CONVOLUTIONAL AUTOENCODER FOR SINGLE IMAGE DEHAZING

Rongsen Chen and Edmund M-K Lai

Department of Information Technology & Software Engineering
Auckland University of Technology, Auckland, New Zealand

ABSTRACT

In this paper, we present a Convolutional AutoEncoder (CAE) for single image dehazing. Our CAE makes us of Densely Connection Networks as its encoder and decoder. It is trained with the corresponding hazy and clean images at the input and output, enabling it to remove the haze without having to rely on an atmospheric scattering model. The CAE is trained and tested with the RESIDE dataset. Experiment results show that this CAE outperforms eight state-of-art methods. The trained CAE is also applied to some real-life hazy images, and decent dehazing results are obtained. Moreover, our method is computationally efficient enough to run on computers without GPU units.

Index Terms — Single image dehazing, autoencoder, convolutional neural network

1. INTRODUCTION

Hazy weather conditions such as those involving fog and mist could greatly lower the visibility and clarity of the scenes captured by a camera. Haze could cause the best computer vision algorithms for object detection and tracking that were trained with clear images to perform poorly. The simplest solution to such problems is to include hazy images in the training data. However, it is infeasible to collect images under all kinds of different hazy weather conditions for all the classes of images in the training set, especially when the number of classes is large. An alternative solution is to remove the haze to obtain relatively clean images.

Dehazing through the use of a single image is a challenging task. Several single image dehazing algorithms have been developed in the past decade [1-6]. They are based on a model of how hazy images are produced from a clean image. This model, known as the atmospheric scattering model, can be expressed mathematically as

$$I(x) = J(x)t(x) + \alpha(x)\left[1 - t(x)\right] \qquad (1)$$

where I is the observed hazy image, J is the clean image to be recovered, t is the medium transmission map, α is the global atmospheric light, and x are the pixel locations in the images. Based on this model, J could be recovered from I if the medium transmission map and the global atmospheric light for the given image is known. Thus a major part of these algorithms involves the estimation of these parameters.

Recently, deep neural network (DNN) based dehazing methods [1-4] have been proposed. In these methods, DNNs are used to find the optimal values of t and α for the given hazy images. Experimental results show that DNNs are able to produce more accurate estimates. However, due to the fact that it is generally difficult to estimate the perfect t and α with single image dehazing, the dehazed images often exhibit strong artifacts.

To overcome this problem, more recent research [5, 6] have proposed the use of DNN methods that perform single image dehazing directly without assuming an atmospheric scattering model. The results reported in [6] is very good for the NITRE challenge dataset [13]. However, it has been shown that this DNN severely overfits the small test data of this challenge [7]. Nevertheless, it still shows that CNN based methods can perform effective dehazing without the atmospheric scattering model.

This paper presents a convolutional autoencoder (CAE) that can perform effective single image haze removal, without needing the help of the traditional atmospheric scattering model. The fundamental idea is that hazy images can be seen as images that have been heavily affected by noise (haze). Therefore, if we train a CAE with a set of clear and hazy images, the CAE should be able to effectively remove haze from the hazy images just like it is able to remove noise from noisy images. To illustrate the effectiveness of this CAE, its performance is compared with other published state-of-the-art methods using the RESIDE standard dataset [19] as well as some real world images.

The rest of this paper is structured as follows. Section 2 gives an overview of convolutional autoencoders and densely connected networks on which our network architecture is based. Our proposed autoencoder is then described in detail in Section 3. In Section 4, the computational experiments are described, together with the results that compare the performance of our method to eight other representative methods. Finally, the conclusions are presented in Section 5.

2. RELATED WORKS

2.1. Convolutional AutoEncoder

Autoencoder is a type of artificial neural network that is capable of learning the representation of the given data through an encoding and decoding process in an unsupervised manner. A Convolutional AutoEncoder is an *AutoEncoder* where its encoder and decoder are convolutional neural networks. It has been proven to be useful in image denoising tasks. For instance, in [8], good denoising performance with a small medical image dataset is achieved using a CAE. Another example can be found in [9] where a CAE with symmetric skip connections outperforms most of the (at the time) state-of-art denoising methods.

