End-to-End Image Super-Resolution via Generative Adversarial Network

Xu Li¹*, Xiangdong You²

¹Key Laboratory Of Universal Wireless Communications, Ministry Of Education, Beijing University Of Posts and Telecommunications, Beijing, China
²Key Laboratory Of Universal Wireless Communications, Ministry Of Education, Beijing University Of Posts and Telecommunications, Beijing, China

*Corresponding author e-mail: lx518@bupt.edu.cn

Abstract. Image super-resolution is to extract information from single or more low-resolution images and use this information to get the corresponding high-resolution image. The ability to capture high texture details of low-resolution images is one of the impressive advantages of generative adversarial networks (GANs). This paper mainly studies the image reconstruction method based on a single low-resolution image, and builds an end-to-end image Super-Resolution method via generative adversarial networks to improve the image resolution. Convolution neural networks use the mean square error (MSE) as loss function. For such loss function, structural similarity (SSIM) and peak signal to noise ratio (PSNR) can be obtained, but high texture details of original images are often lost. The generative adversarial network mainly contains two parts: generator network and discriminator network. Adversarial loss uses a discriminator network to push our solution to a natural image manifold, which is trained to distinguish super-resolution images from photo-realistic images. Additionally, content loss of our network was indicated by perceptual similarity rather than similarity of pixel-wised space.

1. Introduction

Single image super-resolution (SISR) aims to generate high-resolution image with clear texture and good visual effect from its downsampling low-resolution (LR) image. Research of image SR can help to extract more information that hidden in the LR images and restore more accurate high-resolution (HR) image.

Most super-resolution learning-based methods have more concerned about the delivered and community superior performance in super-resolution. Learn a function to represent mapping between LR and original images using auxiliary data is basic idea of these methods. With the wide application of deep learning, researchers have also proposed a variety of image resolution methods based on deep neural networks. For most learning-based methods, minimization of the MSE between the recovered HR image and the ground truth is the common optimization target. Minimizing MSE also maximizes the PSNR, for which is a common method to evaluate and compare SR algorithms. However, The MSE which is defined based on pixel-wised space is very limited to capture perceptually relevant differences [1, 2].

In this paper, we construct a fully end-to-end SR system with the generative adversarial network, in which we performed the upsampling operation in a deep feature space. We jointly optimized with the
other modules of our system. Super-resolution generative adversarial network (which is also called SRGAN) [3] proposed a novel loss named perceptual loss which is not only good at optimizing the MSE, but also can capture more high texture details. We also evaluated the impact of the existence of the discriminator on the final result and compared different neural networks as discriminators and layers of the network to get better visual effects.

2. Related Work

We consider interpolation methods [4, 5], reconstruction methods [6, 7] and learning methods [1, 2] as the main kind of Super-Resolution problems solution. SR methods like bicubic, linear or Lanczos filtering [5] are simple to calculate and can be very fast to get the reconstructed images, but simplify the single SR problem and usually obtain pictures with very simple and smooth textures. Many powerful learning-based methods formulate SR as a nonlinear mapping relations. After the corresponding original images prepared, we train these mapping functions based on example-pairs which totally rely on LR training patches.

Deep neural networks can provide more superior performance if there were complex architectures, which make the image texture richer and the training process more challenging. SR problems commonly use deep-learning methods and have delivered a compelling performance. In [8], a deep learning method based on convolutional neural network is proposed for image super-resolution, which comprises three convolution layers corresponding to input information patch extraction, image vector non-linear mapping and reconstruction modules, respectively. In [9], the generative adversarial networks (GANs) which is first proposed by Goodfellow et al. has been applied to many areas and achieved good results. In Denton et al. [10] and Mathieu et al. [11], the authors employ generative adversarial networks to solve this topic for application of image generation. Johnson et al. [12] and Bruna et al. [13] found another way instead of low-level pixel-wise error measures, for which they raise idea of extracting feature from the prepared VGG network. Generative adversarial network was used to solve the SR and proposed a novel perceptual loss.

3. Methods

To construct a fully end-to-end method, we mainly concentrate on three aspects: generator, discriminator and the optimization target. In SISR, the aim is to restore image $^{\text{SR}}$ from image $^{\text{LR}}$. For our system, the generator will super-resolve a high-resolution image from a LR image to cheat the discriminator, however, the discriminator will also improve the ability to distinguish the real high-resolution images from the generated images.

3.1. Feature Extraction

For most traditional methods, feature extraction module is used to get first and second-order gradients of input patches, such ticks were equivalent to use some well-designed high pass filters to filter the input. In our methods, the local features of high texture details will be extracted automatically rather than manually designing these high pass filters. Whether it is shallow or deep models, most prior methods are based on upsampling LR to the corresponding HR with bicubic interpolation. The extra bicubic interpolation is to obtain a benchmark picture in which pixel size is exactly the same as original high-resolution images. Obviously, this will lead to even damage important low-resolution information. Therefore, we adopt a special method and make feature extraction rely on original low-resolution images information with the output vector of convolution layers.

