International Journal of Computer Applications Technology and Research
Volume 12–Issue 03, 01-05, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1203.1001

Image Super-Resolution Reconstruction Based on Residual Convolution and Double Attention Mechanism

Qinglin Huang
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Congcong He
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Jieyuan Luo
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, China

Abstract: Currently, single-image super-resolution reconstruction based on deep learning has achieved good results. To address the problems that most networks will have long training time, weak image learning ability and not fully utilizing the high frequency information of images, an image super-resolution reconstruction method based on residual convolution and double attention mechanism is proposed. The model performs deep feature extraction of the image by cascading deep convolutional networks, introduces local residual blocks to solve the model degradation brought by too many network parameters, and embeds the dual attention mechanism module in the residual blocks for adaptive calibration to adjust the feature map weights of each channel and the spatial correlation between features, so as to obtain deep texture detail information and reconstruct the feature image by sub-pixel convolutional layers to up sampling to reconstruct the high-resolution image. In the test sets of Set5 and Set14, peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used as evaluation indexes, while comparing SRCNN, FSRCNN and VDSR methods all reconstructed images with better results. The experimental results show that the method can effectively improve the utilization of high-frequency feature information and can increase the reconstruction capability of the images to a certain extent.

Keywords: Deep learning; super-resolution reconstruction; residual convolution; dual attention mechanism

1. INTRODUCTION

Image super-resolution reconstruction (SR) is a challenging and popular research topic in the field of computer vision and image processing, where image resolution is a set of performance parameters used to evaluate the richness of detail information contained in an image. However, in practice, most imaging devices are disturbed by various factors such as hardware and environment, which makes the image resolution not meet the needs of practical applications. Therefore, in order to improve the image resolution without changing the imaging device, researchers have started to try to reconstruct low-resolution (LR) images into high-resolution (HR) images by using image processing and machine learning algorithms. Due to the property that deep learning can adaptively learn the nonlinear mapping relationship between LR images and HR images, the SR algorithm for images based on deep learning is significantly better than the traditional methods, so it has also become the mainstream research method for image super-resolution reconstruction methods [1]. Meanwhile, SR techniques have been widely used and achieved significant results in practical scenarios such as analysis and recognition of medical images, face super-resolution, video surveillance and security, and remote sensing images. Image super-resolution is mainly divided into two categories, single image super-resolution (SISR) and multiple image super-resolution (MISR), which are discussed in this paper [2-3].

With the development of deep learning techniques, convolutional neural network-based approaches have achieved great success and made an important impact in the field of computer vision. In 2014, the pioneering introduction of convolutional neural networks into the field of image super-resolution by Dong [4] et al. proposed the super-resolution convolutional neural network (SRCNN) to learn the mapping relationship between high- and low-resolution images in an end-to-end manner, which greatly simplifies the workflow of super-resolution algorithms. The algorithm combines traditional sparse coding with deep learning as the basis, uses dual cubic interpolation to put a low-resolution image of the target size as the input, and uses a convolutional neural network containing three convolutional layers to fit the nonlinear mapping between high- and low-resolution images, completing the extraction and feature representation, nonlinear mapping, and image reconstruction process, realizing end-to-end image reconstruction, and its reconstruction The effect is significantly better than the traditional super-resolution algorithm, which opens the way for deep learning research in the field of image super-resolution. Later, in 2016, Dong et al [5] proposed a modified fast super-resolution convolutional neural network (FSRCNN) for SRCNN, using the original low-resolution image without processing as the input for training, and up-sampling and reconstructing the high-resolution image by the deconvolution layer at the end of the network. In the same year, Shi et al [6] proposed an efficient sub-pixel convolutional neural network (ESPCN), which achieves the up-sampling operation of LR images by pixel rearrangement, which greatly reduces the computational effort and improves the reconstruction efficiency compared to the inverse convolutional layers. Kim [7] et al also proposed a very deep super-resolution network (VDSR) in the same year by stacking 20 convolutional layers for image features for deep extraction and introduced a residual model to speed up the convergence of the network and improve the reconstruction results.

