ROBUST FEATURE LEARNING BY STACKED AUTOENCODER WITH MAXIMUM CORRENTROPY CRITERION

Yu Qi[1,2], *Yueming Wang*[*1,2], *Xiaoxiang Zheng*[1,3], *and Zhaohui Wu*[2]

[1]Qiushi Academy for Advanced Studies, Zhejiang University, Hangzhou, China
[2]Department of Computer Science, Zhejiang University, Hangzhou, China
[3]Department of Biomedical Engineering, Zhejiang University, Hangzhou, China

ABSTRACT

Unsupervised feature learning with deep networks has been widely studied in the recent years. Despite the progress, most existing models would be fragile to non-Gaussian noises and outliers due to the criterion of mean square error (MSE). In this paper, we propose a robust stacked autoencoder (R-SAE) based on maximum correntropy criterion (MCC) to deal with the data containing non-Gaussian noises and outliers. By replacing MSE with MCC, the anti-noise ability of stacked autoencoder is improved. The proposed method is evaluated using the MNIST benchmark dataset. Experimental results show that, compared with the ordinary stacked autoencoder, the R-SAE improves classification accuracy by 14% and reduces the reconstruction error by 39%, which demonstrates that R-SAE is capable of learning robust features on noisy data.

Index Terms— Unsupervised feature learning, stacked autoencoder, correntropy, deep learning

1. INTRODUCTION

Unsupervised feature learning algorithms aim to find good representations for data, which can be used for classification, reconstruction, visualization and so on. Recently, deep networks such as stacked autoencoders (SAE) and deep belief networks (DBN) have shown high feature learning performance that matches the current state-of-the-art [1, 2, 3, 4].

Despite the progress, robust feature learning is still faced with challenges due to noise and outliers which are commonly appeared in the real-world data. In order to improve the anti-noise ability of the deep networks, efforts have been made. Vincent et al. [5, 6] modified the traditional stacked autoencoder to learn useful features from corrupted data and developed the stacked denoising autoencoder (SDAE). By corrupting the input data and using denoising criterion, the SDAE could learn robust representations and achieve good performance under different types of noises. The SDAE model was then extended by Xie et al. [7] with sparse coding technique (spSDAE). With the reconstruction loss regularized by a sparsity-inducing term, better denoising performance was achieved. Although the existing methods show strength under some noises such as Gaussian noise, they would be fragile in case that the data contain large amounts of outliers. The reason lies that most models are based on mean square error (MSE) criterion which would be sensitive to outliers [8, 9].

Recently, correntropy was proposed as a localized similarity measure based on information theoretic learning (ITL) and kernel methods [10]. It is insensitive to outliers compared with MSE and has been successfully utilized for cost function design in non-Gaussian signal processing [11, 12]. Jeong et al. [13] extended the minimum average correlation energy (MACE) filter to non-linear with correntropy to obtain better distortion tolerance. He et al. [14] proposed a robust principal component analysis method based on maximum correntropy criterion to achieve high performance under outliers. These studies show that, correntropy is robust to outliers so that it is promising for robust algorithm design.

Inspired by the success of correntropy-based approaches in outlier suppression, this paper proposes a robust stacked autoencoder (R-SAE) model with maximum correntropy criterion (MCC). The R-SAE method improves the anti-noise ability of traditional autoencoders by replacing MSE with MCC. Taking advantage of insensitivity of MCC to noise, our method is capable of handling non-Gaussian noises with large outliers. The proposed R-SAE model is tested on the MNIST benchmark dataset with heavy-tailed non-Gaussian noise. Results show that, the MCC-based R-SAE method is superior to standard stacked autoencoders (S-SAE) and shows high performance in feature extraction and denoising under a large number of outliers.

