Improved Denoising Auto-encoders for Image Denoising

Qian Xiang
The College of Information Engineering
Wuchang Institute of Technology
Wuhan, P.R.China

Xuliang Pang
The College of Electric and Information,
Wuhan University of Engineering
Wuhan, P.R.China

Abstract—**Image denoising is an important pre-processing step in image analysis. Various denoising algorithms, such as BM3D, PCD and K-SVD, obtain remarkable effects. Recently a deep denoising auto-encoder has been proposed and shown excellent performance compared to conventional image denoising algorithms. In this paper, we study the statistical features of restored image residuals produced by Denoising Auto-encoders and propose an improved training loss function for Denoising Auto-encoders based on Method noise and entropy maximization principle, with residual statistics as constraint conditions. We compare it with conventional denoising algorithms including original Denoising Auto-encoders, BM3D, total variation (TV) minimization, and non-local mean (NLM) algorithms. Experiments indicate that the Improved Denoising Auto-encoders introduce less non-existent artifacts and are more robustness than other state-of-the-art denoising methods in both PSNR and SSIM indexes, especially under low SNR.**

Keywords- Image Denoising; Auto-encoders; Method noise entropy maximization principle

I. Introduction

Image restoration involves recovering a clean image x from its corrupted observation $\tilde{x}$. Image noises generally include additive noise, multiplicative noise, salt-and-pepper noise and so on. Among them, the Additive White Gaussian Noise (AWGN) is one of the most common noise model and then

$$\tilde{x} = x + n \tag{1}$$

where n is the AWGN introduced in x. Therefore, the task of image denoising can be described as removing n from $\tilde{x}$ while concurrently preserving edge details of the image.

Many algorithms have been proposed for image denoising, including Mean filtering, Gaussian filtering and Bilateral filtering algorithm [1], each using different local area characteristics to remove noise from image. Methods like the, Wavelet-based denoising method [2-3], Total Variate denoising model (TV) [4-5] method, Sparse Dictionary and Decomposition algorithm [6-8], Non-local Self-similarity model [9-10], BM3D algorithm [11-12] are also widely used in image denoising, and obtain remarkable results.

Recently, many researchers try to use deep learning technology for image denoising. Xie et al. [13] successfully improved the performance of image denoising and blind inpainting with a Deep Neural Networks. Mao et al. proposed a symmetrical-layer deep convolutional Encoder-Decoder Networks applied to an image denoising step [14]. Zhang et al. proposed DnCNN [15] denoising algorithm and use it to train the networks with residual outputs. Experiments show that the results obtained by DnCNN at different noise levels are better than those of BM3D. Pascal Vincen et al. [16] build deep networks by stacking layers of denoising autoencoders (SDEA), and shown the value of SDEA to learn useful higher level representations from natural image patches. Zhang et al. [17] design a deep convolution neural network (DCNN) to and employ gradient clipping scheme to train it. Experimental results demonstrate that the proposed denoising method can achieve a better performance compared with the state-of-the-art denoising methods.

In this paper, we begin by introducing the unsupervised Denoising Auto-encoders. Then, we proposed constraint conditions of the training loss function of the Denoising Auto-encoders, and construct an improved auto-encoders training loss function based on Method noise reduction and residual entropy maximization.

II. Model Description

Basic Denoising auto-encoders (DAE) is a two-layer neural network as Fig. 1. Its first layer is used to receive corrupted data x and second layer is used to generate output data $\tilde{x}$. Use $p_{reconstruct}(x|\tilde{x})$ as reconstruction distribution function and $h = f(\tilde{x})$ as the output of encoder, the denoising process can be accomplished by minimizing the loss function with stochastic gradient descent method

$$L = -log p_{decoder}(x|h = f(\tilde{x})) \tag{2}$$

If the conditional probability $p_{decoder}(x|h = f(\tilde{x}))$ obeys the Gaussian distribution, and the training criterion of DAE is to minimize the squared error $||g(f(\tilde{x})) - x||^2$, the output $g(f(\tilde{x}))$ estimate the center of mass of $\tilde{x}$, which can be seen as a reconstruction of clean points [18].

