Robust Feature Learning by Improved Auto-encoder From non-Gaussian Noised Images

Dan Zhao, Baolong Guo, Jinfu Wu, Weikang Ning, Yunyi Yan
School of Aerospace Science and Technology, Xidian University
Xi'an, Shaanxi 710071, China
Email: fengjiran@foxmail.com; blguo@xidian.edu.cn

***Abstract*—Much recent research has been devoted to learning algorithms for deep architectures such as Deep Belief Networks(DBN) and stacks of auto-encoder variants, with impressive results obtained in several areas, mostly on vision and languages datasets. These learning algorithms aim to find good representations for data, which can be used for classification, reconstruction, visualization and so on. Despite the progress, most existing algorithms would be fragile to non-Gaussian noises and outliers due to the criterion of mean square error(MSE) and cross entropy(CE). In this paper, we propose a robust auto-encoder called correntropy-based contractive auto-encoder(C-CAE) to learn robust features from data with non-Gaussian noises and outliers. The maximum correntropy criterion(MCC) is adopted as reconstruction cost function and a well chosen penalty term is added to the reconstruction cost function. By replacing cross entropy with MCC, the proposed method can learn robust features from the data containing non-Gaussian noises and outliers. The penalty term corresponds to the Frobenius norm of the Jacobian matrix of the encoder activations with respect to the input. By adding the penalty term, the antinoise ability of the proposed method is improved. The proposed method is evaluated using the MNIST benchmark dataset. Experimental results show that, compared with the traditional auto-encoders, the proposed method learns robust features, improves classification accuracy and reduces the reconstruction error, which demonstrates that the proposed method is capable of learning robust features on noisy data.**

***Keywords*—feature learning, stacked autoencoder, correntropy**

I. Introduction

In the recent years, unsupervised feature learning with deep architectures such as Deep Belief Networks(DBNs) and stacked auto-encoders(SAEs) has been widely studied and applied. The revival of interest in such deep architectures is due to the discovery of novel approaches [1–3] that proves successful at learning their parameters. Deep architectures have shown high feature learning performance in many fields include classification and regression tasks that involve image [4–6], language [7] and speech [8], because deep architectures can extract more abstract, invariant features from the unlabeled data, and thus are believed to have the ability of yielding higher classification accuracy than traditional, shallower classifier.

Despite the progress of deep architectures, robust feature learning is still faced with many challenges due to noises and outliers which are commonly appeared in the real-world data. The present research begins with the question of how to learn the good intermediate representations from the data with large noises and outliers. In order to improve the antinoise ability of the deep architectures, many efforts have been made. Boureau et al. [9] and Lee et al. [10] proposed the sparsity of representation as a supplemental criterion. The algorithms attempted to learn sparse and good intermediate representations from the input data. Vincent et al. [11, 12] modified the traditional stacked auto-encoder(SAE) to learn more useful features from corrupted data and developed the stacked denoising auto-encoder(SDAE). The SDAE based on stacking layers of denoising autoencoders which are trained locally to denoise corrupted versions of their inputs. The resulting algorithm is a straightforward variation on stacking of traditional auto-encoders. By corrupting the input data and using denoising criterion, the SDAE could learn robust representations and achieved good performance under different types of noises. The SDAE model was extended by Xie et al. [13] with sparse coding technique. With the reconstruction cost function regularized by a sparsity inducing term, better denoising performance was achieved. Rifai et al. [14, 15] proposed a contractive auto-encoder to improve the robustness of learned features by introducing the Frobenius norm of the Jacobian of the non-linear mapping as a penalty term.

Recently, correntropy was proposed as a localized similarity measure based on information theoretic learning(ITL) and kernel methods [16]. As such it has vastly different properties when compared with mean square error(MSE) and cross entropy(CE) that can be very useful in nonlinear, non-Gaussian signal processing. It's insensitive to outliers compared with MSE and cross entropy, and has been successfully utilized for cost function design in non-Gaussian signal processing [17, 18]. He et al. [19] proposed a robust principal component analysis method based on maximum correntropy criterion(MCC) to achieve high performance under outliers. Yu et al [20] proposed a robust stacked auto-encoder model with MCC shows high performance in feature extraction and under a large amounts of outliers. These studies show that, correntropy is robust to outliers so that it is promising for robust algorithm design.

Although the existing learning algorithms 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 MSE and cross entropy criterion which would be sensitive to impulsive noises and outliers.

