved the error rate of 69.33%. Moreover, the other approaches such as the bilinear model, the style mixture model, the FBM and conventional Neural Network are obviously worse than our proposed DNNUO.

Table 1. Recognition rates of different models on face data. The proposed CNN-DNNUO significantly outperforms the other models.

Classifier	Accuracy (%)
Bilinear (Field) [10]	60.00
Style mixture (Singlet) [8]	70.00
Style mixture (Field) [8]	73.33
Nearest class mean [11]	60.00
FDA [11]	69.33
FBM [11]	74.67
linear-SVM	84.00
rbf-SVM	85.33
MLP	81.33
CNN	90.67
CNN-DNNUO	**94.67**
MLP-DNNUO	86.67

4.3 Handwriting Classification

We exploit the benchmark 8-direction histogram feature extraction method to generate for each character image a feature of 512 dimension [6]. These features are reduced to 160 dimensions by FDA and further to 50 by PCA in order to speed up the training. For CNN and DNNUO, the original images are directly resized to 224 × 224.

Table 2. Error rates of different models on handwriting data. Our proposed CNN-DNNUO model achieves the best performance.

Classifier	Accuracy (%)
Nearest class mean	94.44
MQDF	94.44
3-layer perceptron	97.22
FBM	95.49
$\text{SVM}_{linear-kernel}$	96.53
$\text{SVM}_{RBF-kernel}$	96.53
CNN (Alexnet)	98.61
MLP-DNNUO	98.26
CNN-DNNUO	**98.96**

Table 2 shows the recognition rates of various models. Once again, the proposed CNN-DNNUO achieves the best performance of 98.96%. It is superior to the other models including the conventional CNN and the SVM models. Moreover, our CNN-DNNUO also outperforms the MQDF, the state-of-the-art classifier in Chinese character recognition.

5 Conclusions

In this paper, we proposed a novel deep learning framework with unsupervised objective (DNNUO) which can appropriately take advantages of intrinsic information of data. Specifically, we built a novel network with two parts: supervised part and unsupervised part. We proposed to connect the unsupervised part and one hidden layer of CNN, which are exploited to deliver the unsupervised knowledge. We developed a new objective function with additional unsupervised term and modified the stochastic optimization algorithm, which can efficiently optimize the proposed DNNUO model. We conducted experiments on two databases including face and handwriting data. Experimental results showed that our proposed model achieves the best performance on both the data sets compared with the other competitive models.

Acknowledgement. The paper was partially supported by National Natural Science Foundation of China (NSFC) under grant no.61473236, Natural Science Fund for Colleges and Universities in Jiangsu Province under grant no. 17KJD520010, Suzhou Science and Technology Programme under grant no. SYG201712, SZS201613, the UK Engineering and Physical Science Research Council (EPSRC) grant no. EP/M026981/1, and also the UK Royal Society of Edinburgh (RSE) and NSFC under joint-project grant no. 61411130162.

References

1. Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y.: Pcanet: a simple deep learning baseline for image classification? IEEE Trans. Image Process. **24**(12), 5017–5032 (2015)
2. Gourier, N., Hall, D., Crowley, J.: Estimating face orientation from robust detection of salient facial features. In: International Conference on Pattern Recognition (ICPR) (2004)
3. Hinton, G., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
5. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: Casia online and offline chinese handwriting databases. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 37–41 (2011)
6. Liu, C.L., Zhou, X.D.: Online japanese character recognition using trajectory-based normalization and direction feature extraction. In: Tenth International Workshop on Frontiers in Handwriting Recognition (2006)
7. Rasmus, A., Berglund, M., Honkala, M., Valpola, H., Raiko, T.: Semi-supervised learning with ladder networks. In: Advances in Neural Information Processing Systems, pp. 3546–3554 (2015)
8. Sarkar, P., Nagy, G.: Style consistent classification of isogenous patterns. IEEE Trans. Pattern Anal. Mach. Intell. **27**(1), 88–98 (2005)

9. Suddarth, S.C., Kergosien, Y.L.: Rule-injection hints as a means of improving network performance and learning time. In: Almeida, L.B., Wellekens, C.J. (eds.) EURASIP 1990. LNCS, vol. 412, pp. 120–129. Springer, Heidelberg (1990). doi:10.1007/3-540-52255-7_33
10. Tenenbaum, J.B., Freeman, W.T.: Separating style and content with bilinear models. Neural Comput. **12**(6), 1247–1283 (2000)
11. Zhang, X.Y., Huang, K., Liu, C.L.: Pattern field classification with style normalized transformation. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 1621–1626 (2011)