Research on image super-resolution based on attention mechanism and multi-scale

Xiaokang Ren, Xingzhen Li

College of Computer Science & Engineering, Northwest Normal University, Lanzhou, Gansu, 730070, China
2018221760@nwnu.edu.cn

Abstract. In order to solve the problem of the single feature scale of the generated image in the SISR field and the lack of texture information, a parallel generation confrontation network structure based on the attention mechanism and multi-scale is proposed on the basis of SRGAN, which adopts a dual generator and discriminator combined with attention module model. Train the network to learn multi-scale features, and integrate high-frequency information of different scales in the residual network. The experimental results on Set5, Set14, and BSD100 benchmark data sets prove that the algorithm has a good effect in restoring image detail information.

1. Introduction
With the widespread application of deep learning in the field of computer vision, image SR algorithms based on deep learning have become a hot spot. Dong et al. proposed an end-to-end convolutional neural network (Super Resolution Convolutional Neural Network, SRCNN) [1] to learn the mapping relationship between LR and HR images. SRCNN is the first work to apply convolutional network to image SR, which shows the superiority of Convolutional Neural Network (CNN) in feature extraction and representation. After that, Goodfellow et al. [2] first proposed generative adversarial networks (GAN). The success of GAN enabled the idea of adversarial training to be quickly applied to the generative model. Ledig [3] and others used the generative confrontation network to solve the super-resolution problem, and proposed the SRGAN model. The model contains two parts: a generative model and a discriminative model. The target HR image is reconstructed through confrontation training. SRGAN makes HR images clearer and more natural by using perceptual loss and adversarial loss.

2. Related content

2.1. Attention mechanism
In recent years, the attention mechanism has been widely used in deep neural networks. Hu et al. proposed SENet [18] to learn the correlation between channels, and adaptively re-correct the corresponding strength of features between channels through the global loss function of the network. This network structure has achieved significant performance improvements in image classification.

As shown in Figure 1, the channel attention mechanism has two advantages in single-image super-resolution tasks. On the one hand, each filter in the basic block convolutional layer uses a local receptive field, so the output after convolution cannot use context information outside the local area, and the global pooling layer of channel attention contains the global channel spatial information. On
the other hand, different channels in the feature map play different roles in extracting low-frequency or high-frequency components. The introduction of the channel attention layer can adaptively adjust the weight of the channel to effectively improve the reconstruction index. In super-resolution tasks where it is difficult to reconstruct high-frequency details, it is often advantageous to assign higher weights to the channels for extracting high-frequency components, and to assign lower weights to the redundant low-frequency component extraction channels.

Figure 1. Attention mechanism model

2.2. Image super resolution based on gan

The core idea of the SR model based on the generative confrontation network is to make the HR image generated by the generator approximate the original image as much as possible through confrontation training. The working process of the generator G is to input the low-resolution image LR image and output the reconstructed HR image through the end-to-end learning process, and then input the generated HR image into the discriminant model D to judge the authenticity.

3. Method of this article

3.1. The main work

Inspired by the above work, a generative confrontation network model based on attention mechanism and multi-scale feature extraction is proposed. Improved the generative model structure and the discriminant model structure. The new generator is composed of two parallel residual network blocks with attention modules added to different layers, and feature maps of different scales are extracted in different layers of the sub-network, through the fusion network. The HR images of different scale spaces generated by the two sub-networks are weighted and summed, so that the residual information is fused, so as to learn more detailed image information. The new discriminator is a discriminator that adds an attention module on the original basis, strengthens the identification of high-frequency information, and guides the generator to generate high-quality images with rich details.

3.2. Network structure

The model structure of the generator is shown in Figure 2. The model contains two parallel residual network blocks, through this structure to learn the characteristics from the bottom to the high-level, to obtain the edge texture information of different scale spaces. Finally, through the fusion network, the generated HR images of different scale spaces are weighted, so that the residual information is fused to obtain more detailed image information. Both sub-networks are composed of 1 convolution module + 16 residual modules + 1 convolution module + 2 sub-pixel convolution modules + 1 convolution module. Sub-network 1 uses a 3 × 3 convolution core to learn image features, add a channel attention module every 3 residual blocks and extract this layer of image features. Sub-network 2 uses a 5 × 5 convolution kernel, adds a channel attention module every 2 residual blocks and extracts the image features of this layer. Using a multi-scale learning model can give the generative model better learning capabilities.
The discriminant network is equivalent to a feature extraction module, and its network structure is shown in Figure 3. Except for the first convolutional layer, no batch standardization operations are performed. The middle 7 convolutional layers are followed by batch standardization operations. The last convolutional layer is followed by a fully connected layer. After the fully connected layer, the Sigmoid activation function outputs a 0 to 1 probability value between 1 is used to judge the true and false of the image generated by the generator and the real image. For the image generated by the generator, we output a probability value of 0, and for a real high-resolution image, we output a probability value of 1.

3.3. Improved objective optimization function

Existing super-resolution algorithms often choose to minimize the mean squared error (MSE) in the optimization objective. The formula of MSE is expressed as

$$l_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{x,y}^{HR} - G_{\theta G}(I_{x,y}^{LR}))^2$$

In formula (1): W and H represent the image size; $I_{x,y}^{HR}$ represents the real high-resolution image; $G_{\theta G}(I_{x,y}^{LR})$ represents the generated image. This is the most widely used image SR optimization target. Although this method produces a very high peak signal to noise ratio (PSNR), the result often lacks high-frequency information. To avoid the above problems, a new loss function is defined to evaluate solutions based on perceptually related features. Express the perceptual loss function as

$$l_{PD}^{SR} = a l_{C}^{SR} + b l_{S}^{SR} + c l_{F}^{SR}$$

In formula (2): $I_{x}^{SR}$, $I_{y}^{SR}$, and $I_{z}^{SR}$ represent content loss, style loss, and confrontation loss respectively; a, b, and c are the corresponding hyperparameters.

