A Novel Image Super-Resolution Method Based on Attention Mechanism

Da Li¹, Yan Wang¹, Dong Liu¹⁺ and Ruifang Li¹

¹School of information engineering, Wuhan University of Technology, Wuhan, 430070, China

* liudong@whut.edu.cn

Abstract. Image super-resolution processing technology is used to reconstruct high-resolution images from low-resolution images. With the development of deep convolution neural network, image super-resolution processing methods based on deep learning have become the main technology. A novel image super-resolution processing method which is based on attention mechanism is proposed in this paper. The network framework consists of two modules: one is an improved position feature extraction network based on residual network and dense network, and another is a channel feature extraction based on channel attention mechanism. Meanwhile, this paper extracts features directly from low-resolution images, and then amplifies them to reduce the computational complexity. From the experimental results, we can find that the proposed method can enlarge the image better and increase the PSNR on Set5, Set14 and B100.

1. Introduction

Image super-resolution (SR), especially single image super-resolution (SISR), has attracted more and more attention in recent decades. SISR aims to reconstruct a high-resolution image (HR) from a low-resolution image (LR). It is popular in the many fields. Since the method of single frame high-resolution image restoration from low-resolution image sequence was proposed by Tsai and others[1] in 1984, super-resolution restoration technology has been studied widely. However, image SR is an ill question because there may be many solutions for a LR image. In order to solve this problem, many image SR algorithms have been proposed, such as methods based on interpolation, reconstruction and learning.

Dong et al[2]. first used convolutional neural network in image super-resolution processing in 2014, different convolutional network frameworks and their improved structures have been applied to image super-resolution. Inspired by this, this paper adds the low layer input and the high layer output by the skip connection, and takes it as the input of the next layer, so as to ensure that the effective information of the low level features can be more fully input into the deeper layer. And meanwhile the skip connection is used to reduce the problem of gradient-vanishing or gradient-explosion. A feature extraction block is proposed in this paper. The input of the block and the output of each layer in the block are added as the input of the next layer in each block, and finally more Blocks are cascaded to form a deeper network.

Attention mechanism has not been proposed in recent years, but it has become popular in recent years. Especially after Google's Attention Is All You Need[3] published in 2017, which provides a new way of thinking for people. In image processing, attention mechanism is mainly divided into spatial
attention and channel attention. In this paper, on the basis of the previous work, by convoluting the image, we first compress and expand the channel attention feature to enhance the channel attention feature, and finally multiply the original input matrix to obtain the information of attention mechanism.

2. Related work

2.1 Feature extraction network
The key technology of super-resolution based on deep learning lies in the establishment of feature extraction network. Since LeNet-5 network was proposed, various convolutional neural networks have been updated and developed, such as AlexNet[4], VGG, GoogLeNet and so on. He and Ren proposed a residual learning block and a deep residual network ResNet[5] were proposed. The ResNet’s main idea is to add a highway network to the network, so that the original input information can be directly transmitted to a deeper network. This idea of sliding connection protects the integrity of input information. In 2017, Liu Zhuang and others proposed DenseNet[6], a dense connected convolution network, which takes the output of each convolution layer as the input of next layers, so that the back convolution layers can obtain the characteristics of all previous convolutions.

2.2 Image super-resolution
Since SRCNN for image super-resolution processing was proposed in 2014, researchers have applied different neural network and their improved structure to image super-resolution, and achieved different research results. Such as FSRCNN, VDSR and so on. In 2017, Lim [7] optimized the network by removing unnecessary modules in the traditional residual network and proposed the enhanced deep super-resolution network EDSR. The generative adversarial network is first used by Ledig et al[8] in the image super-resolution, and proposed SRGAN. GAN is used to generate details in the image, and perception loss and confrontation loss is used to improve the realism of the restored image.