One typical category to improve sparse feature model performance of the basic DAE is to use deep multi-layer neural networks as its encoder and decoder model. This structure can be illustrated as Fig. 2.

Be sponsored by Green Fan Manufacturing Collaborative Innovation center in Hubei province.

978-1-5386-7604-2/18/$31.00 ©2018 IEEE

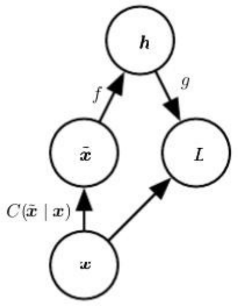

Fig. 1. *The computational graphs of the cost function for a denoising auto-encoder.*

$$f(\boldsymbol{x}) = S(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \quad (3)$$

$$\hat{\mathbf{y}}(\boldsymbol{h}) = S(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{b}') \quad (4)$$

where $S = (1 + \exp(-x))^{-1}$ is sigmoid activation function, and $\boldsymbol{\Omega} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{W}', \boldsymbol{b}'\}$ represents the weights and biases of the network. DAE can be trained to minimize the following loss function with stochastic gradient descent method

$$L(\boldsymbol{x}, \hat{\mathbf{y}}; \boldsymbol{\Omega}) = \sum_{i=1}^{N} ||\boldsymbol{y}_i - \hat{\mathbf{y}}(\boldsymbol{x}_i)||_2^2 \quad (5)$$

To generate sparse coding and avoid over-fitting, sparsity-inducing regularized term can be added to DEA training loss function:

$$L(\boldsymbol{x},\hat{\mathbf{y}};\Omega) = \sum_{i=1}^{N} \frac{1}{2}||\boldsymbol{y}_i - \hat{\mathbf{y}}(\boldsymbol{x}_i)||_2^2 + \beta(||\boldsymbol{W}||_F^2 + ||\boldsymbol{W}'||_F^2) \quad (6)$$

where β is the regularization coefficient.

III. Loss function based on maximum residual entropy

Mean-squared-error (MSE), along with the related quantity of peak signal-to-noise ratio (PSNR), is traditionally the simplest and most widely used objective methods for assessing perceptual image quality, by quantifying the errors (residuals) between a corrupted image and a reference image. But both psychophysical and physiological experiments have proved that the MSE and PSNR can only be seen as a simulation of the functional properties of early stages of human visual system (HVS), and are not very well characterization of perceived visual quality [19].

A. Method noise

Definition 1 (Method noise [9]) Let $\boldsymbol{x}$ be an image and D_h a denoising operator with a filtering parameter ***h***. Then, the method noise $\boldsymbol{M_L}$ can be as defined as the difference

$$M_L = \boldsymbol{x} - D_h * \boldsymbol{x} \quad (7)$$

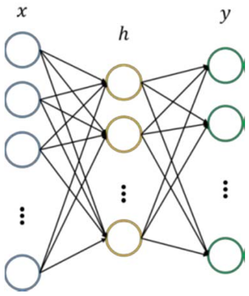

Fig. 2. *The deep multi-layer neural networks model for DAE.*

Therefore, when a non-noisy images is processed by denoising methods, a “good” denoising algorithm should not introduced any changes to the input images, that is, its Method noise should be as small as possible and must look like a noise in the input images which contains as little structure imformation as possible. Method noise indicates another index to evaluate any denoising method, which is not just the traditional quantity of noise romoved from corrupted image. Analysis and calculation results show several classical local smoothing filters such as the Total Variation minimization, the Gaussian filtering, the neighborhood filtering and so on are all exist Method noise.

According to (6), DAE reconstruct image by minimizing the mean-squared-error loss function, and this loss function does not constrain the Method noise. Therefore, DAE also exist Method noise and we should try to reduce its impact.

