A Novel Image Feature Extraction Algorithm Based on the Fusion AutoEncoder and CNN

Ke Xu\textsuperscript{a,}\textsuperscript{c}, Wenhan Long\textsuperscript{a}, Yuan Sun\textsuperscript{b} and Yichao Lin\textsuperscript{a}

\textsuperscript{a} College of Computer Science, South-Central University for Nationalities, Wuhan, China
\textsuperscript{b} Hubei University of Traditional Chinese Medicine, Wuhan, China.
\textsuperscript{c} Shenzhen Research Institute of Huazhong University of Science and Technology, Shenzhen, China

Abstract. Deep learning method has very excellent ability of image feature extraction. In order to get rid of disadvantages of traditional methods require a priori knowledge, this paper proposed an image feature extraction algorithm based on the fusion AutoEncoder and convolutional neural networks (CNN). The method introduces a fast sparsity control technique to AutoEncoder and utilizes AutoEncoder to train the basic elements of image and initialize the convolution kernel of CNN. Meanwhile, the algorithm adds filtering mechanism to the CNN network to keep the sparsity of output characteristics. The results of experiments point out that this method has achieved good performance on the Minist handwritten digital library and the Yale face database. Furthermore, the advanced experimental outcomes indicate that the feature extraction model included the filtering technique is more effective than the model without filtering mechanism by using cross-validation with T Test.

1. Introduction

Deep learning is a branch of machine learning which attempts to learn high-level representations of data by utilizing hierarchical structures. Researches in this area attempt to make better representations and create models to learn these abstractions from large-scale unlabeled data. It has been widely applied in traditional artificial intelligence fields like computer vision \cite{1}, automatic speech recognition \cite{2}, natural language processing \cite{3} and bioinformatics \cite{4}, where they have been shown to produce state-of-the-art results on diverse problems. Various deep learning methods are applied to unsupervised learning tasks. This is an important benefit because unlabeled data are usually more abundant than labeled data.

Recently many deep learning approaches have been extensively investigated and discussed for different applications in various domains \cite{5, 6}. Schmidhuber emphasized shallow and deep learners are distinguished by the depth of their credit assignment paths which are chains of possibly learnable, causal links between actions and effects \cite{5}. Ref. \cite{6} proposed to examine some of these challenges centering on the questions of scaling deep learning algorithms to much larger models and datasets and presented a few forward-looking research directions. Deep networks have achieved good performance on computer vision tasks for they can extract appropriate features \cite{7}. Deep learning approaches have been widely adopted by different researchers and achieved top accuracy scores in the 2014 ImageNet competitions \cite{8}. Generally, these deep learning techniques can be divided into four categories including Autoencoder, convolutional neural networks (CNN), restricted boltzmann machines (RBM) and sparse coding respectively.
In our paper, we focus on the ability of Autoencoder to pre-training CNN and discuss the filtering mechanism is useful or not. Our main contributions can be summarized as follows:

(1) We propose an algorithm for image feature extraction built on the CNN initialized by the Autoencoder rather than random weight. The method utilizes Autoencoder to train the basic elements of image and initialize the convolution kernel of CNN. It achieves better performance than the random initialized method on image classification while the number of layers of the CNN grows rapidly.

(2) We add the filtering mechanism to the CNN network to conquer the sparse problem of image features extraction according to the following assumption. We put forward an assumption on image features abstraction that the basic blocks in the image is only activated on a few positions and the basic blocks which appear frequently in the representative processing is useless or harmful for features extraction.