2. OUR METHOD

In this section, we first briefly introduce the correntropy measure. Then we combine the correntropy and autoencoder into

* Corresponding author (e-mail: ymingwang@zju.edu.cn). This work was partly supported by the National 973 Program (No. 2013CB329500), National High Technology Research and Development Program of China (No. 2012AA020408), National Natural Science Foundation of China (No. 61103107), and Research Fund for the Doctoral Program of Higher Education of China (No. 20110101120154).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

a MCC-based robust autoencoder which is capable of dealing with non-Gaussian noises. After that, we stack the robust autoencoders to build a deep network for high-level feature learning.

2.1. Correntropy

Correntropy, as proposed in [10], is defined as a localized similarity measure. In comparison with traditional second-order statistics such as MSE, correntropy is less sensitive to outliers. The correntropy between two random variables X and Y is defined by:

$$V_\sigma(X,Y) = E[\kappa_\sigma(X-Y)], \tag{1}$$

where $E[.]$ denotes the mathematical expectation and $\kappa_\sigma(.)$ is the Gaussian kernel with σ as the kernel size:

$$\kappa_\sigma(.) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(.)^2}{2\sigma^2}). \tag{2}$$

The correntropy induces a new metric that, as the distance between X and Y gets larger, the equivalent distance evolves from 2-norm to 1-norm and eventually to zero-norm when X and Y are far apart [11]. Therefore, the correntropy measure has good property of outlier rejection.

In practice, the joint probability density function is unknown and only a finite set of samples of $\{(x_i, y_i)\}_{i=1}^N$ are available for X and Y respectively. Then the estimated correntropy can be obtained by:

$$\widetilde{V}_\sigma(X,Y) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(x_i - y_i). \tag{3}$$

2.2. Robust Autoencoders

Autoencoders aim to learn a compressed representation of data with minimum reconstruction loss. An autoencoder is a three-layer network including an encoder and a decoder. The encoder maps the input vector $\mathbf{x}$ to the hidden layer with a non-linear function:

$$\mathbf{x}' = s(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}), \tag{4}$$

where s(.) is the sigmoid function. The decoder maps the hidden layer to the output layer that has the same number of units with the input layer:

$$\mathbf{y} = s(\mathbf{W}^{(2)}\mathbf{x}' + \mathbf{b}^{(2)}). \tag{5}$$

In order to reconstruct the input data from the output layer, the parameter set $\theta = \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}\}$ is optimized by minimize the reconstruction loss. In the standard autoencoder model, the reconstruction loss is defined by the MSE between the input vector $\mathbf{x}$ and the output vector $\mathbf{y}$. However, it is sensitive to outliers so that the feature learning ability would be fragile given highly noised data. Therefore, we modify the reconstruction loss by MCC for a more robust model. The cost function of robust autoencoder is defined as follows:

$$J_{cost}(\theta) = -J_{MCC}(\theta) + J_{weight}(\theta) + J_{sparse}(\theta). \tag{6}$$

In the formulation of $J_{cost}(\theta)$, we employ a MCC-based loss function and two constraint terms. The reconstruction loss function of our method is defined as:

$$J_{MCC}(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} \kappa_\sigma(x_i^j - y_i^j), \tag{7}$$

where m is the number of training samples and n is the length of each training samples. The optimal parameter θ is obtained when $J_{MCC}(\theta)$ is maximized.

The weight decay term $J_{weight}(\theta)$ is added to prevent overfitting. It is defined as follows:

$$J_{weight}(\theta) = \frac{\lambda}{2}\sum_{l=1}^{2}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}} (w_{ji}^{(l)})^2, \tag{8}$$

where $w_{ji}^{(l)}$ represents an element in $\mathbf{W}^{(l)}$, λ is the parameter to adjust the weight of $J_{weight}(\theta)$ and s_l denotes number of units in layer l.

The sparsity penalty term $J_{sparse}(\theta)$ is employed as in [7] for better denoising ability. It is defined by:

$$J_{sparse}(\theta) = \beta\sum_{i=1}^{s_2} KL(\rho\|\hat{\rho}_i), \tag{9}$$

where β is the weight adjustment parameter, $\hat{\rho}_i$ is the activation value for the i^{th} hidden layer unit and ρ is a small number. The sparsity penalty term constrains that the value of $\hat{\rho}_i$ should be near ρ under Kullback-Leibler divergence.