2.2. Densely Connected Network

The Densely Connected Network (DenseNet) proposed in [10] is a CNN that is capable of overcoming the vanishing gradient problem. It also enables feature reuse and is able to learn effectively with fewer parameters. It has been shown to work well with relatively small training sets.

Recently, in [4], Zhang and Patel combined DenseNet and U-Net to build a Generative Adversarial Network (GAN) to jointly predict the medium transmission map and the global atmospheric light of a given hazy image. The result is a much more accurate estimation compared to previous methods [1-3], and thus produced better dehazing results. In [6], DenseNet and the Residual Network (ResNet) are combined to build a specially designed generative network. This network achieved first place in the *NTIRE 2018 Dehaze Challenge* [11] in the indoor task and second place in the outdoor task. In the same challenge, a network called DenseNet for Dehaze (DND) is ranked third place in the indoor task and first place in the outdoor task. These works showed that DenseNet is a very suitable type of CNN for image dehazing.

3. METHOD

3.1. Overall Architecture

The main feature of our proposed CAE for single image haze removal is that the mapping between the hazy and the clean images is learnt directly without having to learn the atmospheric scattering model in (1), in contrast to most previous CNN based methods [1-4][3-6]. Hence, our CAE network takes a set of hazy images as input and directly generate the corresponding haze-free images as output.

Figure 1 illustrates the overall architecture of our CAE for an input image size of 512×512 pixels. The input to the encoder of the CAE is a hazy image. The encoder extracts the desired features through a series of convolutional layers. The decoder then takes the extracted feature maps as input, and use them to reconstruct the desired output, which in this case, is the haze-free version of the input.