2. Related Works
In recent years, deep neural networks have won numerous contests in pattern recognition and machine learning. Unlike shallow learning models, for example, support vector machine (SVM) and softmax, deep learning is about learning multiple levels of representation and abstraction that helps to make sense of data. Some theoretical investigations demonstrate that deep models are more concise and expressive than shallow models in representing most learning functions [1,5,6]. The deep learning could be utilized to solve hard learning problems with the benefit of its effectiveness, for example, identifying the semantic class of images from low-level visual features. Different from previous image classification models, the bilinear deep belief network (BDBN) focuses on providing human-like judgment by simulating the mechanism of the human visual system and the procedure of intelligent perception [9]. Hayat introduced a deep learning framework without the assume that images of a set lie on a certain geometric surface and designed a template deep reconstruction model (TDRM) to discover the underlying geometric structure automatically [10]. In [11], this paper proposed a PCA network combined with binary hash and block histogram for image classification, experiments indicated the PCANet would be regarded as the potential baseline for image classification and object recognition thought it was a concise but highly competitive model. To address the problem of lack complete data in image recognition, they designed field effect bilinear deep networks based on field effect RBM to model the reliability of the delivered information referred to the availability of the features [12].

3. Theory and Method
The CNN is usually initialized by random weight parameters. However, it gradually increases the difficulty of network training when the number of layers of the network grows rapidly. For the CNN shares the weight coefficient, so the input data can be used as the original image and etc.. The Autoencoder has the ability to quickly extract the basic components of the image. In this paper, we propose a method based on Autoencoder pre-training the CNN network and using Autoencoder to set the CNN with a better initial value according to the advantages of these two approaches. We proceed to discuss the initialization network can be divided into the feature space as well as for feature extraction.

The features extracted method proposed in this section is described by the pseudo-code in Algorithm 1, the input and output of which is original image and features of the image respectively. The pseudo-code of the key function for initializing the convolutional kernel of CNN is described in Algorithm 2.
We select nine training and testing group according to the proportion of training sample accounting for 10%, 20%, ..., and 90% respectively for each dataset. By using the SVM and Softmax classifier, the results are shown in from Fig. 5 to Fig. 14, where the "proportion" represents the proportion of the training set for total sample. As shown in Fig. 5, the difference on the Yale data is very obvious in the two models when the training set is small, and the difference reduced gradually while the proportion of the training samples increased rapidly. The MWF and MWOF achieves the best results on Yale data when the training sample accounting for 90%. We can see from Fig. 6, the difference of the two models in the 9 division of the training set on the MIT data do not change obviously, and the MWF model achieve the best results while the proportion is 0.8 , meanwhile, the MWOF model get the best results when the proportion of the training set is 0.9. As shown in Fig. 7, on the Minst digital library, the difference of training samples proportion have leaded big impact to the two models' performance, for example, the MWF and MWOF have achieved the best results when the training sample accounted for 90%. The comprehensive analysis shows that the three kinds of data have achieved good results when the proportion of the training set is more than 80%, and the error rate of MWF model is lower than MWOF model. Furthermore, the experimental outcomes indicated that Autoencoder can initialize convolution kernel initialization of CNN with better parameters and filtration mechanism reduce the error rate. In Fig. 8, the three datasets case big difference on the MWOF model when training sample proportion is less than 60%. For example, the Yale data achieved better result than MIT or Minst data on the MWOF model when the proportion more than 80%. As shown in Fig. 9, it can be found that the three datasets on MWOF model have big differences in case of the proportion of the training sample is less than 60% . The results of the Yale obtained better than the other two datasets when the proportion is more than 60%. Those results and analysis show that the results obtained by Yale face images are better than the other two datasets, which show that our proposed method can process large images with rich information. For the Softmax classifier, the trend of the MWOF and the MWF on the Yale dataset is similar shown in Fig. 10, but the error rate of the latter one is lower than the former one. As shown in Fig. 11, it can make a conclusion that the MWF achieves better performance than the MWOF on the MIT dataset according to the position of the two polyline. The polyline of both of the
MWOF and the MWF on the Minist dataset classified with Softmax are fluctuant according to the illustration in Fig. 12. But the growing trend of the two polylines change consistently, the best results of both are obtained while the proportion value is 0.9. As shown in Fig.13, the three polylines of the MWOF model intersect with each other while the proportion value varies from 0.5 to 0.8. The MWOF of Yale dataset achieves the best result when the proportion is greater than 0.8, which is consistent with the classification result for the SVM classifier. We can see from Fig. 14, the MWF polylines of those three datasets intersect with each other near the proportion value of 0.6. However, the error rate of those datasets are distinguished obviously while the proportion is over 0.7 and are the lowest level when proportion equals to 0.9. In summary, we can conclude from the experimental results that the Autoencoder can better initialize the convolution kernel of CNN and the filtering mechanism could reduce the error rate of classification.