2.3 Attention mechanism
Attention is summed up by human's habit of observing the environment. Attention mechanism has become popular in recent years, especially after Google's Attention Is All You Need in 2017, which provides a new way of thinking for people. In image SR, attention mechanism is mainly divided into spatial attention and channel attention. S. woo and J. Park proposed convolution block attention module. Zhang and Wang involved in significant target detection by gradually integrating multi-level context information. The extruding and exciting (SE) blocks is proposed by Hu and others[9], which is used to model relationships in channels and achieve significant image classification performance improvements.

3. Propose network structure
In this section, the network structure proposed in this paper for image SR will be introduced. This network is divided into two parts: feature extraction network and channel attention extraction network.

3.1 Feature extraction block
The feature extraction block is used to obtain image features from low-resolution image. \(L_{i-1}\) represents the input of the \(i\)-th convolution layer, and the result of the \(i\)-th layer is expressed as follows:

\[
L_i = \sigma(W_i \ast L_{i-1} + B_i)
\]

(1)

where \(\ast\) represents convolution operation, \(\sigma\) represents linear activation function ReLU, \(W_i\) and \(B_i\) represent weight matrix and bias matrix in the \(i\)-th layer respectively.

The input of each convolution layer includes not only the output of the previous one but also the original input of the feature extraction block. Therefore, the input of each convolution layer is
expressed as:

\[ L_{i-1} = s \times L_{oi-1} + I \] (2)

where \( s \) is the scale factor, and its value range is \((0, 1]\) which is used to control the input of each layer to adopt the scale of the output of the previous layer.

\[\text{Figure 1. Feature extraction block}\]

The structure of feature extraction block is shown in Figure 1. The original input of each black is used as the input of each convolution layer in this block, so the output \( F_i \) of each block can be expressed as:

\[ F_i = \varphi_i(H_i, W_i, s_i) + H_i \] (3)

Where \( s_i \) represents the scale factor in the block, \( W_i \) represents the weight set in the block, \( H_i \) represents the original input of the block, and \( \varphi_i \) represents the function of the block. Each block can contain multiple rollups. The function of this block can be expressed as follows by two rollups:

\[ \varphi_i(H_i, W_i, s_i) = \sigma_2(W_i^2 \times (s_i \sigma_1(W_i^1 \times H_i) + H_i)) \] (4)

where \( W_i^1 \) and \( W_i^2 \) represent the weights of these two convolutions respectively, while the bias matrix is omitted. \( \sigma_1 \) and \( \sigma_2 \) represent the linear activation function ReLU and regularization respectively.

We draw lessons from the dense network architecture, and take the original input of the block as the input of the next layer after linear calculation with the output of each layer.

### 3.2 Channel attention mechanism

It is very important for super-resolution reconstruction to recover high-frequency information. In order to obtain more global features and rich context dependence, a channel attention mechanism is proposed in this paper, which aims to analyze image features from the channel.

The first thing to consider is the number of channels. An RGB image with three channel is not enough for channel attention design. So we first convolute the original image to increase the number of channel. And then the features on each channel are pooled globally:

\[ z_c = H_{gp}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \] (5)

where \( x_c(i, j) \) represents the position \((i, j)\) of the point on the channel \( c \), and \( H_{gp} \) represents the global average pooling function. Through global average pooling, the feature information of each channel is focused on one point to achieve the purpose of attention. Then, we compress and enlarge the channel attention feature by convolution, as shown in the Figure 2. Here, we refer to the idea of U-net.
3.3 Network structure
In the first two sections, we propose the feature extraction block and channel attention mechanism. We combining the above two modules to proposed network structure in this section.

As shown in the Figure 3, we cascade feature extraction blocks, with the output of each block as the input of the next block. One channel attention mechanism is used in several cascaded blocks. The number of blocks can be determined flexibly, which is regarded as a unit. The channel attention mechanism is parallel to the cascaded block with same input features. The outputs are integrated by matrix addition as the input of the next unit. Finally they are used to reconstruct high-resolution image.

4. Experiments and results

4.1 Data set and Setting
This paper uses DIV2K set as training dataset to train. Then we test on Set5, Set14 and B100 dataset. In this paper, we set the network structure as follows: the number of convolutions in feature extraction block is set to 3, the scaling times in channel attention mechanism is 2, the number of feature extraction blocks is 5, and the number of units is 6. Meanwhile, we train directly on LR rather than on Bicubic-image, because the complexity of training will be greatly increased by enlarging the image first.

