Text Image Super Resolution Using Deep Attention Neural Network

Yun Liu
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
yuinn@asagi.waseda.jp

Remina Yano
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
yano.remina@toki.waseda.jp

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

Takuya Suzuki
SHARP Corporation
Chiba, Japan
takuya.suzuki@sharp.co.jp

Takeshi Chujoh
SHARP Corporation
Chiba, Japan
chujoh.takeshi@sharp.co.jp

Tomohiro Ikai
SHARP Corporation
Chiba, Japan
ikai.tomohiro@sharp.co.jp

***Abstract*—In this paper, we propose a super-resolution method for text images to improve the accuracy of optical character recognition (OCR). The accuracy of OCR is closely related to the resolution of the image, and when OCR is applied to low resolution text images, satisfactory results are often not obtained. In the proposed method, we extract more representative feature information from text images by combining channel and spatial attention. Furthermore, we propose a new loss function called "edge loss". Experimental results show that the recognition accuracy of text images by our SR method is 5.87% higher than that of the original low-resolution images, and also higher than the results of BICUBIC and the baseline model.**

Keywords—super resolution, Optical Character Recognition (OCR), text image, attention, convolutional neural network

I. INTRODUCTION

OCR is an important task in computer vision and has a wide range of applications. It can be used in card number recognition, license plate recognition, document content extraction and so on. However, one of the prerequisites for character recognition accuracy is that the recognized image is clear enough, and if the text image to be recognized is of low resolution, then the recognition accuracy will not be satisfactory. An intuitive idea is to introduce super-resolution method, which pre-processes the low-resolution text images to generate more clear ones, and then we can perform OCR operations on the processed images. In this paper, we propose a super-resolution approach for text images. Our method combines channel attention and spatial attention mechanisms to enhance the expressiveness of neural network. In addition, we use an edge loss to reconstruct images with sharper edges, thus making the OCR process easier.

II. REALTED WORKS

A. Image Super resolution

Image super-resolution methods fall into three main categories, interpolation-based, reconstruction-based and learning-based methods. In recent years, with the development of artificial intelligence, a number of deep learning-based methods have been proposed, achieving impressive results. SRCNN [1], VDSR [2] , SRGAN [3] are some of the pioneer works. RCAN [4] brings channel attention mechanism into super-resolution networks, which improves the expressive power of the network and further improves the performances, and it is also the baseline method used in this paper. Although there are countless image super-resolution methods, approaches specifically for text images are rare. The method proposed in this paper is targeted at text images and can improve the accuracy of optical character recognition of text images.

B. Attention Mechanism

Originally used in machine translation, the attention model (AM) has now become an important concept in the field of neural networks. It has recently been introduced to the field of computer vision as well. SENet [5], the champion of the ImageNet 2017 image classification challenge, proposes a channel-based attentional SE block that automatically learns the importance of each feature channel when extracting features, and then focuses on more useful features according to that importance and suppresses features that are less useful. The SE block can be simply embedded in existing networks to improve the feature extraction capability and optimize the overall performance of the task. CBAM [6], an attention model proposed in 2018, adds an additional spatial attention module compared to SENet, which considers the importance of features in different regions of the image, encouraging the neural network to focus on key information. In our work, we include the spatial attention module because we want the network to focus more on text areas and suppress background areas.

C. Loss functions for image super resolution

L1 loss, also known as Least Absolute Deviation (LAD) , is calculated as the mean of the sum of the absolute differences of the corresponding pixel values between the actual image and the target image. L2 loss, also known as mean squared error loss (MSE), is calculated as the mean of the sum of the squares of the corresponding pixel differences between the actual image and the target image. Perceptual loss is calculated by comparing the difference between the real image and the generated image by feeding them into a trained feature network to obtain the respective feature maps. The use of perceptual loss can improve the subjective visual perception, but the PSNR may be compromised. In the super- resolution