Inspired by the success of correntropy-based approaches in outlier suppression and contractive auto-encoders, this paper proposes a correntropy-based contractive auto-encoder(C-CAE) method. By replacing MSE or cross entropy with correntropy, the anti-noise ability of C-CAE is improved. The

978-1-4799-8633-0/15/$31.00 ©2015 IEEE

proposed method is tested on MNIST benchmark dataset. Results show that the C-CAE is superior to standard auto-encoder and shows high performance in feature extraction and denoising under a large amounts of outliers.

The organization of the paper is as follows. First, we give a detail description of the proposed method in section II. Then in section III, we present experimental results to compare the ability of antinoise between traditional auto-encoder and the proposed method. Finally, section IV summarizes the main conclusions.

II. Correntropy-based contractive auto-encoder

In this section, we first introduce the correntropy measure, and then we combine the correntropy and contractive auto-encoder into the correntropy-based contractive auto-encoder(C-CAE) which is capable of dealing with non-Gaussian noises. After that, we stack the C-CAE to build a deep network for high-level feature learning.

A. Correntropy

Recently, the concept of correntropy was proposed in ITL [17] to process non-Gaussian noises and impulsive noises. The correntropy is directly related to the Renyi's quadratic entropy [21] in which the Parzen windowing method is used to estimate the data distribution [22]. It is a local similarity measure between two arbitrary random variables X and Y, defined by

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

where $E[\cdot]$ denotes the mathematical expectation and $\kappa_\sigma(\cdot)$ is a normalized Gaussian kernel function that satisfies Mercer theory [23], with σ as the kernel size:

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

With a clear theoretic foundation, the correntropy is symmetric, positive, and bounded.

The correntropy induces a new metric that, as σ increases, the high-order moments decay faster, so the second-order moment tends to dominate and correntropy approaches correlation. That is, 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 [17]. Therefore, the correntropy measure has good property of outlier rejection.

In practice, the joint probability density function of X and Y is often unknown and only a finite number of samples $\{(x_i,y_i)\}_{i=1}^N$ of the variables X and Y are available. Therefore, it is more common to use the sample estimator for estimating the correntropy:

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

The sample-based correntropy criterion was further extended for a general similarity measurement between two discrete vectors. That is, the correntropy induced metric(CIM) [17] was introduced for any vectors $A=(a_1,...,a_N)^T$ and $B=(b_1,...,b_N)^T$ as follows:

$$\begin{aligned} CIM(A,B) &= \left(g(0) - \frac{1}{N}\sum_{i=1}^{N} g(e_i)\right)^{\frac{1}{2}} \\ &= \left(g(0) - \frac{1}{N}\sum_{i=1}^{N} g(a_i-b_i)\right)^{\frac{1}{2}} \end{aligned} \tag{4}$$

where the error e_i is defined as $e_i = a_i - b_i$, $g(x)$ is Gaussian kernel $g(x) \triangleq \exp\left(-\frac{x^2}{2\sigma^2}\right)$. For adaptive systems, the below correntropy of error e_i,

$$max\frac{1}{N}\sum_{i=1}^{N} g(e_i) \tag{5}$$

is called the maximum correntropy criterion(MCC).

B. The proposed method

Auto-encoders aim to learn a compressed representation of data by minimizing the reconstruction cost function. The proposed method correntropy-based contractive auto-encoder(C-CAE) is a three-layer network including an encoder and a decoder as shown in Fig.1. It has one visible layer of d inputs, one hidden layer of h units, one reconstruction layer of d units, and an activation function. Correntropy-based contractive auto-encoders aim to learn the robust representations of data with noises and outliers.

The encoder maps the input vector $x \in R^d$ to the hidden layer and produces the latent activity $y \in R^h$. Then, y is mapped by a decoder to an output layer that has the same size of the input layer, which is called reconstruction of the input. The reconstructed values are denoted as $z \in R^d$. Mathematically, these two steps can be formulated as

$$y = f(W_y x + b_y) \tag{6}$$

$$z = f(W_z y + b_z) \tag{7}$$

where W_y and W_z denote the input-to-hidden and hidden-to-output weights, respectively, b_y and b_z denote the bias of hidden and output units, and $f(\cdot)$ denotes the activation function. To get a nonlinear mapping, the activation function $f(\cdot)$ is set to be a sigmoidal function in both the encoder and decoder.