In order to analyze the detail change in the local area, we proceed to select the two data group, for example, one group includes training set (TR) accounting for 70% and test set (TE) accounting for 30% (TR=70%, TE=30%), and the other group includes training set (TR) accounted for 80% and test set (TE) accounted for 20% (TR=80%, TE=20%). Then, this paper introduces the feature extraction model with filtering mechanism (MWF) and the feature extraction model without filtering mechanism (MWOF) to measure the performance of the proposed method. The error rates of difference classifiers based on the SVM and the Softmax are shown in Table 2 and Table 3.

| Dataset name | (TR=70%,TE=30%) | (TR=80%,TE=20%) |
|--------------|------------------|------------------|
|              | MWF   | MWOF  | MWF   | MWOF  |
| Minist       | 17.00  | 20.00 | 14.00  | 15.33 |
| Yale         | 10.13  | 15.33 | 8.66   | 11.03 |
| MIT          | 12.09  | 15.00 | 10.68  | 12.66 |

| Dataset name | (TR=70%, TE=30%) | (TR=80%, TE=20%) |
|--------------|------------------|------------------|
|              | MWF   | MWOF  | MWF   | MWOF  |
| Minist       | 20.00  | 24.35 | 16.48  | 19.00 |
| Yale         | 12.46  | 14.28 | 10.44  | 14.18 |
| MIT          | 16.80  | 17.32 | 13.36  | 14.89 |

Table 2. Classified results of SVM. (%)

Table 3. Classified results of Softmax. (%)

| Dataset name | (TR=70%, TE=30%) | (TR=80%, TE=20%) |
|--------------|------------------|------------------|
|              | MWF   | MWOF  | MWF   | MWOF  |
| Minist       | 12.76  | 16.00 | 15.08  | 17.48 |
| Yale         | 4.97   | 6.57  | 9.83   | 12.65 |
| MIT          | 10.68  | 13.00 | 11.54  | 14.00 |

Table 4. The best classified results of SVM and Softmax. (%)

As shown in Table 3 and Table 2, the feature extraction model proposed in this paper based on deep learning to achieve better results on these three datasets. The best error rate of the Yale face is 0.0866. The best error rate of Minist is 0.14, and the best error rate of MIT is 0.1068. Those results...
indicate that the extracted features can be used to identify these data better. Therefore, we can conclude that the initialization of convolution method is feasible and effective. The error rate of MWF is smaller than MWOF on the two cases of (TR=70%, TE=30%) and (TR=80%, TE=20%), thus, the filtering mechanism has reduced the error rate. The best results of the two classifiers including SVM and Softmax are shown in Table 4, the SVM classifier achieves better performance than the Softmax on those three datasets. The best error rate of classification on Minst, Yale and MIT is 0.127, 0.0497 and 0.1068 respectively by using SVM.

In order to verify the generalization performance of the proposed method, this paper has randomly selected 1000 samples from the Minst database and carries out 5*2 cross-validation by t-Test [26] for those randomly disrupting samples. According to the results using the T Test, we proceed to evaluate if there is obviously difference between the MWF and the MWOF and to analyze which model is better. The error rate of classification using cross-validation are shown in Table 5.