The feature extraction module in our methods consists of convolution layer and corresponding to Parameters Rectified Linear Unites (pReLUs) [14] make a non-linear mappings. The convolution layer can be expressed as the fellow equation:

$$ F_l = \begin{cases} \alpha_l(W_l * F_{l-1} + b_l), & \text{if } W_l * F_{l-1} + b_l \leq 0 \\ W_l * F_{l-1} + b_l, & \text{if } W_l * F_{l-1} + b_l > 0 \end{cases} $$

in which $b_l$ and $W_l$ denote bias of the $l$-th layer and the weights of convolution filter, respectively. At the same time, $F_l$ represent output feature-map for $l$-th layer, $F_0$ denoting low-resolution images. And kernel size of the convolution layer is $9 \times 9$ and it will generate feature maps of 64 channels.
3.2. Combining Generator and Discriminator

The ultimate target of our system is to train the generation function $G$, for the corresponding input, the super-resolution can be generated. The final result we want to achieve is to use the generated image generated by the generator network to fool the discriminant network in the case of that the discriminant network has strong discriminating power.

The general idea of GANs is to optimize adversarial min-max problem, in which we define a discriminator network $D_{\theta_D}$ to correct the result of generator network $G_{\theta_G}$:

$$\min_{\theta_G} \max_{\theta_D} E_{I_{HR} \sim P_{train}(I_{HR})}[\log D_{\theta_D}(I_{HR})] + E_{I_{LR} \sim P_{G}(I_{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I_{LR})))].$$

This formulation allows the generator to learn to perform images that are highly similar to original high-resolution images and also improve the ability of $D$ to classify origin images and generation. The core parts of generator network $G$, which is introduced by C. Ledig et al. [3] and illustrated in Figure 1. As is shown, generator networks are residual blocks which have an identical layout. There are two special convolutional layers that have small 64 feature maps and $3 \times 3$ kernels followed by Parametric-ReLU [14] and batch-normalization layers [17] as the activation function.

![Generator Network](image1)

**Figure 1.** Generator’s architecture, in which indicates for each conv layer with corresponding stride(s), num of feature maps(n) and kernel size(k).

After the generator network generate HR images from LR images, the discriminator network will discriminate original high-resolution images from generated super-resolution images.

![VGG net’s architecture](image2)

**Figure 2.** VGG net’s architecture, each convolutional layer with detailed parameters.

The discriminant network architecture is based on VGG net, which is shown in Figure 2. We have modified the VGG net architecture, and we inherit the architectural guidelines that were explained by Radford. Then we use Leaky-ReLU, in which the alpha is indicated as 0.2, and drop max-pooling layer in the discriminator network as shown in Figure 3. The advantage of VGG net [19] is that the structure has been simplified, and the whole network uses the same maximum pool size (2 x 2) and convolution kernel size (3 x 3). It is a better chosen for us to assemble several small filter (3 x 3) convolution layers rather than a large filter (5 x 5 or 7 x 7) convolution layer, and the VGG net module can improve discriminating performance by continuously deepening the neural network structure, but it also has certain disadvantages. Fully connected layer contributes most of the params, while VGG has three fully connected layers.
3.3. Loss Function

The pixel-wised MSE for the super-resolution problem is defined as follows:

$$I_{MSE}^{SR} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I_{x,y}^{HR}) - G_{\theta_G}(I_{x,y}^{LR})^2$$  \hspace{1cm} (3)

This formulation is the most commonly used method to optimize super-resolution problem and can always get higher PSNR \[22, 23\]. However, this optimization target will lead to the missing of high texture details.

According to the neural network architecture introduced above, there will be two part of loss functions and two parameters defined as content loss ($l_X^{SR}$) and adversarial loss ($l_{Gen}^{SR}$):

$$l^{SR} = \alpha l_X^{SR} + \beta l_{Gen}^{SR}$$  \hspace{1cm} (4)

The novel loss function makes SRGAN gets better high texture details, rather than just considering MSE. SRGAN proposed using perceptual loss and adversarial loss to make the generated image and the target image more similar in semantics and style. Although the PSNR value is not as high as bicubic or SRResNet, however, the generated image’s visual effect has been improved.

To avoid the loss of image high texture details caused by pixel-wised loss and establish a loss function closer to human visual perception, SRGAN changed the loss function. Based on the Leaky-ReLU active layer of the pre-trained 19-layer VGG model, we use $\Phi_{i,j}$ to represent the feature map of the $j$-th layer convolution after the $i$-th layer max-pooling layer in the VGG network. Then define VGG loss as the Euclidean distance between the reconstructed image and the reference image:

$$l_{VGG}^{i,j} = \frac{1}{W_iH_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (\Phi_{i,j}(I_{x,y}^{HR}) - \Phi_{i,j}(G_{\theta_G}(I_{x,y}^{LR}))^2$$  \hspace{1cm} (5)

where $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the corresponding feature maps in the VGG model.

In addition to the loss of content described above, the loss of the generated network is also an important part of the perceived loss, and the generation loss of the generated network is based on the discriminator $D_{\theta_G}(G_{\theta_G}(I_{LR}))$ of all samples:

$$l_{Gen}^{SR} = \sum_{n=1}^{N} - \log(D_{\theta_G}(G_{\theta_G}(I_{LR})))$$  \hspace{1cm} (6)

where $D_{\theta_G}(G_{\theta_G}(I_{LR}))$ is the probability that the reconstructed image is judged as the original high-resolution images \[9\].