In this paper, the following constraint holds

$$W_y = W_z' = W \tag{8}$$

It is so called that the C-CAE has tied weights. Thus, there remain three groups of parameters to learn: $\theta = \{W, b_y, b_z\}$. In order to reconstruct the input data from the output layer, the parameter set $\theta = \{W, b_y, b_z\}$ is optimized by minimizing the reconstruction cost function. In the standard auto-encoder model, the reconstruction cost function is defined by the MSE or cross entropy between the input vector x and the output vector z. However, the MSE and cross entropy are sensitive to outliers and impulsive noises so that the feature learning ability would be fragile given highly noised data. To encourage robustness of representations to outliers and impulsive noises of a training input x, the correntropy of input

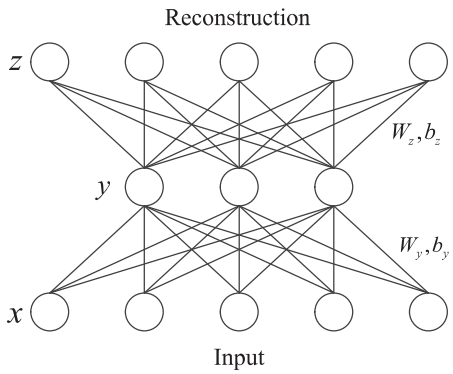

Fig. 1. Single layer auto-encoder. The auto-encoder learns a hidden features y from input x by reconstructing it on z. Corresponding parameters are denoted in the network.

x and output z is adopted as the reconstruction cost function. So the reconstruction cost function of C-CAE is defined as follows:

$$J_{cost}(\theta) = L(x, z) + \lambda \|J_f(x)\|_F^2 \tag{9}$$

In our implementation, the cost is actually computed on mini-batch of input since mini-batch update strategy is adopted for the large dataset. The formulation of $J_{cost}(\theta)$ consists of a correntropy-based cost function and a regularization term. λ is a positive hyperparameter that controls the strength of the regularization. The reconstruction cost function is defined as:

$$L(x, z) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{n} \kappa_\sigma(x_{ik} - z_{ik}) \tag{10}$$

where m is the number of training samples of a mini-batch and n is the length of each training sample.

In order to encourage robustness of representation, Frobenius norm of the Jacobian $J_f(x)$ of the non-linear mapped by encoding function f to hidden representation $h = f(x) \in R^{d_h}$, this sensitivity penalization term is the sum of squares of all partial derivatives of the extracted features with respect to input dimensions. In the case of a sigmoid nonlinearity, the penalty on the Jacobian norm has the following simple expression:

$$\begin{aligned} \|J_f(x)\|_F^2 &= \sum_{i=1}^{d_h} \sum_{j=1}^{d_x} \left(\frac{\partial h_i}{\partial x_j} \right)^2 \\ &= \sum_{i=1}^{d_h} \sum_{j=1}^{d_x} \left(h_i(1 - h_i) \cdot W_{ij} \right)^2 \\ &= \sum_{i=1}^{d_h} \left(h_i(1 - h_i) \right)^2 \cdot \sum_{j=1}^{d_x} W_{ij}^2 \end{aligned} \tag{11}$$

Computing this penalty(or its gradient) is similar to and has about the same cost as computing the reconstruction cost(or, respectively, its gradient). The overall computational complexity is $O(d_x \times d_h)$.

Penalizing $\|J_f(x)\|_F^2$ encourages the mapping to the feature space to be contractive in the neighborhood of the training data. The flatness induced by having low valued first derivatives will imply an invariance or robustness of the representation for small variations of the input.

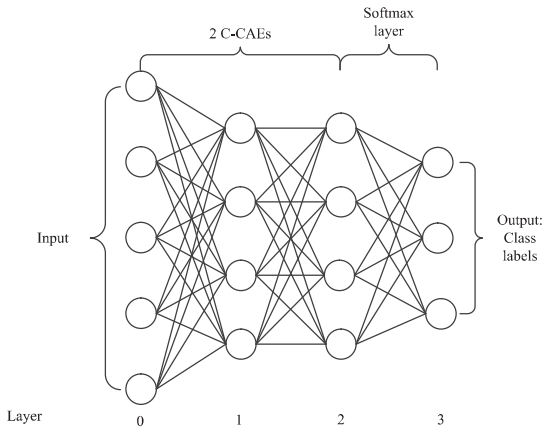

Fig. 2. Instance of a stacked C-CAE connected with a softmax regression layer. It has four layers: one input layer, two hidden layer, and an output layer.

