Coarse-to-Fine Image Super-Resolution Using Convolutional Neural Networks

Liguo Zhou[1], Zhongyuan Wang[1] (*IEEE Member*), Shu Wang[2], Yimin Luo[3]

[1] Computer School of Wuhan University, China
[2] Life Science and Biomedical Engineering of King's College London, UK
[3] Remote Sensing Information Engineering School of Wuhan University, China

Abstract. A deep and wide learning based single image super resolution (DWSR) is proposed in our paper. It is inspired by the previous SRCNN & VDSR method in 2015 & 2016 respectively, while the modification of traditional convolutional neural network (CNN) networks improves the image quality clearly in terms of PSNR & SSIM, also, the necessity of increasing the deep network's kernel size is explored by simulation while the network's depth limit is tested through level changing experiments. Compared to traditional bicubic interpolation, SRCNN & VDSR, our DWSR network shows higher reconstruction accuracy. The dataset is classic ImageNet and all the tested images are of RGB three channels, which means the proposed network is a general applicable method to deal with real world problems. However, the time cost issue and weaker performance on larger testing set should be improved in future researches.

Keywords: Single Super Resolution (SR), Convolutional Neural Network (CNN), Bicubic Interpolation, Super Resolution Convolutional Neural Network (SRCNN), Very Deep Super Resolution (VDSR).

1 Introduction

1.1 Background of Image Super Resolution

Single image super resolution (SR) [1] is a technique that uses some prior models or matching mechanism to seek the matched details of the specified image from given external resources, and then adds to the original one to improve its resolution, while multiple image super resolution (SR) [2] reconstructs the high resolution (HR) image from several low resolution (LR) images, and its core concept is to sacrifice time resolution (more frames) for spatial resolution (clearer images), realize the transferring of two different spaces. SR's main application field contains satellite imaging, video surveillance, virtual reality and biomedical image analysis, etc. After years of development, SR has become more mature and complete, while there are still some obstacles and shortcomings unsolved: its high cost of hardware implementation and less realistic degradation models make it hard to be expanded in realization [3], and the existing algorithms have relatively high complexity and time cost, while now, the reconstruction effect is difficult to improve in terms of PSNR [4] or MSE [5].

Current SR methods can be divided into two classes: 1. SR based on reconstruction method; 2. SR based on learning method. The first class consists of frequency domain methods, like aliasing elimination reconstruction method [6], and spatial domain methods, like bicubic interpolation [7], iterative back projection (IBP) method [8], projection on convex set (POCS) method [9], statistic reconstruction methods [10, 11]. The second class consists of classification methods, like Laplacian Pyramids (LP) method [12], and regression based method like support vector regression (SVR) method [13], RAISR [14] has even drifted SR to be likely in industrial use for its fast reconstruction speed. Learning methods on SR has outperformed other methods as many machine learning and pattern recognition algorithms can be embedded into existing ones, and the recovery object can be chosen to reduce amount of computation, the recovery effect can also be improved in certain level when learning samples are enough.

1.2 Learning based SR and Our Proposed Method.

SRCNN [15] first proposed image super resolution using convolutional neural network, the author designed a shallow end-to-end convolutional neural network, after upscaling the low resolution image by bicubic interpolation method to high resolution image, put the high resolution image to the convolution neural network to get better reconstruction quality compared to traditional methods. Since there is no edge filling in [15], the output image has some missing edges. VDSR [16] made certain improvements based on SRCNN, it designed a very deep network, and added edge filling during convolution process, making the size of images processed by such a deep convolution neural network remain unchanged. VDSR also used residual networks, the convolutional neural network learned the high frequency information from low quality interpolated images, and then added this information to the LR image to obtain corresponding HR image.

As the above 2 methods both acquire HR images from several interpolated LR images, we furthermore proposed a coarse-to-fine multiple layer SR image optimization method, obtain even better quality HR images from LR images by optimizing the HR images after bicubic interpolation with making the layer of the network deeper and wider

2 Related Works

2.1 . Super Resolution & Convolutional Neural Network

Tsai and Huang first propose the concept of super resolution (SR) in [17] in 1984 as an advanced technique to enhance the low resolution (LR) image, in the next few decades, it trigs booming attention and passion in both academic and industrial field of optical imaging, signal processing, even computer vision. Most SR algorithms focus on how to extract the high frequency information that is embedded in the LR images' aliasing spectrum, and estimate the motion blur model accurately for the convenience of restoration. [18] cited that among those algorithms, example-based SR algorithm [19] has achieved state-of-the-art performance. [20] illustrated a general external example-based

method that can learn the mapping between LR/HR patches from external datasets. Then the focus of SR research turns to be the learning of LR/HR patches related dictionary, nearest neighbor [21] and manifold embedding [22] techniques have appeared to solve this reconstruction problem. And certainly, there are some other methods like random forest and kernel regression are applied later to improve the learning performance in terms of speed and accuracy. The proposition of super resolution convolutional neural network (SRCNN) first appears in [18] and it realizes the super resolution image reconstruction with the use of deep learning network.

The basis of deep learning is commonly recognized as convolutional neural network (CNN), which is proposed by LeCun Yang in 1989 in [23]. It imitates the biological neuron propagation procedure, trains the computer to learn needed parameter for the use of image recognition, recovery, analysis, etc. Generally, there are 3 parts of CNN: input layer, output layer, and multiple hidden layers. Hidden layers can be classified as convolutional layer, pooling layer, and fully connected layer. Convolutional layer emulates the real response of biological neuron stimulation, it operates a convolutional kernel with arbitrary patch size to the input and pass it to the next layer, one convolutional layer will be followed by one pooling layer in the classic CNN structure stated in [24], after several times of convolution and pooling, the output will be the input of fully connected layer, then all the hidden identity features are learned and the final estimation results will be shown from the probability distribution of softmax layer.