2.3. Stacking Robust Autoencoders to Build Deep Networks

The robust autoencoders are stacked into R-SAE for high level feature learning. Stacking the robust autoencoders works in the same way as stacking the ordinary autoencoders. In the R-SAE model, each layer is trained separately with a robust autoencoder.

3. EXPERIMENTAL RESULTS

In this section, experiments are carried out to compare the feature learning performance of standard stacked autoencoder (S-SAE) and R-SAE under noises. The experiments include three parts: (1) we visualize the trained models to inspect the feature learning effect; (2) we employ the classification accuracy to evaluate the feature extraction performance; (3) we use the reconstruction error to measure the denoising ability.

3.1. Dataset

The experiments are carried out with the MNIST benchmark dataset of ten classes of handwritten digits (from 0 to 9) [15]. The dataset includes a training set of 60000 samples and a test set of 10000 samples. The gray scaled images of digits are in size of 28×28 and normalized to $[0, 1]$.

In order to test the feature learning ability under outliers and non-Gaussian noise, we adopt a typical heavy-tailed noise, i.e. Cauchy distributed noise, to corrupt the original images. With the Cauchy noise, outlying data could appear in abrupt large values. The Cauchy noise is centered at 0 and scaled by S, with a bigger S indicating a higher degree of noise.

3.2. Algorithm Settings

In our experiments, same stacked architectures are applied for both S-SAE and R-SAE. For the S-SAE, the cost function is defined as:

$$J_{cost}(\theta) = J_{MSE}(\theta) + J_{weight}(\theta) + J_{sparse}(\theta), \quad (10)$$

where the loss function $J_{MSE}(\theta)$ is formulated with mean square error:

$$J_{MSE}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\frac{1}{2}\|\mathbf{y}_i - \mathbf{x}_i\|^2), \quad (11)$$

and $J_{weight}(\theta)$ and $J_{sparse}(\theta)$ are formulated the same as R-SAE.

Three-layer stacked models are applied with 784 input units, 200 hidden units and 200 output units. The parameters are set as $\lambda = 0.003$, $\beta = 3$ and $\rho = 0.1$ for both methods and $\sigma = 0.2$ for R-SAE. The networks are initialized randomly and trained layer-wisely using back propagation to minimize the cost functions.

3.3. Visualization

We first show the visualizations of the learned representations from the input weights to inspect the feature learning performance. The filters (bases) learned by S-SAE and R-SAE are shown in Fig. 1. With the original images, both S-SAE and R-SAE learn useful features as shown in Fig. 1 (a-b). The filters learned by both methods are similar while sharper patterns appear in R-SAE's results. When the images contain heavy-tailed noise, the traditional S-SAE can't learn effectively that no recognizable structure is shown in the learned filters as in Fig. 1(c). That is because, the MSE-based reconstruction loss could be dominated by large values caused by outliers, therefore the S-SAE model couldn't be well trained. In contrast, the R-SAE method is more robust to outliers and keep high learning performance. As shown in Fig. 1(d), pen-stroke-like patterns are learned by R-SAE with data containing large amounts of outliers. Therefore, the proposed R-SAE has better feature learning ability under non-Gaussian noises.

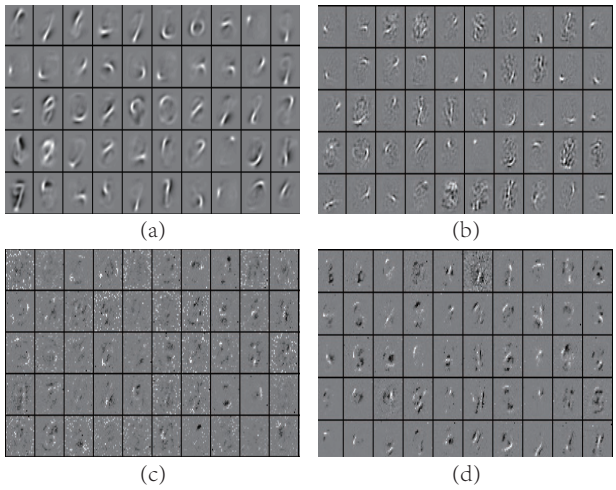

Fig. 1. Visualizations of subsets of filters learned with S-SAE and R-SAE. (a) and (b) are filters learned from original images by S-SAE and R-SAE, respectively; (c) and (d) are filters learned from images with additive Cauchy noise of S = 0.03 by S-SAE and R-SAE, respectively.