C. Stacking correntropy-based contractive auto-encoders to build deep networks

Stacking the input and hidden layers of C-CAEs together layer by layer constructs a stacked C-CAE model is used to generate deep features of the data. Fig.2 shows a typical instance of a stacked C-CAE connected with a subsequent softmax regression classifier.

The first C-CAE maps inputs in 0th layer to a first layer feature in first layer. After training the first layer C-CAE, subsequent layers of C-CAEs are trained via the output of its previous layer. Finally, parameters throughout the whole architecture can be adjusted slightly while we are training the softmax classifier. This step is called fine-tuning. These steps are so called greedy layer-wise training method of deep architectures[24].

III. Experimental results

In this section, experiments are carried out to evaluate the feature learning performance of standard stacked auto-encoder and the C-CAE. We use a Python library called Theano[25, 26] to implement the following experiments. 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 learning performance; (3) we use the reconstruction error to measure the denoising ability.

A. Dataset

The experiments are carried out with the MNIST benchmark dataset of ten classes of handwritten digits(from 0 to 9). The MNIST dataset consists of handwritten digit images and it is divided in 60000 examples for the training set and 10000 examples for testing. All digit images have been size-normalized and centered in a fixed size image of 28×28 pixels. The gray scaled images of digits are normalized to [0, 1]. In order to test the feature learning ability under outliers, we adopt the impulsive Gaussian mixture noises.

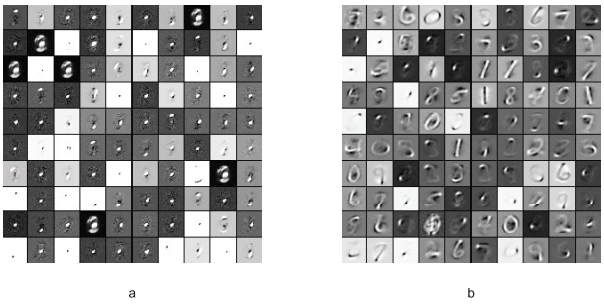

Fig. 3. Visualization of filters learned by auto-encoders on MNIST. (a) are filters learned by the traditional auto-encoder with additive impulsive Gaussian mixture noises; (b) are filters learned by C-CAE method with additive impulsive Gaussian mixture noises

B. Visualization of Features

Let the noise probability density function be an impulsive Gaussian mixture $p_Z(z) = 0.9 \times N(0, 0.1) + 0.1 \times N(4, 0.1)$. Fig.3 show the weights(called filters) of the first layer learned from images with additive impulsive Gaussian mixture noises. When the images contain impulsive Gaussian mixture noises, the traditional auto-encoder can't learn effectively that no recognizable structure is shown in the learned filters as in Fig.3(a). That is because the cross entropy based reconstruction cost function could be dominated by large values caused by outliers, therefore the traditional auto-encoder couldn't be well trained with additive impulsive Gaussian mixture noises. In contrast, the C-CAE method is more robust to outliers and keep high learning performance. As shown in Fig.3(b), penstroke-like patterns are learned by C-CAE with data containing large amounts of outliers. Therefore, the proposed C-CAE method has better feature learning ability under large amounts of outliers.

C. Classification Accuracy

In this section, we evaluate the extracted features by classification accuracy. Once the stacked models of traditional auto-encoder and C-CAE are built and trained, the output from the highest layer is used to trained a stand-alone classifier and the classification accuracy can be obtained on the test dataset. Here we use a multi-class softmax classifier for classification. Five layers stacked models are applied in traditional auto-encoder and C-CAE with 784 inputs units, 200-200-100 hidden units and 10 output units.

With the original images without noises, C-CAE and traditional auto-encoder achieve the classification accuracy of 98.10% and 97.91%, respectively. All of them show high performance in classification. But when the images are highly corrupted with large amounts of outliers, the C-CAE method achieves the classification accuracy of 98.03%, and the result of traditional auto-encoder is 95.44%. The results show that the C-CAE has highly anti-noise ability than the traditional auto-encoder. Benefit from correntropy-based reconstruction cost function and the penalty term, the feature learning ability under outliers has been improved.