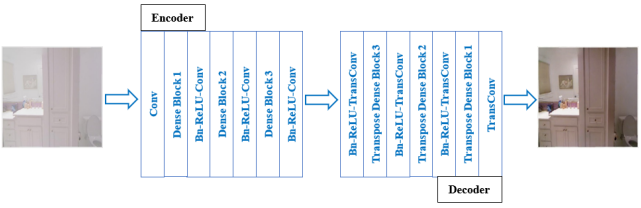

Figure 1. The overall architecture of our *CAE*

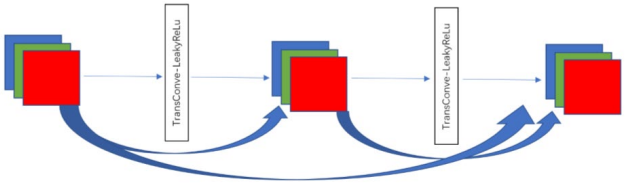

Figure 2. Overview of the Transpose Dense Block

TABLE I. THE DETAIL OF THE CONVOLUTIONAL AND TRANSPOSE CONVOLUTIONAL SETTING IN THE NETWORK

	Layer Name	Output Size	Layer setting
Encoder	Conv	$256 \times 256, 16$	$7 \times 7\ conv, strid\ 2$
	Dense Block 1	$256 \times 256, 34$	$\begin{pmatrix} 1 \times 1\ conv, strid\ 1 \\ 3 \times 3\ conv, strid\ 1 \end{pmatrix} \times 3$
	Transition 1	$256 \times 256, 32$	$1 \times 1\ conv, strid\ 1$
	Dense Block 2	$256 \times 256, 56$	$\begin{pmatrix} 1 \times 1\ conv, strid\ 1 \\ 3 \times 3\ conv, strid\ 1 \end{pmatrix} \times 4$
	Transition 2	$128 \times 128, 48$	$1 \times 1\ conv, strid\ 2$
	Dense Block 3	$128 \times 128, 78$	$\begin{pmatrix} 1 \times 1\ conv, strid\ 1 \\ 3 \times 3\ conv, strid\ 1 \end{pmatrix} \times 5$
	Transition 3	$128 \times 128, 48$	$1 \times 1\ conv, strid\ 1$
Decoder	Transition' 3	$128 \times 128, 48$	$5 \times 5\ transconv, strid\ 1$
	Transposed Dense Block 3	$128 \times 128, 78$	$\begin{pmatrix} 5 \times 5\ conv, strid\ 1 \\ 5 \times 5\ conv, strid\ 1 \end{pmatrix} \times 5$
	Transition' 2	$256 \times 256, 32$	$5 \times 5\ transconv, strid\ 2$
	Transposed Dense Block 2	$256 \times 256, 56$	$\begin{pmatrix} 5 \times 5\ conv, strid\ 1 \\ 5 \times 5\ conv, strid\ 1 \end{pmatrix} \times 4$
	Transition' 1	$256 \times 256, 16$	$5 \times 5\ transconv, strid\ 1$
	Transpose Dense Block 1	$256 \times 256, 34$	$\begin{pmatrix} 5 \times 5\ conv, strid\ 1 \\ 5 \times 5\ conv, strid\ 1 \end{pmatrix} \times 3$
	TransConv	$512 \times 512, 3$	$5 \times 5\ transconv, strid\ 2$

The encoder part of the CAE starts with a convolutional layer, followed by three dense blocks. There is also a transition blocks after each dense block. Each transition block consists of a series of three processes – batch-normalization [12], ReLU [13], and convolution.

The decoder basically has the same structure as the encoder, with the convolution operations replaced by transposed convolution operations [14] and larger size filters. Thus the

dense blocks in the decoder are called transposed dense blocks. A transposed dense block is illustrated in Figure 2. Since the transposed convolution operation is the inverse of convolution operation, the decoder has the ability to restore the convolutional results to its previous form. Moreover, since the transposed convolution has learnable filters that can be adjusted through backpropagation, it is capable not only of restoring the data, but also restoring the data in a way that we want it to present. Therefore, the decoder has the ability to generate haze-free images from the hazy inputs if it is trained properly. Table I provides the dimensional details of each layer.

3.2. Loss Function

Choosing an appropriate loss function is a critical part in designing CNNs as it directly affects the quality of learning during the backpropagation process. It is known that using only L_2 loss functions will produce blurry outputs [6]. Furthermore, recent research has shown that using multiple loss functions in DNN based dehaze network will help improve the quality of the result [15]. Based on this knowledge, a combined loss function is chosen for training our CAE. More specifically, this combined loss function is given by

$$L = L_2 + L_f \tag{2}$$

where L_2 is the traditional L_2 loss function defined by

$$L_2 = \sum_{i=1}^{n} \left(I_i - J_i \right)^2 \tag{3}$$

Here, I is the ground truth (clean) image, J is the dehazed image and n is the number of pixels.

The second part of (2), L_f, is the *perceptual loss* introduced in [16]. *Perceptual loss* has been selected because it enables the CAE to identify the feature loss during image reconstruction in the decoding step. It is defined as

$$L_f = \sum_{i=1}^{n} \left(\left\| V\left(I_i\right) - V\left(J_i\right) \right\|_2^2 + \left\| G\left(V\left(I_i\right)\right) - G\left(V\left(J_i\right)\right) \right\|_2^2 \right) \tag{4}$$

where V represents a CNN structure that extracts the low-level features of the given ground truth image I and the dehazed result J. G is the *grim matrix* [17] and i is the selected layer of the CNN structure. Similar to previous works [4, 5] we use a pretrained VGG-16 network to extract the low-level features. Furthermore, the selected layer l has been set to conv 3_1.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setting

4.1.1. Dataset

Most of the previous CNNs based dehazing methods [1-4] were trained with customized synthesized haze dataset. As the dataset is customized it could give the proposed method advantage in comparison to other methods, which is not fair. To overcome this unfairness, we will use the benchmark dataset that has recently has been introduced to the research community called the RESIDE dataset [17]. The RESIDE dataset has two versions – the RESIDE standard, and the RESIDE-β. The one we used in our experiments is the RESIDE standard dataset.