Table 5. The results of cross-validation (%)

| Number of times | MWF 1-fold | MWF 2-fold | MWOF 1-fold | MWOF 2-fold |
|-----------------|------------|------------|-------------|-------------|
| 1               | 9.50       | 9.33       | 15.00       | 14.67       |
| 2               | 9.66       | 10.50      | 14.33       | 15.80       |
| 3               | 9.66       | 10.00      | 15.63       | 15.00       |
| 4               | 11.50      | 10.67      | 14.34       | 15.90       |
| 5               | 10.33      | 9.80       | 15.67       | 15.50       |

Each 2-fold cross-validation contains four types of error rate. According to the t-Test formula in [26], the cross validation value is 6.385, which is greater than the value of t_{0.05/2,5} equal 2.570. From the outcome and analysis above, we make a conclusion that there is significant difference between the MWF and the MWOF model and the average error rates of the former is lower. In addition to, the generalization ability of the MWF is better than the MWOF.

5. Conclusion
This paper combines the advantages of CNN and Autoencoder and utilizes the Autoencoder to initialize CNN convolution kernel. The experimental results show that the proposed method has achieved good result on three kinds of different data sets. Furthermore, the outcomes indicate that it is feasible to use Autoencoder to initialize CNN to extract image feature. This also provides a way to train the multi-layer of CNN step by step. Last but not least, we introduce the filter layer to the CNN, the experimental results show that this feature extraction model based on deep learning has achieved a good performance. The best error rate of classification on Minst, Yale and MIT is 0.127, 0.0497 and 0.1068 respectively. So we can conclude that the filtering mechanism reduces the error rate. We also can make a conclusion that the proposed feature extraction method based on deep learning can be used to divide the feature space as good as possible.

6. Acknowledgments
This research is supported by the Science and Technology Development Center Fund of the Ministry of Education of China (2018A02037), and the Industry-university Collaborative Education Project (Baidu Project, the second batch in 2018) for Higher Education Department of the Ministry of Education (A preliminary study on the training mode of artificial intelligence innovative talents based on the second class learning based on Paddle Paddle).

7. References
[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, pp.1106-1114, 2012.
[2] Z. Huang, S. M. Siniscalchi, C. H. Lee, “A unified approach to transfer learning of deep neural
networks with applications to speaker adaptation in automatic speech recognition,” Neurocomputing, vol.218, pp.448-459, 2016.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality,” Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, pp.3111-3119, 2013.

[4] S. Cheng, M. Guo, C. Wang, X. Y. Liu, Y. Liu, X. J. Wu, “MiRTDL: a deep learning approach for miRNA target prediction,” IEEE/ACM Trans. Comput. Biology Bioinform, vol.13, no.6, pp.1161-1169, 2016.

[5] J. Schmidhuber, “Deep learning in neural networks: an overview,” Neural Networks, vol.61, pp.85-117, 2015.

[6] Y. Bengio, “Deep learning of representations: looking forward,” Proceedings of the 1th International Conference on Statistical Language and Speech Processing, Tarragona, Spain, pp.1-37, 2013.

[7] Y. LeCun, “Learning invariant feature hierarchies,” ECCV workshop, 2012.

[8] <http://www.image-net.org/challenges/LSVRC/2014/results>.

[9] Y. Chen, H. Jiang, C. Li, X. Jia, P. Ghamisi, “Deep feature extraction and classification of hyperspectral images based on convolutional neural networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232-6251, 2016.

[10] L. Wang, T. Liu, G. Wang, K. L. Chan, Q. X. Yang, “Video tracking using learned hierarchical features,” IEEE Transactions on Image Processing, vol. 24, no. 4, pp. 1424-1435, 2015.

[11] J. Zahálka, S. Rudinac, M. Worring, “Interactive multimodal learning for venue recommendation,” IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2235-2244, 2015.

[12] Y. Gong, L. Wang, R. Guo, S. Lazebnik, “Multi-scale orderless pooling of deep convolutional activation features,” European Conference on Computer Vision, vol. 7, pp.392-407, 2014.