D. Reconstruction Error

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

$$Error = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} - z_{ij})^2 \tag{12}$$

where x_{ij} and z_{ij} are pixels from original images and reconstruction images, respectively.

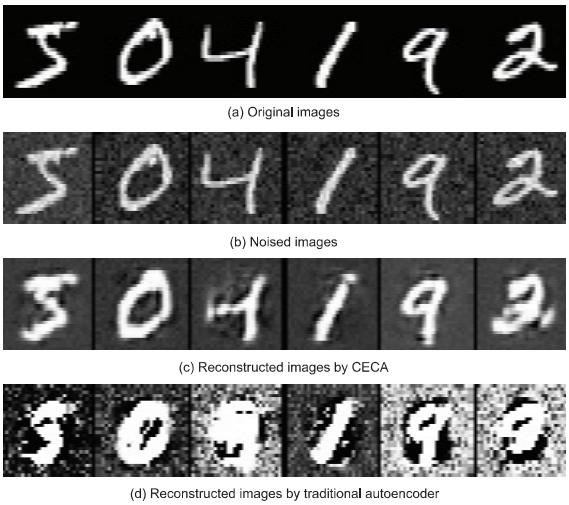

Fig. 4. Comparison of image reconstruction performance of C-CAE and traditional auto-encoder. (a) The original images without noises; (b) images corrupted by impulsive Gaussian mixture noise; (c) reconstructed images by C-CAE method; (d) reconstructed images by traditional auto-encoder.

With the original images, low reconstruction error of 2.55% is achieved with traditional auto-encoder which outperforms 4.68% obtained with C-CAE. However, with the noises added, the reconstruction error obtained by C-CAE is 2.76% lower than 3.76% obtained by traditional auto-encoder , which indicates strong denoising ability of C-CAE under large amounts of outliers.

Experiments of reconstructions from noised images are illustrated in Fig.4. The reconstructed images with C-CAE preserve clear features of digits with noises removed compare with the noised images. However, with the traditional auto-encoder, the reconstructed images are noised with blur and some digits are hard to recognize(such as the 3rd and 6th digits). The proposed C-CAE method provides more robust reconstruction and denoising performance under noises compared with traditional auto-encoder.

IV. Conclusion

In this paper, the proposed correntropy-based contractive auto-encoder method modified the traditional auto-encoders using maximum correntropy criterion and the penalty term corresponds to the Frobenius norm of the Jacobian matrix of the encoder activations with respect to the input. Taking advantages of outliers immunity of correntropy and the penalty term, the modified method obtains high feature learning performance under large amounts of outliers. The proposed method is capable of dealing with non-Gaussian noise and outliers so that it is promising to provide robust unsupervised feature learning in practice.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China under Grants No. 61105066, No. 61305041, No. 61305040.

REFERENCES

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

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

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

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, Conference Proceedings, pp. 1097–1105.

[5] Z. Zuo and G. Wang, "Learning discriminative hierarchical features for object recognition," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1159–1163, 2014.

[6] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 2094–2107, 2014.

[7] D. Yu, L. Deng, and S. Wang, "Learning in the deep-structured conditional random fields," in *Proc. NIPS Workshop*, 2009, Conference Proceedings, pp. 1–8.

[8] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, Conference Proceedings, pp. 5060–5063.

[9] Y.-l. Boureau and Y. L. Cun, "Sparse feature learning for deep belief networks," in *Advances in neural information processing systems*, 2008, Conference Proceedings, pp. 1185–1192.

[10] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Advances in neural information processing systems*, 2008, Conference Proceedings, pp. 873–880.

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

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[13] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, 2012, Conference Proceedings, pp. 341–349.

[14] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, Conference Proceedings, pp. 833–840, cAE.

[15] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, *Higher order contractive auto-encoder*. Springer, 2011, pp. 645–660.

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

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

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

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

[20] Q. Yu, W. Yueming, Z. Xiaoxiang, and W. Zhaohui, "Robust feature learning by stacked autoencoder with maximum correntropy criterion," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, Conference Proceedings, pp. 6716–6720.

[21] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.

[22] I. Santamarła, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization," *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 2187–2197, 2006.

[23] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[24] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[25] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio, "Theano: new features and speed improvements," *arXiv preprint arXiv:1211.5590*, 2012.

[26] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a cpu and gpu math expression compiler," in *Proceedings of the Python for scientific computing conference (SciPy)*, vol. 4. Austin, TX, 2010, p. 3.