The HR image generated by the generator is added to the perceptual loss through the process of VGG network feature extraction, and then the generator and the discriminator are alternately optimized, and the gradient feedback given by the discriminator prompts the generator to capture the high-frequency information lost on the LR image. Improve the visual quality of the reconstructed image. The adversarial loss $l_{G}^{HR}$ in this process is defined as

$$l_{G}^{SR} = \sum_{n=1}^{N} - logD_{\theta D}(G_{\theta D}(I_{LR}))$$

In formula (3): $G_{\theta D}(I_{LR})$ is defined as the probability that the reconstructed image $D_{\theta D}(G_{\theta D}(I_{LR}))$ is a natural HR image. For better gradient performance, we use $- logD_{\theta D}(G_{\theta D}(I_{LR}))$ to minimize the generation loss function and better gradient update instead of $log \left(1 - D_{\theta D}(G_{\theta D}(I_{LR}))\right)$. 
4. experiments

4.1. Experimental environment and data set
This article is based on the Pytorch deep learning framework for algorithm building model, the corresponding version torch1.4.0, Python3.6. The operating system is Windows 10, and the GTX 1080TI graphics card is used for acceleration. This article uses the COCO2014 data set for training, and uses the Set5, Set14 and BSD100 data sets for testing.

4.2. Training details
Load an image, crop a 96x96 sub-block from any position, and use the sub-block as the original high-resolution image hr_img; perform bilinear down-sampling (4 times) on hr_img to obtain a 24x24 sub-block. As the initial low score image lr_img; preprocess lr_img according to the ImageNet data set, and convert hr_img to [-1,1]; return lr_img and hr_img as a pair of training pairs in the actual training, using the already trained The residual network structure is used as the initialization of the generator to prevent the model from appearing locally optimal results. Since each operation of the convolution process will reduce the size of the feature map, the edge pixel information is preserved by zero-filling the image after the convolution, while ensuring that the size of the feature map before and after the layer jump connection structure is the same, so that it can be accurate Calculate the center pixel value. The model is a parallel network structure. The initial learning rate of the two sub-networks is 0.0001. After 50 iterations of update training, the learning rate starts to decrease. We use the ADAM optimizer to train the model. The slope of the activation function LeakyReLU is set to 0.2.

4.3. Experimental comparison
We compared the network structure proposed in this article with the four image super-resolution methods of Bicubic, SRCNN, SRResNet, and SRGAN. The implementation of the comparison method comes from the author's open source code. The following table shows the average PSNR and SSIM averages of different methods on the Set5, Set14, and BSD100 test sets. When the amplification factor is 4, our method has a significant improvement in the average value of PSNR and SSIM on each test data set. This shows that our network performance has been significantly improved compared to these algorithms.

| Table1. set5 test set evaluation index average results. |
|---------------------------------|---------|--------|
| method                      | PSNR    | SSIM    |
| Bicubic                     | 27.45   | 0.765   |
| SRCNN                       | 30.13   | 0.863   |
| SRResNet                    | 31.55   | 0.902   |
| SRGAN                       | 29.02   | 0.839   |
| Ours                        | 30.43   | 0.773   |

| Table2. set14 test set evaluation index average results. |
|---------------------------------|---------|--------|
| method                      | PSNR    | SSIM    |
| Bicubic                     | 25.70   | 0.715   |
| SRCNN                       | 27.18   | 0.785   |
| SRResNet                    | 28.49   | 0.591   |
| SRGAN                       | 28.65   | 0.755   |
| Ours                        | 28.92   | 0.773   |

| Table3. BSD100 test set evaluation index average results. |
|---------------------------------|---------|--------|
| method                      | PSNR    | SSIM    |
| Bicubic                     | 25.47   | 0.687   |
| SRCNN                       | 26.82   | 0.725   |
| SRResNet                    | 27.52   | 0.725   |
| SRGAN                       | 24.78   | 0.755   |
| Ours                        | 26.43   | 0.783   |
In addition, we selected 2 pictures from the test set for testing. Enlarge a specific area of the image to better observe the reconstruction effect of texture details. Figure 4 shows the reconstruction effect of this algorithm and several other traditional algorithms. From the perspective of subjective visual effects, compared with other methods, the image reconstructed by this method restores more high-frequency details and produces clearer edge effects. Therefore, combining subjective effects and objective indicators, the algorithm in this paper can obtain better reconstruction results than mainstream super-resolution reconstruction algorithms.

![Figure 4. The reconstruction effect of various algorithms when the magnification factor is 4 times.](image)

5. Conclusion
This paper proposes an adversarial neural network based on attention mechanism and multi-scale feature fusion. Through the attention mechanism module, the inherent attribute features of the image are fully explored and the rich features are adaptively learned, which improves the network recognition and learning ability. Use different scale convolution kernels to extract image features and fuse multi-scale features. The experimental results show that the image quality reconstructed by this method is improved in both visual and quantitative indicators. The next step will continue to study and improve the network structure to improve the image reconstruction effect.

References
[1] Chao Dong, C. Loy, Kaiming He and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 2, pp. 295-307, 1 Feb. 2016.
[2] Goodfellow I J, Pougetabadie J, Mirza M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014 (3): 2672-2680.
[3] Zhang Y, Li K, Li K, et al. Image Super-Resolution Using Very Deep Residual Channel Attention Networks[J]. 2018.
[4] Jie Hu, Li Shen and G. Sun. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7132-7141.