Based on the consideration of the training time of most networks, the performance of image reconstruction effect and the utilization of high-frequency image information perspectives, an image super-resolution reconstruction method based on residual convolution and double attention mechanism is proposed in the paper. The method takes the low-resolution image input model after double triple interpolation, and firstly uses the convolution layer with smaller convolution kernel to extract the shallow features of the input image; then extracts the deep features of the image by cascading deep convolutional network, introduces the local residual block to solve the problem of model degradation and gradient caused by too many parameters of the network, and adds the double attention mechanism module in the residual block to give the network feature weights Finally, a sub-pixel convolutional layer is used as the up sampling method at the end of the network to reconstruct a high-resolution image of the target size. The experimental results show that this method can effectively improve the utilization of high-frequency information, recover image details, and improve the image reconstruction effect.

2. RELATED JOBS

In order to enable the network to pay more attention to the ground high frequency information and ignore the irrelevant information in the network data, the attention mechanism has been reapplied in the field of computer vision. The attention mechanism is an information processing mechanism that originated from the study of human vision. In the human vision system, the received visual information is not processed all at once, but some of the key information is selected for processing, and resources are allocated rationally to solve the information overload problem, thus improving efficiency [8]. Since Hu et al. proposed the channel attention (CA) mechanism in SENet [9] in 2018, it has been widely used in deep learning, and although it increases the number of certain parameters, the performance has been well improved.

The Convolutional Block Attention Module [10] (CBAM) is a dual attention mechanism that combines the attention mechanisms of space and channel, which can achieve better results than the CA mechanism that focuses only on the channel. CBAM starts from two scopes, channel and space, and introduces both spatial CBAM starts from two scopes of action, channel and space, and introduces two analysis dimensions, spatial attention and channel attention, to achieve a sequential attention structure from channel to space. The spatial attention allows the neural network to focus more on the pixel regions in the image that play a decisive role in classification and ignore the irrelevant regions, while the channel attention is used to deal with the assignment relationship of feature map channels, and the simultaneous attention allocation to both dimensions enhances the effect of the attention mechanism on model performance. the structure of the CBAM module is shown in Figure 1.

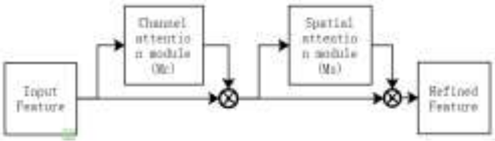

Figure 1. The structure of the CBAM module

Given an intermediate feature map F∈R(C × H × W) as input, CBAM sequentially derives a 1D channel attention map Mc∈ R(C × 1 × 1) and a 2D spatial attention map Ms ∈ R(1 × H × W), and the whole attention process can be summarized in Equation as：

$$F' = M_c(F) \otimes F, F'' = M_s(F') \otimes F'$$

where ⊗ denotes element-by-element multiplication, during which the attention values are propagated accordingly and F^" is the final refined output.

In the channel attention module, each channel of the feature map is considered as a feature detector, and the channel attention focuses on "what" is meaningful in a given input image. The global maximum pooling and global average pooling are applied to the input feature maps, and the feature maps are compressed based on two dimensions to obtain two different dimensional feature descriptions, F_{avg}^c and F_{max}^c; the pooled feature maps share a multilayer perceptron network (MLP) and a hidden layer, and the hidden activation size is set to R(C/r×1×1) to reduce the parameter overhead, where r is the scaling rate; the two feature maps are combined using element-by-element summation to merge the output feature vectors, and the weights of each channel of the feature map are normalized by the sigmoid activation function, and the normalized weights are multiplied with the input feature map. the flowchart of the channel attention mechanism module in CBAM is shown in Figure 2.

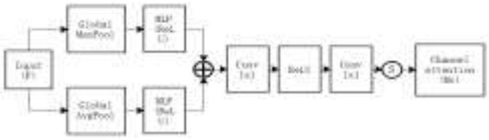

Figure 2. The channel attention mechanism module

The channel attention of channel attention is calculated by equation as follows：

$$M_c(F) = \sigma(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{max}^c\right)\right))$$

where σ is the sigmoid activation function, W_0 ∈ R(C/r×C), and W_1 ∈ R(C×C/r), noting that the MLP weights W_0 and W_1 are shared for both inputs.