4. Methods

4.1. Data and Training Details

For the experiment part, our experiments are operated with a scaling factor of 4 times between LR images and HR images. Experience shows that a larger patch is better for training a deep neural network, this is because a relatively large receptive field can help to extract more semantic information. At the same time, it will also consume more computing resources and cost more training time. For the training details, we mainly use the same dataset for training and test sets \[23\], which is a 2K resolution dataset for image super-resolution tasks. We have scaled the low-resolution image vector to [0, 1] for the high-resolution image to [-1, 1]. So we calculate the MSE in the interval of [-1, 1], and
the corresponding VGG feature map will also be adjusted to get the corresponding VGG loss. During the retraining process, our learning rate is initially set to 0.0001 and the decay will be set as 0.1 for every 100 epoch. For optimization, we use Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$. We alternately update the generator and discriminator network until the model converges. We have also attempted different layers of networks and evaluate its impact of final PSNR.

4.2. **Comparison with state-of-the-art**

We have compared the performance of SRResNet and SRGAN to bicubic interpolation and visual the detail results are shown in Figure 4. The PSNR of the picture below is 21.59db (bicubic), 23.53db (SRResNet) and 21.15db (SRGAN). From these figures we can see that the texture details of the picture with the best PSNR is not the best. To see this more clearly, we separately select a part of the picture to enlarge the details shown in the Figure 5.

![Figure 4. Super-Resolution images of bicubic, SRResNet, SRGAN and the original picture.](image)

![Figure 5. The details of bicubic, SRResNet and SRGAN compare to the original picture.](image)

The existence of adversarial networks is crucial to the texture details of the restored images. Without the adversarial network component, we can get a higher score, but the texture details are performed very poor. Table 1 is the summation of our experiment results and we also provide detail visual images in Figure 5. As is shown, MSE will indicate the solutions to highest PSNR score. At the same time, the texture details are perceptually rather smooth and less convincing.
Table 1. Average PSNR on test datasets.

| DataSet | SRResNet-MSE | SRResNet-VGG20 | SRGAN-MSE | SRGAN-VGG20 | SRGAN-VGG52 |
|---------|--------------|----------------|-----------|-------------|-------------|
| Set5    | 32.11        | 30.63          | 30.64     | 29.84       | 29.40       |
| Set14   | 28.52        | 27.33          | 27.98     | 26.46       | 26.12       |

5. Conclusion
In this paper, we introduce an end-to-end image super-resolution via generator adversarial that extracts picture information and restores the corresponding super-resolution image for a given low-resolution picture. We compare the effects of different network layers on the image restoration effect, and prove that a certain degree of deepening the network layer can achieve a more robust recovery effect. However, the jpg compression storage algorithm leads to the loss of pixel information. The future work is to optimize the algorithm to have a good recovery effect on jpg images.

6. Acknowledgments
Thanks to Beijing University of Posts and Telecommunications Innovation and Entrepreneurship Fund for its sponsorship.

References
[1] G. Freedman and R. Fattal, “Image and video upscaling from local self-examples,” ACM Transactions on Graphics, vol. 30, no. 2, p. 12, 2011.
[2] C.-Y. Yang and M.-H. Yang, “Fast direct super-resolution by simple functions,” in Proceedings of IEEE International Conference on Computer Vision, 2013, pp. 561–568.
[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), Honolulu, HI, 2017, pp. 105-114.
[4] R. G. Keys, “Cubic convolution interpolation for digital image processing,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 29, no. 6, pp. 1153–1160, 1981.
[5] C. E. Duchon, “Lanczos filtering in one and two dimensions,” Journal of Applied. Meteorology, vol. 18, no. 8, pp. 1016–1022, 1979.
[6] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient. profile prior,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[7] K. Zhang, X. Gao, D. Tao, and X. Li, “Single image super-resolution with non-local means and steering kernel regression,” IEEE Transactions on Image Processing, vol. 21, no. 11, pp. 4544–4556, 2012.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional. network for image super-resolution,” in Proceedings of European Conference on Computer Vision, 2014, pp. 184–199.
[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680, 2014.
[10] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction. beyond mean square error. In International Conference on Learning Representations (ICLR), 2016.
[11] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image. models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems (NIPS), pp. 1486–1494, 2015.
[12] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (ECCV), pp. 694–711. Springer, 2016.
[13] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep. convolutional sufficient statistics. In International Conference on Learning Representations (ICLR), 2016.
[14] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034, 2015.
[15] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In European Conference on Computer Vision (ECCV), pages 694–711. Springer, 2016.
[16] S. Gross and M. Wilber. Training and investigating residual nets, online at. http://torch.ch/blog/2016/02/04/resnets.html. 2016.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of The 32nd International Conference on Machine Learning (ICML), pages 448–456, 2015.

[18] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1874–1883, 2016.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.

[20] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.

[22] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transactions on Image Processing, vol. 19.