B. Residual entropy Maximum

According to the principle of information entropy, the degree of ordering of a system is inversely proportional to its information entropy. The information entropy $\boldsymbol{H(p)}$ of a random variable $\boldsymbol{x}$is defined as:

$$\boldsymbol{H(p)} = \sum_{x} p(x)\boldsymbol{log}_2\left(\frac{\mathbf{1}}{\boldsymbol{p(x)}}\right) \quad (8)$$

Under same variance, a Gaussian distribution random variable has the maximum information entropy, which is equal to its variance, that is, a Gaussian distribution random variable contains the least amount of information [20]. Therefore, to make the residual $\hat{\boldsymbol{e}} = g(f(\boldsymbol{x})) - \boldsymbol{x}$ contains as little information of $\boldsymbol{x}$ as possible$\hat{\boldsymbol{e}}$ should be as close to a Gaussian distribution white noise as possible.

C. Residual Gaussian measure based on high order statistics

Because it is difficult to calculate the Gaussian of a random variable directly with information entropy, it is necessary to search other Gaussian measure methods. Therefore the higher-order statistics of $\hat{\boldsymbol{e}}$ can be choose as it”s Gaussian measure.

If we set $\{x(n)\}$ to be a zero mean stationary random process, its kth-order cumulant can be defined as

$$C_{k,x}(t_1,\cdots,t_{k-1}) = cum(x(n),x(n+t_1),x(n+t_{k-1}) \quad (9)$$

where $t_i\,(\,i = 1,\ldots,k-1)$ is time delay.

The *k*th-order moment of $\{x(n)\}$ can be defined as

$$m_{k,x}(t_1,\cdots,t_{k-1}) = mom(x(n),x(n+t_1),x(n+t_{k-1})(10)$$

Because $\{x(n)\}$ is the kth-order stable, so $c_{k.x}$ and $m_{k.x}$ are functions of $m_1, m_2,\cdots,m_{k-1}$ and have nothing to do with the time delay, its second-, third- and fourth-order cumulants are

$$C_{2,x}(t) = \mathbb{E}[x(n)x(n+t)] \qquad (11)$$

$$c_{3,x}(t) = \mathbb{E}[x(n)x(n+t_1)x(n+t_2)] \qquad (12)$$

If $\{x(n)\}$ is a Gaussian random process, then

$$c_{k,x} = 0, k = odd \qquad (14)$$

$$m_{k,x} = 0, k \geq 3 \qquad (15)$$

Therefore, high-order statistics of a random variable can be used to measure its Gaussian.

D. *Unbiased estimate*

Assume that $\hat{e}$ will converge to a parameter e after training, i.e. e is an estimate of $\hat{e}$. Next, we consider the Mean squared error (**MSE**) of $\hat{e}$.

$$\mathbf{MSE}(\hat{e},e) = \mathbb{E}(\hat{e}-e)^2 = var(\hat{e}) + (\mathbb{E}(\hat{e}) - e)^2 (16)$$

Thus, $\mathbf{MSE}(\hat{e},e)$ is composed of both the variance and the square of bias of the estimator. For image denoising, if the mean of the image residuals is equal to 0, the mean of $\hat{e}$ will be an unbiased estimate of 0 and $\hat{e}$ will be a zero-mean random process, then we have

$$\mathbf{MSE}(\hat{e},0) = var(\hat{e}) \qquad (17)$$

which means **MSE** of $\hat{e}$ will converge to its variance.

In conclusion, if the residual $\hat{e}$ obeys the zero-mean Gaussian distribution, then the $\mathbf{MSE}(\hat{e})$ is equal to $var(\hat{e})$, i.e., information entropy of $\hat{e}$ In this case, the training of DAE via a gradient descent algorithm, which reduces the $\mathbf{MSE}(\hat{e})$, will make the reconstructed image retain more information of the reference image by removing that from $\hat{e}$ as much as possible, and finally we can obtain better image reconstruction quality.