The spatial attention mechanism mainly processes the spatial domain of the output feature map of the channel attention mechanism, and uses the spatial attention module after the channel attention module to extract the spatial features between channels and generate the spatial attention map. When extracting the feature information, the spatial attention module shifts the focus of the network model from "what features are meaningful" to "where features are meaningful", thus further extracting the spatial feature information from the output features of the channel attention module. The spatial attention mechanism module in CBAM is shown in Figure 3.

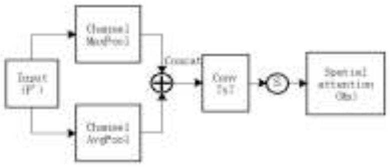

Figure 3. The spatial attention mechanism module

The feature map F^' output from the channel attention module is used as the input feature map of this module. First, the feature maps are subjected to maximum pooling and average pooling based on the channel dimension to obtain two H×W×1 2D maps: the average pooling feature F_{avg}^s and the maximum pooling feature F_{max}^s across channels, respectively; then the two output feature maps are stacked in the channel dimension are connected and convolved by a standard convolution layer, and the spatial features are filtered using a 7×7 convolution kernel to produce a 2D spatial attention map; finally, the final attention map is obtained by normalizing the weights with a sigmoid activation function. The spatial attention is calculated by equation as follows.

$$M_s(F) = \sigma(f^{7\times7}([F^s_{avg}, F^s_{max}]))$$

where σ is the sigmod activation function and $f^{7\times7}()$ is the 7×7 convolution operation.

3. METHODOLOGY OF THIS PAPER

3.1 Network Structure

Although increasing the network depth and number of layers can improve the image reconstruction, it also makes the network difficult to train and hard to converge. For image super-resolution reconstruction tasks, high-frequency features are more valuable for the reconstruction of high-resolution images. By introducing the attention mechanism, the model can focus on the extraction of high-frequency features among the many input information, which can improve the accuracy and efficiency of the model processing. Considering the image reconstruction performance, the training time of the network and the utilization rate of high-frequency feature information of the image, an image super-resolution reconstruction method based on residual convolution and double attention mechanism is proposed in the paper. The local residual block is introduced to solve the information overload and model degradation caused by too many network parameters, and then the dual attention mechanism module is embedded in the residual block to adaptively calibrate the network features and adjust the feature map weights on each channel and spatial domain to improve the model reconstruction effect, and its network structure model is shown in Figure 4.

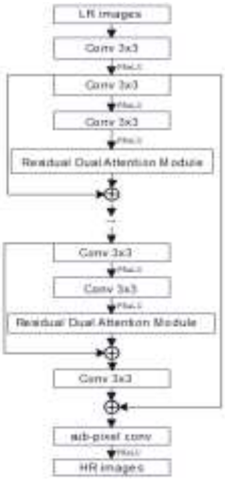

Figure 4. Network Structure Diagram

The method incorporates a residual double attention module, which takes the low-resolution image after double triple interpolation as the input of the model, first extracts the shallow feature information of the image using a convolutional layer, maps the image features nonlinearly into a high-dimensional vector, then learns the depth feature information of the image by cascading residual blocks, nests a residual double attention module in each local residual block to adaptively calibrate the generated image features to suppress redundant information, and finally up samples the features through a sub-pixel convolutional layer [11] to reconstruct the high-resolution image.

3.2 Residual Dual Attention Module

The residual double attention module adds a double attention mechanism module on the basis of the residual module to give different weights corresponding to the importance of different feature outputs and suppress the relatively redundant features, so as to extract the key features that are beneficial to image reconstruction.

In order to enhance the nonlinearity of the network, a convolutional layer with smaller convolutional kernel and PReLU activation function are added to the module to deepen the depth of the module to extract the depth features of the image, and its specific module structure framework is shown in Figure 5.

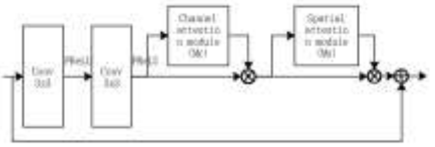

Figure 5. Residual double attention module structure diagram

Let the input of the module be x_0, then the output y of the module can be expressed as equation:

$$y = x_0 + M_s(M_c(x_l) \otimes x_l) \otimes (M_c(x_l) \otimes x_l)$$

where $M_c(\cdot)$ denotes the channel attention mechanism, $M_s(\cdot)$denotes the spatial attention mechanism, and $x_l$denotes the input of the residual dual attention module, obtained by the convolution operation.