978-1-6654-3676-2/21/$31.00 ©2021 IEEE

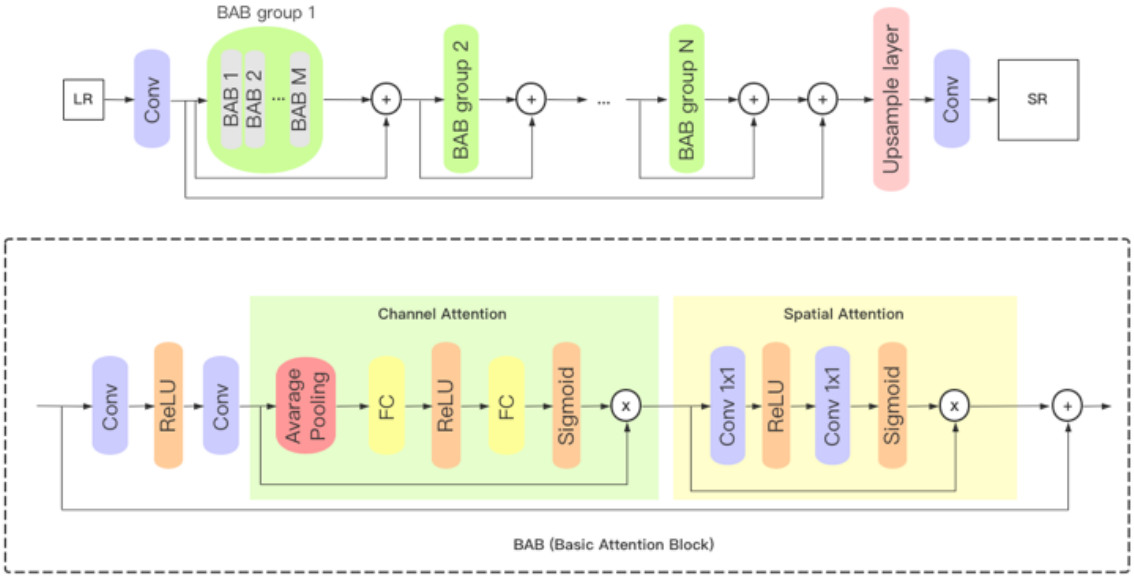

Fig. 1. Network structure of the proposed method.

approach using generative adversarial networks, another type of loss, adversarial loss, is also used to generate images that are close to the real high-resolution image through an adversarial process of generators and discriminators. In this paper we use a combination of L1 loss and a new edge loss to achieve a better reconstruction by forcing the edge map of the generated image to be closer to that of the real image.

III. Proposed Method

Our baseline model is RCAN [4] and we have made the following adjustments: a spatial attention module has been added and a new edge loss function has been used.

A. Model Structure

The structure of the model is shown in Figure 1, where the input image is a low-resolution text image LR and the output is the reconstructed image. A convolution is first performed on the input image to extract the shallow features, and then there are N BAB groups, each with M BABs (Basic Attention Block). Every BAB contains a channel attention and a spatial attention module, and the two modules are connected sequentially. The channel attention module allows the neural network to focus on more important channel features, while the spatial attention forces the network to concentrate on more important features at spatial locations. Finally we use the sub-pixel convolution to enlarge the image.

B. Loss Function

The loss function used in this paper is a combination of L1 loss and edge loss.

$$L1(Y,\hat{Y}) = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||Y(i,j) - \hat{Y}(i,j)|| \quad (1)$$

$$L_{edge}(Y,\hat{Y}) = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||S[Y](i,j) - S[\hat{Y}](i,j)|| \quad (2)$$

$$L_{total} = L1 + L\_edge \quad (3)$$

Where Y is the generated SR image and $\hat{Y}$ represents the real high-resolution image (ground truth), S [*] stands for the edge map extracted by Sobel edge extractor.

IV. Experiment

A. Dataset

The dataset used in this paper is ICDAR 2015 TextSR [7], which consists of 708 sets of images extracted from French TV videos, each group includes one HD image, one HR image (2x down-sampling result), and one LR image (4x down-sampling result). The dataset was divided into a training set consisting of 567 images and a test set containing 141 images.

B. Evaluation Methods

We used PSNR and SSIM, which are commonly used in super-resolution tasks, as evaluation tools. In particular, we also calculated the OCR accuracy of each method.

PSNR (Peak Signal-to-Noise Ratio) is often used as a measure of signal reconstruction quality in areas such as image compression, and is an important tool for objectively evaluating the performance of super-resolution, and it can be defined in terms of Mean Squared Error (MSE).

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||X(i,j) - Y(i,j)||^2 \quad (4)$$

$$PSNR = 10 \cdot log_{10}\frac{MAX_I^2}{MSE} \quad (5)$$

SSIM (structural similarity) is a measure of image similarity and can also be used to determine the quality of a compressed image. It measures the similarity between images in terms of brightness, contrast and structure.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (6)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (7)$$