To ensure the residual $\hat{e}$complies with a zero-mean Gaussian distribution, we use higher-order statistics as constraint conditions. For simplicity, $m_{1,\hat{e}}$, $c_{3,\hat{e}}$ ($m_{3,\hat{e}}$), and $c_{4,\hat{e}}$ are selected as constraint conditions, and these statistics should be equal to 0. According to the Lagrange multiplier method, the improved loss function is defined as

$$L(w) = ||x, g(f(\tilde{x})||_2^2 + \alpha|m_{1,\hat{e}}| + \beta|c_{3,\hat{e}}| + \lambda|c_{4,\hat{e}}| \quad (20)$$

where α, β and λ are the weights of constraints. Their value determines how close the corresponding item is to 0 [18]. A larger weight corresponds to an item closer to 0, which has better Gaussian. However, excessive weight will reduce the weight of the mean-square-error and may introduce image distortions.

IV. Experiments

In this section, we evaluate the denoising performance of our models against a few state-of-the-art methods through various experiments.

A. *Experimental data*

The experiments are carried out on the MINIST benchmark dataset, which contains ten classes of handwritten digits (from 0 to 9) [21]. Among them, the training set includes 60000 samples and the test set includes 10000 samples. All these samples are gray scaled digits images with size 28×28 and normalized to [0,1].

To test the denoising performance in different noise types, we add Gaussian noise and salt-and-pepper noise (with different noise levels) respectively into the MINIST samples to construct the training and test set.

B. *Network structure and parameter setting*

The experiments are carried out by using convolutional Encoder-Decoder Networks, the Encoder section consists of two 32×3×3 convolutional layers and two 2×2 maximum pooling layers; the Decoder section consists of two 32×3×3 convolutional layers and two 2×2 upper sampling layers. The value of weights in loss function are set to α=10, β=5, λ=20.

C. *Comparison and analysis*

NLM, BM3D, DAE (with MSE loss function) are compared with our method. The Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) indexes are evaluated by applying an average of denoising processes to 200 images selected randomly from the test dataset.

The image denoising results in Gaussian noise, with a standard deviation (σ=50), are shown in Fig.3. The image denoising results in salt-and-pepper noise, with a noise density (d=0.09), are shown in Fig.4.

$$c_{4,x}(t_1,t_2,t_3) = \mathbb{E}[x(n)x(n+t_1)x(n+t_2)x(n+t_3)]c_{2,x}(t_1)c_{2,x}(t_2-t_3) - c_{2,x}(t_2)c_{2,x}(t_3-t_1)c_{2,x}(t_3)c_{2,x}(t_1-t_3) \quad (13)$$

TABLE 1. PSNR AND SSIM WITH VARIOUS DENOISING METHODS FOR GAUSSIAN NOISE AT DIFFERENT NOISE LEVELS.

σ	10		30		50		80		125	
method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NLM	27.42	0.649	23.81	0.474	20.17	0.399	15.07	0.278	11.44	0.112
BM3D	**30.22**	0.635	24.01	0.468	20.16	0.398	16.13	0.336	12.63	0.225
DAE	21.62	0.933	20.82	0.920	20.54	0.905	19.25	0.874	17.45	0.801
IDAE	25.41	**0.975**	**24.30**	**0.969**	**23.38**	**0.954**	**21.81**	**0.935**	**19.13**	**0.876**

TABLE 2. PSNR AND SSIM WITH VARIOUS DENOISING METHODS FOR SALT-AND-PEPPER NOISE AT DIFFERENT NOISE LEVELS.

d	0.01		0.04		0.09		0.16		0.25	
method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NLM	**25.53**	0.941	20.65	0.766	20.01	0.590	17.41	0.367	13.37	0.187
BM3D	23.80	0.733	20.91	0.656	19.03	0.405	16.85	0.343	14.68	0.293
DAE	21.14	0.912	20.60	0.895	19.94	0.890	18.98	0.852	18.04	0.825
IDAE	24.64	**0.968**	**23.86**	**0.958**	**23.02**	**0.946**	**21.89**	**0.934**	**20.22**	**0.904**

