SCIENCE CHINA
Information Sciences

February 2020, Vol. 63 120111:1–120111:2
https://doi.org/10.1007/s11432-019-2722-3

• LETTER •
Special Focus on Deep Learning for Computer Vision

Discriminative stacked autoencoder for feature representation and classification

Yiping GAO, Xinyu LI & Liang GAO*

School of Mechanical, Huazhong University of Science and Technology, Wuhan 430072, China

Received 10 August 2019/Revised 30 September 2019/Accepted 1 November 2019/Published online 14 January 2020

Citation Gao Y P, Li X Y, Gao L. Discriminative stacked autoencoder for feature representation and classification. Sci China Inf Sci, 2020, 63(2): 120111, https://doi.org/10.1007/s11432-019-2722-3

Dear editor,
Recently, deep learning (DL) has become a hot research topic and as one of the most well-known DL models, stacked autoencoder (SAE) [1] has received increasing attention. In SAE, layer-wise pretraining is the basic mechanism for automatic feature extraction and it can also avoid gradient vanishing while constructing deep architectures. Based on this pretraining, SAE has achieved outstanding performance in several applications such as fault diagnosis [2], medical image [3] and sense image [4].

Previous research on SAE has considerably improved feature representation and classification tasks [5], but most of them had two drawbacks. First, the essential mechanism of layer-wise pretraining, which is unsupervised, has not yet been improved. Thus, a sample label cannot be fully utilized and it cannot guide feature learning. Second, the learned feature is indiscriminative, and this influences the classification results of SAE. Therefore, these drawbacks limit the performance of SAE.

Method. To overcome these drawbacks, we proposed a new discriminative stacked autoencoder (DSA) that uses hybrid pretraining to replace the original unsupervised one. The proposed DSA is based on a convolutional autoencoder and it contains two steps: hybrid pretraining and global finetuning. The purpose of hybrid pretraining is to learn an abstract and discriminative feature by maximizing the inter-class similarity, while minimizing the intra-class similarity. Hybrid pretraining contains two stages. In the first stage, SAE is trained by minimizing the reconstruction error. This stage is unsupervised and an abstract feature h is learned. In the second stage, the centroid target h^* is first found using the maximum inter-class similarity. The feature is finetuned under supervision to approach its corresponding centroid target, h^*, by minimizing the intra-class similarity. Based on the learned feature, a multi-layer perceptron is connected as the classifier and the whole network is globally finetuned to minimize the classification error E_c. The diagram of the proposed DSA is presented in Figure 1(a).

Assuming an input x, x' denotes the output of SAE, and the object function E_r of the unsupervised pretraining is defined as

$$E_r = \sum \|x - x'\|^2 . \tag{1}$$

After the unsupervised pretraining, a supervised pretraining step is applied for feature finetuning. In this stage, the centroid target h^* is selected by maximizing the inter-class similarity. Assuming h_j is the learned feature of class j, the corresponding centroid target h_j^* is selected from h_j, which exhibits the maximum inter-class similarity. It can be expressed as

$$h^* = \arg\max \sum_J \left\| h_j^* - h_{-j}^* \right\|^2 , \tag{2}$$

where J is the number of classes and $h_{-j}^* \cup h_j^* = h^*$. The optimization of h^* is a combinatorial opti-

[*] Corresponding author (email: gaoliang@mail.hust.edu.cn)