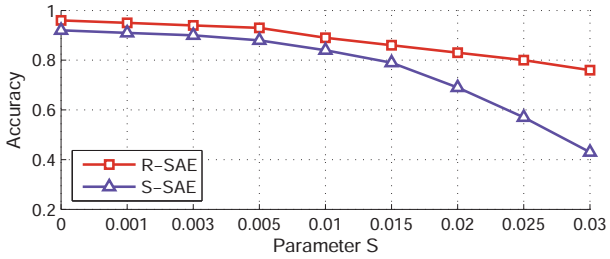

Fig. 2. Comparison of classification accuracy of R-SAE and S-SAE under different scale parameter S for Cauchy distributed noise. As the noise becomes severer, the performance of S-SAE decreases rapidly while more robust performance shows with R-SAE.

3.4. Comparison of Classification Accuracy

In this experiment, we evaluate the extracted features by classification accuracy. Once the stacked models of S-SAE and R-SAE are built and trained, the output from the highest layer is used to train a stand-alone classifier and the classification accuracy can be obtained on the test set. Here, we use the multi-class softmax model for classification.

In order to eliminate the effects of randomness in network initialization and noise generation, we present all results averaged over 10 trials. The classification results are shown in Table 1. With the original images, both S-SAE and R-SAE show high performance of 0.92 and 0.96, respectively. When the images are highly corrupted with heavy-tailed noise, R-SAE outperforms S-SAE that the classification accuracies increase by 5%-14%. In particular, as the noise gets severer, the

Table 1. Results of classification accuracy with S-SAE and R-SAE.

Method	Original	$S = 0.005$	$S = 0.010$	$S = 0.015$	$S = 0.020$
S-SAE	$0.92 \pm 1 \cdot 10^{-3}$	$0.88 \pm 4 \cdot 10^{-3}$	$0.84 \pm 6 \cdot 10^{-3}$	$0.78 \pm 5 \cdot 10^{-3}$	$0.69 \pm 2 \cdot 10^{-2}$
R-SAE	$0.96 \pm 2 \cdot 10^{-3}$	$0.93 \pm 1 \cdot 10^{-3}$	$0.90 \pm 3 \cdot 10^{-3}$	$0.86 \pm 4 \cdot 10^{-3}$	$0.83 \pm 6 \cdot 10^{-3}$

Table 2. Results of mean square reconstruction error with S-SAE and R-SAE.

Method	Original	$S = 0.005$	$S = 0.010$	$S = 0.015$	$S = 0.020$
S-SAE	$0.039 \pm 8 \cdot 10^{-5}$	$0.043 \pm 2 \cdot 10^{-4}$	$0.047 \pm 5 \cdot 10^{-4}$	$0.052 \pm 6 \cdot 10^{-4}$	$0.057 \pm 1 \cdot 10^{-3}$
R-SAE	$0.010 \pm 3 \cdot 10^{-4}$	$0.016 \pm 4 \cdot 10^{-4}$	$0.023 \pm 8 \cdot 10^{-4}$	$0.029 \pm 1 \cdot 10^{-3}$	$0.035 \pm 1 \cdot 10^{-3}$

performance of S-SAE decreases rapidly to a low accuracy of about 40% when $S = 0.03$; while the strength of R-SAE keeps robust with highly noised images as shown in Fig. 2. Results show that, the MCC-based reconstruction loss function improves feature extraction ability of S-SAE under outliers.