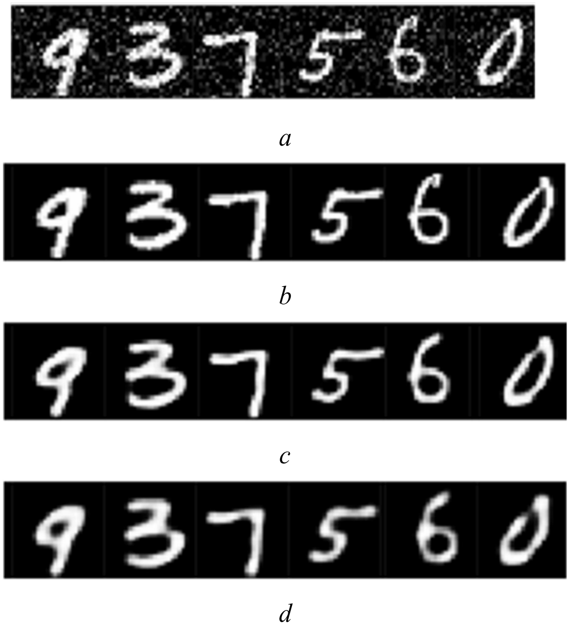

Fig.3. *Comparison of the denoising performance of Gaussian noise (σ=50).*
(a) Noisy Image (b) Original Image (c) IDEA (d) DAE

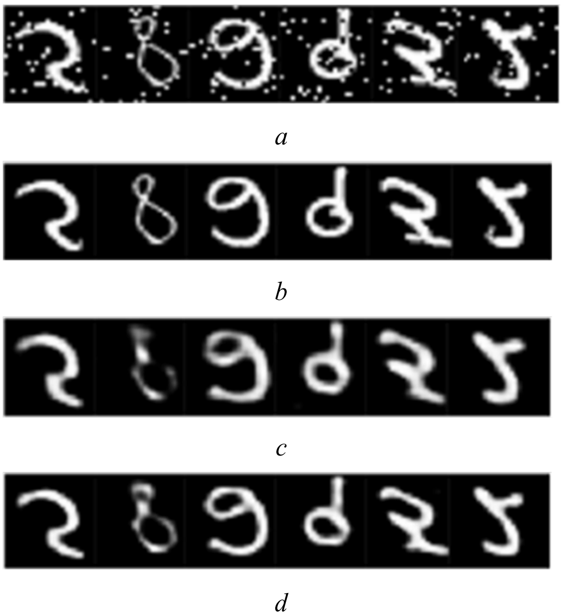

Fig.4. *Comparison of the denoising performance of salt-and-pepper noise (d=0.09).*
(a) Noisy Image (b) Original Image (c) DAE (d) IDEA

We achieve the following results from the experiments.

The IDAE can obtain structural and edge information of the original image from a noisy image more effectively than other state-of-the-art denoising methods, such as BM3D, and get better image quality recovery.

II. Conclusion

A. Conclusion

In this paper, we developed and evaluted an loss function construction method for denoising auto-encoders based on Method noise, residual entropy maximization, and residual variance unbiased estimate. We begin with the idea that a good denoising algorithm should only introduce Method noise as little as possible. Then, analysis from the point-of-view of information theory are carried out, and We reaching a conclusion that Gaussian distribution should be the best distribution for a image residual generated by denoising method. Furthermore, High-order statistics are introduced as constraint conditions for residual Gaussian. Finally, the statistical properties of $\mathbf{MSE}$ are analyzed. Residual mean value is introduced in the loss function as a constraint condition for residual variance unbiased estimation. Experimental results in Gaussian noise and salt-and-pepper noise show that our method has better robustness under low SNR and outperforms several other existing state-of-the-art algorithms such as BM3D, NLM, TV and DAE. The restored image obtained by our algorithm has better structural similarity, which means it can retain the real image information while removing noise more effectively.

B. Future work

To improve generalization ability of our method, the DAE can be trained on dataset which mixed the image samples with different noise levels and types.

Since the speech signal also has structural information, the IDEA is not only limited to image denoising, but also can be used in speech denoising.