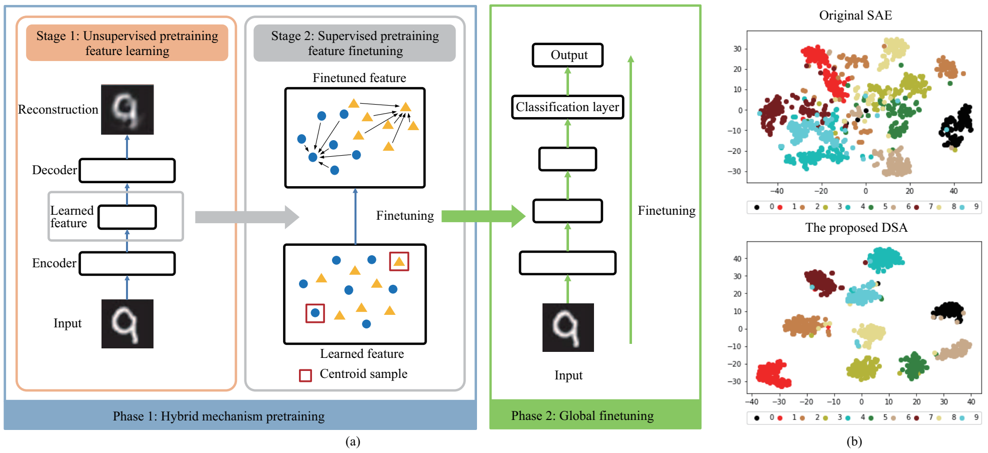

Figure 1 (Color online) (a) The diagram of the proposed DSA. (b) The data visualization of the learned feature; the top shows the feature learned from the original SAE, and the bottom is the feature learned from the proposed DSA.

mization and it is searched by a genetic algorithm. When h^* is found, the feature is finetuned under supervision to approach its corresponding centroid target. Because h^* exhibits the maximum inter-class similarity and this finetuning is meant to minimize the intra-class similarity, this process can map features from different categories into a discriminative space, in which features in the same category have a contracted distribution. The object function of the feature finetuning is expressed as

$$E_d = \sum_{j \in J} \sum_{i \in n_j} \left\| h_j^* - h_i \right\|^2, \tag{3}$$

where h_i and n_j denote the training sample and sample number in the j-th class, respectively.

After the hybrid pretraining, the whole network is globally finetuned to minimize the classification error. The object function E_c of the global finetuning is defined as

$$E_c = -y \log y' - (1-y) \log (1-y'), \tag{4}$$

where y is the sample label and y' is the classified label.

To evaluate the algorithm's performances, we test the proposed method on MNIST and CIFAR-10. The experiments include data visualization, supervised learning, and semi-supervised learning. The data visualization is based on MNIST and the result is presented in Figure 1(b). This result indicates that the proposed DSA can learn a discriminative feature distribution. Based on the learned feature, the proposed DSA achieves an error rate of 0.39% on MNIST and 9.96% on CIFAR-10. Additionally, the error rates represent a 0.32% and 11.84% improvement from the original SAE. In the semi-supervised learning task, the proposed DSA achieves 5.22%, 2.48%, and 0.81% error rates on MNIST with 300, 1000, and 10000 labelled samples, respectively.

Conclusion. In SAE, layer-wise pretraining is an important mechanism. However, this pretraining is unsupervised; therefore, the label is not fully utilized and the learned feature might be indiscriminative. These conditions can influence the model's performance. To overcome this problem, we propose a discriminative stacked autoencoder (DSA) whose main contributions can be summarized as follows. First, hybrid pretraining is introduced to replace the unsupervised pretraining in SAE. Second, the label can be fully utilized in the pretraining phase. Finally, the experimental results suggest that the proposed DSA can learn a discriminative feature representation, and the classification results are improved as a result of the learned feature.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 51721092), Natural Science Foundation of Hubei Province (Grant No. 2018CFA078), and the Program for HUST Academic Frontier Youth Team (Grant No. 2017QYTD04).

References

1 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507

2 Gao Y P, Gao L, Li X Y, et al. A zero-shot learning method for fault diagnosis under unknown working loads. J Intell Manuf, 2019, 3: 1–11

3 Mehta J, Majumdar A. RODEO: robust DE-aliasing autoencoder for real-time medical image reconstruction. Pattern Recogn, 2017, 63: 499–510

4 Zhu Z T, Wang X G, Bai S, et al. Deep learning representation using autoencoder for 3D shape retrieval. Neurocomputing, 2016, 204: 41–50

5 Fan Y J. Autoencoder node saliency: selecting relevant latent representations. Pattern Recogn, 2019, 88: 643–653