4.2 Analysis of experimental results
Two image quality evaluation standards are PSNR (Peak signal-to-noise ratio) and SSIM (Structural
Similarity). PSNR evaluates the image by comparing the gray values of the corresponding pixels of the two images. SSIM is used to evaluate the similarity of the two images from the brightness and structure.

In this paper, we test on set5, set14 and B100 datasets, and calculate the PSNR and SSIM values of our network under the magnification of ×2, ×3, ×4. Finally we compare our results with other networks.

### Table 1. Comparison between our method and other methods

| Method   | Scale | Set5 PSNR(dB) | SSIM | Set14 PSNR(dB) | SSIM | B100 PSNR(dB) | SSIM |
|----------|-------|---------------|------|----------------|------|---------------|------|
| Bicubic  |       | 33.66         | 0.930| 30.24          | 0.870| 29.56         | 0.843|
| SRCNN    |       | 36.66         | 0.954| 32.45          | 0.907| 31.36         | 0.892|
| FSRCNN   |       | 37.05         | 0.956| 32.66          | 0.909| 31.53         | 0.892|
| VDSR     | ×2    | 37.53         | 0.959| 33.05          | 0.913| 31.90         | 0.896|
| DRCN     |       | 37.63         | 0.959| 32.98          | 0.913| 31.85         | 0.894|
| LapSRN   |       | 37.52         | 0.959| 33.08          | 0.913| 31.08         | 0.895|
| Ours     |       | 37.73         | 0.962| 33.16          | 0.914| 31.89         | 0.896|
| Bicubic  |       | 30.39         | 0.868| 27.55          | 0.774| 27.21         | 0.738|
| SRCNN    |       | 32.75         | 0.909| 29.30          | 0.822| 28.41         | 0.786|
| FSRCNN   |       | 33.18         | 0.914| 29.37          | 0.824| 28.53         | 0.791|
| VDSR     | ×3    | 33.67         | 0.921| 29.78          | 0.832| 28.83         | 0.799|
| DRCN     |       | 33.82         | 0.923| 29.76          | 0.833| 28.80         | 0.797|
| LapSRN   |       | 33.82         | 0.923| 29.87          | 0.832| 28.82         | 0.798|
| Ours     |       | 33.89         | 0.925| 29.93          | 0.833| 28.84         | 0.799|

5. Conclusion
The feature extraction block and attention mechanism are proposed in this paper. The feature extraction block uses the spatial features of the image since the sliding connection. The attention mechanism uses the structure of U-shaped network to extract the channel features of the image after the average pooling. The two modules are connected by mathematical calculation. Then we train our network on div2k data set, and test it on data set set5, set14 and B100. Compared with the results of other networks, as shown in Table 1 and Figure 4, the proposed method improves both PSNR and SSIM. In the future, we will improve this paper’s structure, train and test the remote sensing image to achieve better results.
| Method     | 
|------------|
| Bicubic    |
| SRCNN      |
| VDSR       |
| DRCN       |
| Ours       |
| HR         |

**Figure 4.** Comparison between our method and other methods on ×2 and x3 SR

References

[1] Tsai R Y, Huang T S. Multipleframe image restoration and registration. Advances in Computer Vision and Image Processing, 1984:317-339.

[2] Chao Dong, Chen Change Loy, Kaiming He, et al. Learning a Deep Convolutional Network for Image Super-Resolution. European Conference on Computer Vision, 2014: 184-199.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. Neural Information Processing Systems 30, 2017.

[4] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25, 2012.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2016

[6] Huang H, Liu Z, Maaten Lad. Densely Connected Convolutional Networks. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[7] Lim B, Son S, Kim H. Enhanced Deep Residual Networks for Single Image Super-Resolution. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[8] Christian Ledig, Lucas Theis, Ferenc Husz’ar. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[9] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. IEEE Conference on Computer Vision and Pattern Recognition, 2018.