1. References

[1] B.,Zhang, J. P. Allebach, "Adaptive bilateral filter for sharpness enhancement and noise removal", IEEE Trans. Image Processing, 2008, 17, (5), pp. 664–678

[2] O. Hari, B.Mantosh, "An improved image denoising method based on wavelet thresholding", Journal of Signal and Information Processing, 2012, 3, (1), pp. 109–116

[3] W. Habib, T. Sarwar, A. M. Siddiqui, I. Touqir, "Wavelet denoising of multiframe optical coherence tomography data using similarity measures", IET Image Processing, 2017 , 11, (1), pp. 64 - 79

[4] D. Mahdi, N. F. Isabel, R. G. Gil, "Spatially adaptive total variation deblurring with split Bregman technique", IET Image Processing, 2018, 12, (6), pp. 948 - 958

[5] S. Ivan, L. G. Harry, S. P. Douglas, B. Randall, "Simultaneous low-pass filtering and total variation denoising", IEEE Trans. Signal Processing, 2014, 62, (5), pp. 1109–1124

[6] Z. Huang, Q. Li, H. Fang, " Iterative weighted sparse representation for X-ray cardiovascular angiogram image denoising over learned dictionary ", IET Image Processing, 2018, 12, (2), pp. 254 - 261

[7] G. Shan, W. Long, Y. Guowei, "Sparse representation based on vector extension of reduced quaternion matrix for multiscale image denoising ", IET Image Processing, 2016, 10, (8), pp. 598 - 607

[8] J. Ji, F. Ren, HF. Ji, YF. Yao , GF Hou, "Generalised non-locally centralised image de-noising using sparse dictionary", IET Image Processing, 2018, 12, (7), pp. 1072 - 1078

[9] A. Buades, B. Coll, J.M.Morel, "A non-local algorithm for image denoising", IEEE Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, pp. 60–65

[10] J. Darbon, A. Cunha, TF. Chan, S. Osher, GJ. Jensen,"Fast nonlocal filtering applied to electron cryomicroscopy", IEEE Inter Symposium. Biomedical Imaging: from Nano to Macro, Paris, France, 2008, pp.1331–1334.

[11] K. Dabov, A. Foi, V. Katkovnik, "Image denoising by sparse 3-D transform-domain collaborative filtering", IEEE Trans. Image Processing, 2007, 16, (8), pp. 2080–2095,

[12] YJ. Li, J. Zhang, J. Wang, "Improved BM3D denoising method", 2017, 11, (12), pp. 1197 – 1204

[13] J. Xie, L. Xu, E. Chen, "Image denoising and inpainting with deep neural networks", Advances in neural information processing systems, Nevada, USA, 2012, pp. 341–349.

[14] X. J. Mao, C. Shen, Y. B. Yang, "Image restoration using very deep convolutional Encoder-Decoder networks with symmetric skip connections ", https://arxiv.org/pdf/1603.09056, accessed 1 Sep 2016

[15] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising", IEEE Trans. Image Processing, 2017, 26, (7), pp. 3142–3155

[16] V. Pascal, L. Hugo, L. Isabelle, M. Pierre-Antoine: "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion", Journal of Machine Learning Research, 2010, 11, (12), pp. 3371–3408

[17] F. Zhang, N. Cai, J. Wu, G. Cen, H. Wang, "Image denoising method based on a deep convolution neural network", IET Image Processing, 2018, 12, (4), pp. 485 – 493

[18] G. Ian, B. Yoshua, C. Aaron, "Deep Learning", MIT Press, Cambridge, MA, USA, 2017, pp. 512–517

[19] Z. Wang , AC. Bovik , HR. Sheikh, EP. Simoncelli, "Image quality assessment: from error visibility to structural similarity", IEEE Trans. Image Process, 2004, 13, (4), pp. 600–612

[20] P. Athanasios, "Probability, random variables, and stochastic processes", McGraw-Hill, NewYork, USA, 2002, 3rd ed., pp. 533–569.

[21] Y. LeCun, C. Cortes, "The MNIST database of handwritten digits", http://yann.lecun.com/exdb/mnist/, accessed, 1998References.