3.3 Loss Function

The task of image super-resolution reconstruction is to obtain the reconstructed image through a series of learning by convolutional neural networks, so that the difference between the reconstructed image and the original high-resolution image is as small as possible.

The method in this section uses mean square error (MSE) as the loss function to optimize the parameter training model, which is shown in equation.

$$MSE = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W}(x(i,j) - y(i,j))^2$$

Where, H and W denote the height and width of the image, respectively, and x(i,j) and y(i,j) denote the pixel points corresponding to the reconstructed image and the original image, respectively.

4. EXPERIMENTAL ANALYSIS

4.1 Data Set and Experimental Setup

The experiments use the publicly available image super-resolution dataset DIV2K, which contains 800 training images, 100 validation images, and 100 test images. The method in the paper uses 800 training images of this dataset as the training dataset. Set5 and Set14 benchmark datasets are used as the test datasets.

The LR images in the training set after double triple interpolation are randomly cropped into 96×96 size image blocks, and data enhancement is achieved by random rotation of 90°, 180°, 270° and horizontal flip. In the training phase, the network parameters were optimized using the Adam optimizer with parameters $\beta 1$ and $\beta 2$ set to 0.9 and 0.999, respectively, and ϵ set to 10^(-8). The learning rate was initialized to 0.0001 and the learning rate was halved every 200 cycles. Each batch input was set to 64.

4.2 Evaluation Criteria

The evaluation criteria of single-frame image super-resolution methods are usually divided into subjective and objective evaluations. The subjective evaluation is performed by the human eye visually comparing the original image with the generated image. To verify the quality of the model, objective evaluation criteria such as peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) are usually used to evaluate the reconstruction quality of the generated image for different models.

The peak signal-to-noise ratio measures the image reconstruction quality by calculating the error between the corresponding pixel points, which is calculated as follows:

$$PSNR = 10lg\frac{MAX^2}{MSE}$$

Where MAX indicates the maximum value of the image signal, i.e., the peak value, expressed as (2^n-1), and n is the number of image bits per pixel, generally 8. MSE is the mean square error between the original image and the generated image. the unit of PSNR is dB, and the larger the value, the smaller the image distortion.

Structure similarity measures the similarity of the image from three perspectives: brightness, contrast and structure. Assuming that x and y denote the original high-resolution image and the recovered high-resolution image, respectively, the calculation is shown below:

$$SSIM(x,y) = l(x,y) \cdot c(x,y) \cdot s(x,y)$$

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

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

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

where l(x,y) denotes brightness comparison, c(x,y) denotes contrast comparison, and s(x,y) denotes structure comparison; μ_x and μ_y denote the pixel mean of the two images, σ_x and σ_y denote the standard deviation of the two images, and σ_{xy} denotes the covariance of the pixel blocks in the two images; C_1, C_2, and C_3 are constants to avoid the systematic error when the denominator is 0. SSIM takes values in the range of [0, 1], and the closer the result is to 1, the smaller the distortion is; when the result is 1, it means that the input image and the output image are identical.

4.3 Analysis of Results

To verify the performance of the method in the paper, experimental comparisons and data analysis are performed for different models of single-image super-resolution reconstruction with different data sets and reconstruction magnifications.

The peak signal-to-noise ratio and structural similarity of the algorithms such as SRCNN, FSRCNN, VDSR and the method in the paper were compared at different reconstruction magnifications using the trained models for super-resolution reconstruction of low-resolution images at 2x, 3x and 4x, and the test results are shown in Table 1 and Table 2, respectively.

Table 1. Average PNSR of different algorithms at different reconstruction multiples

Datasets	Multiples	SRCNN	FSRCNN	VDSR	Ours
Set5	×2	36.66	36.87	37.33	37.49
	×3	32.37	33.05	33.56	33.78
	×4	30.07	30.46	31.19	31.24
Set14	×2	32.45	32.57	32.69	32.96
	×3	29.01	29.26	29.61	29.74
	×4	27.49	27.69	27.95	28.01

From the data in the table, it can be seen that the algorithm in the paper achieves better super-resolution reconstruction performance by improving both the PNSR average and SSIM average at different reconstruction multiples compared with other algorithms.