The RESIDE standard dataset contains one training set and two test sets. More specifically, the training set called Indoor Training Set (ITS) contains 13990 synthetic indoor hazy images, generated using 1399 clear images from NYU2 [18] and Middlebury stereo [19]. The first test set, known as the Synthetic Objective Testing Set (SOTS) contains 500 synthetic indoor hazy images, generate using images in NYU2 that are different from the training images. The second test set is known as the Hybrid Subjective Testing Set (HSTS), which contains 10 synthetic outdoor hazy images, and 10 real-world outdoor images to evaluate qualitative visual performance.

4.1.2. Training setting

We use ADAM [20] to help optimize our CAE during training, with a standard learning rate of 1×10^{-4}. All training images are in size of 460×620 pixels as provided in the dataset. We feed the training images into the CAE with a batch size of 1 and train the CAE for 20 epochs.

4.1. Result

Table II shows the SSIM and PSNR scores of our CAE compare to eight other state-of-the-art methods [1-3, 21-25] for the SOTS test set. The results of previous methods are from [17]. Nonetheless, since we are running our method on the exact same dataset, the compare is fair. As shown in this table, our CAE has significantly outperformed all the other methods. More specifically, the PSNR score for our CAE is 3.42 higher than the second highest score (the DehazeNet) and the SSIM score is 0.0622 higher than the second highest score (the GRM). These results show that our CAE is much more capable than other methods listed in the table, in dehazing indoor hazy image.

To further test whether our CAE has overfitted the indoor synthetic hazy environment after training, we directly apply the trained network to the HSTS test set. In contrast to ITS, the HSTS contains outdoor synthetic hazy images only. Thus we are testing the performance of the network trained with indoor images on outdoor images. The results are shown in Table III. In this case, the DehazeNet achieved the highest PSNR and SSIM scores while our CAE is third highest, with AOD-Net being second.

TABLE II. DEHAZE RESULT ON SOTS test set.

	DCP [2]	FVR [22]	BCCR [23]	GRM [24]	NLD [25]	DehazeNet [1]	MSCNN [2]	AOD-Net [3]	Our CAE
PSNR	16.10	17.18	16.91	18.64	17.52	21.14	17.57	19.06	**24.56**
SSIM	0.8158	0.7483	0.7913	0.8553	0.7489	0.8472	0.8102	0.8504	**0.9126**

TABLE III. DEHAZE RESULT ON HSTS test set

	DCP [2]	FVR [22]	BCCR [23]	GRM [24]	NLD [25]	DehazeNet [1]	MSCNN [2]	AOD-Net [3]	Our CAE
PSNR	14.84	14.48	15.08	18.54	18.92	**24.48**	18.64	20.55	**20.08**
SSIM	0.7609	0.7624	0.7382	0.8184	0.7411	**0.9153**	0.8168	0.8973	**0.8169**

TABLE IV. Average run-time per image

	DCP [2]	FVR [22]	BCCR [23]	GRM [24]	NLD [25]	DehazeNet [3]	MSCNN [2]	AOD-Net [3]	Our CAE
Time (s)	1.62	6.79	3.85	83.96	9.89	2.51	2.60	**0.65**	**1.13**

INPUT NLD [25] MSCNN [2] DehazeNet [1] AOD-Net [3] Our CAE

Figure 3. Example of dehaze results on real-world hazy image from HSTS test set

The reason why our CAE performance dropped for the outdoor images can be attributed to the fact that the light conditions between outdoor and indoor environment are vastly different. Hence the extracted feature maps will exhibit huge differences, impacting the recovery process and consequently produce worse results.