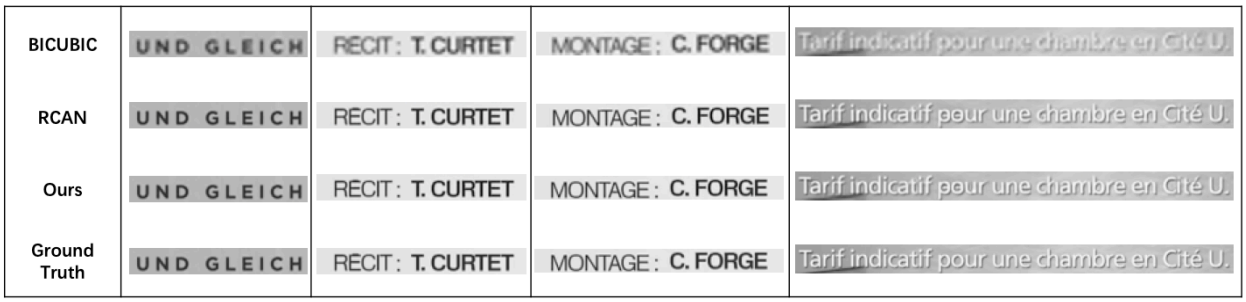

Fig. 2. Qualitative comparison results.

$$s(x,y) = \frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3} \tag{8}$$

$$SSIM(x,y) = f\big(l(x,y),c(x,y),s(x,y)\big) \tag{9}$$

$$= \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2\sigma_y^2+C_2)}$$

$\mu_x, \mu_y, \sigma_x, \sigma_y$ are the mean, variance of x and y, respectively; σ_{xy} is the covariance of x and y; C_1, C_2 and C_3 are constants.

OCR accuracy is an evaluation metric used for text super-resolution scenarios. By comparing the OCR results of text images generated by different methods, it is possible to know which SR method works better. The OCR tool used in this paper is pytesseract 0.3.7, a python library for Tesseract-OCR.

$$OCR_{acc} = 1 - \frac{d}{K} \tag{10}$$

K is the total number of characters, d is the edit distance between two strings, i.e. the total number of operations required to convert one string to another, these operations include deletion, substitution, addition, etc.

TABLE I. RESULTS OF DIFFERENT METHODS.

Method	Evaluation Indicator		
	PSNR/dB	***SSIM***	***OCR accuracy/%***
LR (without SR)	-	-	78.29
BICUBIC	23.51	0.9368	78.70
RCAN	**37.63**	**0.9968**	83.03
OURS	37.58	0.9965	**84.16**
GROUND TRUTH	-	-	85.42

C. Training details

During training, our up-sampling scale is set to 2. M and N in the model structure are set to 20 and 10, respectively. And the batchsize is set to 32. We then trained the network for 400,000 iterations using the ADAM optimizer on a Tesla T4 GPU. The learning rate was initialized to 0.0001 and reduced by half for every 100,000 iterations.

D. Result Analysis

As we can see from Table I, the super-resolved text images can achieve higher recognition accuracy compared to the low-resolution text images without any processing. Among these methods, although the PSNR and SSIM of images generated by our method are slightly lower than RCAN, they are much better than BICUBIC, and it is worth mentioning that our method achieves the highest OCR accuracy, which is 5.46% and 1.13% higher than BICUBIC and RCAN methods, respectively. Experiments show that our super-resolution method is effective in improving the accuracy of low-resolution text recognition, and the accuracy of image recognition processed by this method is close to that of recognition using real high-resolution text images. Some of the qualitative results of several methods can be viewed in Figure 2.

V. CONCLUSION

In this paper, we propose a super-resolution approach for text images, aiming to improve the OCR accuracy of low-resolution text images. Channel attention and spatial attention modules are embedded in the deep network to improve the expressive power of the neural network and make the network focus on the textual features in the image. Furthermore, in addition to the L1 loss function, an edge loss is used to assist the reconstruction by forcing the edge information of the generated image and the target image to be closer, so that sharper edges can be obtained, thus benefiting the optical character recognition process.

REFERENCES

[1] C. Dong, C. C. Loy, K. He and X. Tang, "Image super-resolution using deep convolutional networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 1 Feb. 2016.

[2] J. Kim, J. K. Lee and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646-1654, June 2016.

[3] C. Ledig et al., "Photo-realistic single image super-resolution using a Generative Adversarial Network," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105-114, July 2017.

[4] Y. Zhang, K. Li, K. Li, et al. , "Image super-resolution using very deep residual channel attention networks," proceedings of the European conference on computer vision (ECCV), pp. 286-301, Sep. 2018.

[5] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, Aug. 2020.

[6] S. Woo, J. Park, J. Lee, et al., "Cbam: Convolutional block attention module," proceedings of the European conference on computer vision (ECCV), pp. 3-19, Sep. 2018.

[7] C. Peyrard, M. Baccouche, F. Mamalet and C. Garcia, "ICDAR2015 competition on text image super-resolution," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1201-1205, Nov. 2015