Table 2. Average SSIM of different algorithms at different reconstruction multiples

Datasets	Multiples	SRCNN	FSRCNN	VDSR	Ours
Set5	×2	0.9452	0.9521	0.9543	0.9623
	×3	0.8972	0.9036	0.9126	0.9147
	×4	0.8590	0.8557	0.8830	0.8894
Set14	×2	0.9031	0.9061	0.9181	0.9207
	×3	0.8169	0.8153	0.8317	0.8337
	×4	0.7534	0.7456	0.7674	0.7832

Specifically, in the Set5 data set, the PNSR improved by 0.16 dB, 0.22 dB, 0.05 dB, and SSIM improved by 0.0080, 0.0021, and 0.0064, respectively, compared with the VDSR method at magnifications of 2, 3, and 4; in the Set14 data set, the PNSR improved by 0.27 dB, 0.13 dB, 0.06 dB, and SSIM improved by 0.0026, 0.0020, and 0.01 dB, respectively, compared with the VDSR method at magnifications of 2, 3, and 4. In the Set14 data set, the PNSR is improved by 0.27 dB, 0.13 dB and 0.06 dB, and the SSIM is improved by 0.0026, 0.0020 and 0.0158, respectively, compared with the VDSR method. It can be seen that the reconstruction effect of the method in the paper is overall better than the other three methods.

5. CONCLUSIONS

In order to improve the image reconstruction accuracy and make full use of the image high-frequency information, an image super-resolution reconstruction algorithm based on residual convolution and double attention mechanism is proposed in the paper. Then, we learn the depth features of the image with the help of residual blocks, embed a residual double attention module in each residual block to adaptively calibrate the generated image features, fully utilize the high-frequency features and suppress the invalid information, and finally up sample the features through a sub-pixel convolution layer to reconstruct a high-resolution image of the target size. The experimental results show that compared with SRCNN, FSRCNN and VDSR methods, the method in the paper has a great improvement in both peak signal-to-noise ratio and structural similarity. In the subsequent research work, the method can be tried to be applied to a specific field, such as medical imaging and satellite remote sensing. However, the image reconstruction effect of this network still needs to be improved, and the network design will be further optimized in the future to further improve the super-resolution reconstruction accuracy.

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., “Table 1” or “Figure 2”), please note that the word for Table and Figure are spelled out. Figure’s captions should be centered beneath the image or picture, and Table captions should be centered above the table body

6. REFERENCES

[1] Xu Mengxi,Yang Yun. Super-resolution image video restoration methods and applications [M]. Beijing:People's Post and Telecommunications Publishing House,2020:1-3.

[2] Liu Y. X., Duan T. T.. Research on image super-resolution reconstruction technology based on deep learning[J]. Technology and Innovation,2018(23):40-43.

[3] YANG J，WANG Z，LIN Z，et al. Coupled dictionary training for image super － resolution[J].IEEE Transactions on Image Processing,2012, 21(8):3467－3478.

[4] DONG Chao，LOY C C，HE Kaiming, et al.Image super-resolution using deep convolutional networks［J］.IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(2):295-307.

[5] DONG Chao，LOY C C，TANG Xiaoou．Accelerating the super-resolution convolutional neural network［C］//Computer Vision–ECCV 2016.Amsterdam，The Netherlands: Springer，2016:391-407.

[6] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York:IEEE,2016：1874-1883.

[7] KIM J，LEE J K，LEE K M，et al.Accurate image super-resolution using very deep convolutional networks［C］//Proceeding of the 2016 IEEE conference on computer vision and pattern recognition．Las Vegas，NV，USA:IEEE，2016:1646-1654.

[8] Corbetta M,Shulman G.L.Control of goal-directed and stimulus-driven attention in the brain[J].Nature Reviews Neuroscience.2002,(3)：3201-215.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks.Proceeding of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018:7132–7141.

[10] Woo S,Park J,Lee J Y,et al.CBAM: Convolutional Block Attention Module[J]. Springer,Cham,2018:32-49.

[11] Li Lan,Zhang Yun,Du Jia,Ma Shaobin. Research on super-resolution image reconstruction method based on improved residual sub-pixel convolutional neural network[J]. Journal of Changchun Normal University, 2020(39):23-29.