While the objective scores for our CAE for outdoor images are not the best, Figure 3 shows how it behaves compared with 5 other methods with two real-world outdoor hazy images. As can be observed, the methods from [1-3] failed to remove haze and the one from [25] produces visually poor results. It seems that DehazeNet [1] and AOD-Net [3], the two methods that outperform our CAE in Table III, is overfitting the synthetic hazy conditions. Meanwhile, our CAE is able to produce a decent result. Therefore, it is confident to say our method is better in real-world situations.

Table IV shows the average run-time per image on an Intel Core i7 processor with no GPU support. It shows that our CAE requires on average 1.13s per image, which is the second fastest among the methods compared. Thus it is among the most computationally efficient methods for single image dehazing.

5. CONCLUSIONS

This paper presented a CAE with DenseNet as encoder and decoder for effective single image dehazing. Experimental results have demonstrated that it outperforms other state-of-the-art methods, especially in dehazing real-world hazy images. Furthermore, our CAE is computationally efficient enough to run on computers without GPU support.

6. REFERENCES

[1] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, 2016, pp. 5187-5198.

[2] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," *Computer Vision* -

ECCV 2016, Lecture Notes in Computer Science, vol. 9906, Springer, pp. 154-169.

[3] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-Net: All-in-one dehazing network," *IEEE International Conference on Computer Vision*, 2017, pp. 4780-4788.

[4] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3194-3203.

[5] L. T. Goncalves, J. D. O. Gaya, P. Drews, and S. S. D. C. Botelho, "DeepDive: An End-to-End Dehazing Method Using Deep Learning," *30th SIBGRAPI Conference on Graphics, Patterns and Images* (SIBGRAPI), 2017, pp. 436-441.

[6] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1015-1024.

[7] Y. Gandelsman, A. Shocher, and M. Irani, "Double-DIP: Unsupervised image decomposition via coupled deep-image-priors," *arXiv preprint* arXiv:1812.00467, 2018.

[8] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," *IEEE 16th International Conference on Data Mining Workshops* (ICDMW), 2016, pp. 241-246.

[9] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," *Advances in Neural Information Processing Systems*, 2016, pp. 2810-2818.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261-2269.

[11] C. Ancuti et al., "NTIRE 2018 challenge on image dehazing: Methods and results," *IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2018, pp.891-901.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd International Conference on Machine Learning*, 2015, pp. 448-456.

[13] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *29th International Conference on Machine Learning*, 2013, 5 pages.

[14] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2528-2535.

[15] H. Sim, S. Ki, J.-S. Choi, S. Seo, S. Kim, and M. Kim, "High-resolution image dehazing with respect to training losses and receptive field sizes," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 912-919.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Computer Vision - ECCV 2016*, Lecture Notes in Computer Science, vol. 9906, Springer, pp. 694-711.

[17] B. Li et al., "RESIDE: A benchmark for single image dehazing," *arXiv preprint* arXiv:1712.04143, 2017.

[18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," *Computer Vision - ECCV 2012*, Lecture Notes in Computer Science, vol. 7576, pp. 746-760: Springer.

[19] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," *18th International Conference on Pattern Recognition*, 2006, 4 pages.

[20] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *3rd International Conference on Learning Representations* (ICLR), 2015.

[21] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, 2011, pp. 2341-2353.

[22] J.-P. Tarel and N. Hautiere, "Fast visibility restoration from a single color or gray level image," *IEEE 12th International Conference on Computer Vision*, 2009, pp. 2201-2208..

[23] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization," *IEEE International Conference on Computer Vision*, 2013, pp. 617-624.

[24] C. Chen, M. N. Do, and J. Wang, "Robust image and video dehazing with visual artifact suppression via gradient residual minimization," *Computer Vision - ECCV 2016*, Lecture Notes in Computer Science, vol. 9906, Springer, pp. 576-591.

[25] D. Berman and S. Avidan, "Non-local image dehazing," *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1674-1682.