3.5. Comparison of Reconstruction Error

In this experiment, we measure the denoising performance under criterion of reconstruction error on the test set. The reconstruction error is defined as the pixelwise mean square error between the reconstructed images and original images without noise:

$$Err = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} - \hat{x}_{ij})^2, \tag{12}$$

where x_{ij} and $\hat{x}_{ij}$ are pixels from original images and reconstructed images, respectively.

The results of reconstruction error are also averaged over 10 trails for randomness elimination. As shown in Table 2, with the original images, low reconstruction error of 0.010 is achieved with R-SAE which outperforms 0.039 obtained with S-SAE. Moreover, with the noises added, the reconstruction errors obtained by R-SAE are 39%-63% lower than results of S-SAE, which indicates strong denoising ability of R-SAE under large amounts of outliers.

Examples of reconstructions from noised images are illustrated in Fig. 3. The reconstructed images with R-SAE preserve clear features of digits with the noises removed. However, with the S-SAE, the reconstructions are noised with blur and even with errors (such as the 2^{nd} and 5^{th} digits). The proposed R-SAE model provides more robust reconstruction and denoising performance under noises compared with S-SAE.

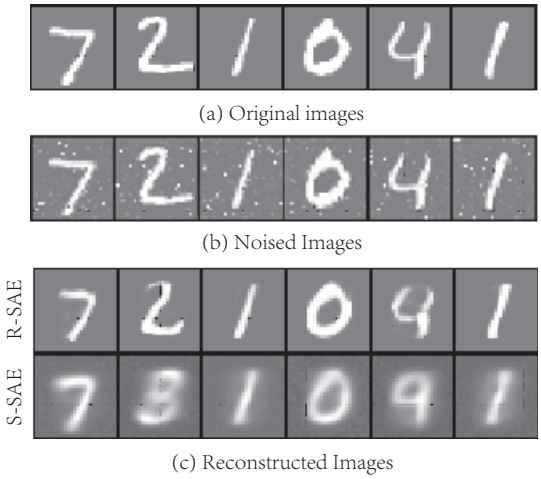

Fig. 3. Comparison of image reconstruction performance of R-SAE and S-SAE. (a) Images from test set; (b) images corrupted by Cauchy noise with S = 0.01, note that the outliers out of range of $[0, 1]$ are manually set to 0 and 1 for plotting; (c) reconstructed images with R-SAE and S-SAE.

4. CONCLUSION

In this paper, we modified the traditional stacked autoencoders using maximum correntropy criterion. Taking advantage of outlier immunity of correntropy, the modified R-SAE model obtains high feature learning performance under large amounts of outliers where the traditional S-SAE would fail. The proposed R-SAE is capable of dealing with non-Gaussian noises and outliers so that it is promising to provide robust unsupervised feature learning in practice.

5. REFERENCES

[1] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[2] R. Hinton, G.and Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[3] Y. Boureau and Y. Cun, "Sparse feature learning for deep belief networks," in *Advances in neural information processing systems*, 2007, pp. 1185–1192.

[4] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[5] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[6] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[7] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 350–358. 2012.

[8] A. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 6, pp. 748–763, 2002.

[9] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 337–350, 2006.

[10] W. Liu, P. Pokharel, and J. Principe, "Correntropy: A localized similarity measure," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 2006, pp. 4919–4924.

[11] W. Liu, P. Pokharel, and J. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *Signal Processing, IEEE Transactions on*, vol. 55, no. 11, pp. 5286–5298, 2007.

[12] R. He, W. Zheng, B. Hu, and X. Kong, "A regularized correntropy framework for robust pattern recognition.," *Neural Computation*, vol. 23, no. 8, pp. 2074–2100, 2011.

[13] K. Jeong, W. Liu, S. Han, E. Hasanbelliu, and J. Principe, "The correntropy mace filter," *Pattern Recognition*, vol. 42, no. 5, pp. 871 – 885, 2009.

[14] R. He, B. Hu, W. Zheng, and X. Kong, "Robust principal component analysis based on maximum correntropy criterion," *Image Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1485–1494, 2011.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.