2.2 Previous Learning based Super Resolution Methods

In this paper, we mainly investigate three previous super resolution image enhancement methods: bicubic [7], SRCNN [15], and VDSR [16], SRCNN and VDSR are the most cutting-edge learning based SR techniques among all the algorithms.

Bicubic interpolation is one basic algorithm used for scaling images or video display, as it can preserve more details of the image than bilinear and nearest neighbor algorithms, it is a widely used SR method for a long time. The core formulation can be found in [25], which is proposed by Keys in 1981.

However, bicubic interpolation algorithm has relatively high computation, and it cannot calculate the first row, first column, last two rows and last two columns, besides, the accuracy of its restoration is still limited.

Compared to traditional SR algorithms which only take images as a form of signal, learning based SR algorithms also take the content and structure of the image into consideration, and use data related prior knowledge to provide stronger restriction, thus, it can achieve better reconstruction performance.

[15] introduces a famous SR algorithm manipulating deep learning network and its name is super resolution convolutional neural network (SRCNN). The core method of SRCNN is adding a three layers convolutional neural network after doing bicubic interpolation, thus, it can be regarded as an equal framework of sparse coding based SR. (补充SRCNN 算法概述)

[16] then deepens the network to 20 layers and utilizing a method named residual learning in SR recovery, the higher PSNR result proves its feasibility and proposed

theory: the deeper, the better. However, the method's extension is limited by its difficulty of creating such a deep network and the relevant high time cost. (补充VDSR 算法概述)

3 Proposed Methodology

As image super-resolution can be viewed as a mapping problem from the low resolution image to the high resolution image, we can use Y=*f* (X) to represent the process, where X represents the LR image, Y represents the HR image, and the function *f* is the mapping relationship between them. For the deep learning based SR method, *f* stands for a nonlinear mapping, and our network structure is *f* (•), the training process is to solve the nonlinear regression problem. Next, we will explain our method in terms of the network structure and training procedure.

3.1 Method Framework

SRCNN and VDSR have shown that fine high-resolution images can be obtained from coarse interpolated images. We design a cascaded convolutional neural network, which makes the quality of the image can be refined from coarse to meticulous, and finally, the quality of the image is better than that of only one stage neural network.

Our network consists of three stages which is shown in Figure 1: the first stage takes the interpolated image as the input and the output image's quality can be slightly improved, the second and third stage takes the last layer's output as the input respectively, then the image quality is gradually refined. Each CNN stage contains ten convolutional layers with 3×3 kernel size, and each convolutional layer produces 64 feature maps except for its output layer. The output layers' channel number should be equivalent to the channel number of input image. To keep the output image size unchanged, we pad zeros around the border.

In order to speed up the training process, inspired by VDSR, we also use residual learning. Each CNN stage's output is the high frequency detail of the predicted high resolution image, which is then added to the stage's input to synthesize the SR image.

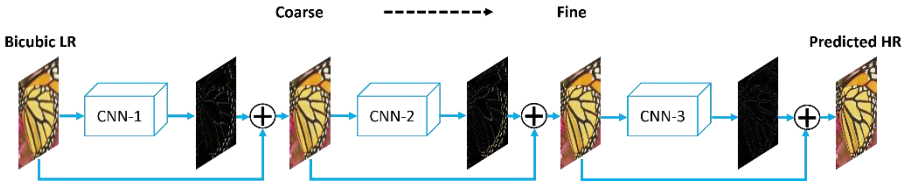

Fig. 1. The Core Network of Our Coarse-to-Fine Super Resolution Method

3.2 Training

For a given training sample pair (X_i, Y_i), X_i represents the coarse image processed by bicubic interpolation, and Y_i represents the ground truth high resolution image. Putting

X_i into the network, a predictive high-resolution image Y_i' will be yielded. The purpose of training is to make the difference between Y_i' and Y_i as small as possible. To judge the similarity between two images, PSNR (1) is the most popular evaluation standard, the larger the PSNR value is, the more similar the two images are. Therefore, our training is to increase PSNR value between Y_i' and Y_i, and this is equivalent to minimize the two images' MSE (2), or minimize their Euclidean distance, so our loss function can be expressed in (3).

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{(2^n-1)^2}{MSE}\right) \tag{1}$$

$$\text{MSE} = \frac{1}{width \times height} \|Y_i - Y_i'\|^2 \tag{2}$$

$$\text{Loss} = \frac{1}{2N}\sum_{i=1}^{N}\|Y_i - Y_i'\|^2 \tag{3}$$

Where n is the number of bits of the image pixel, generally $n = 8$, and N is the batch size.

Similar to VDSR, we also train a general model that can zoom in the LR image at different ratios. SSIM (4) is commonly used in image similarity calculation.

$$\text{SSIM}\ (Y_i, Y_i') = \frac{(2\mu_{Y_i}\mu_{Y_i'}+c_1)(2\sigma_{Y_i,Y_i'}+c_2)}{(\mu_{Y_i}{}^2+\mu_{Y_i'}{}^2+c_1)(\sigma_{Y_i}{}^2+\sigma_{Y_i'}{}^2+c_2)} \tag{4}$$

Where μ is the mean value, σ is the variance, and $\sigma_{Y_i,Y_i'}$ is the covariance, c_1=$(0.01L)^2$, c_2=$(0.03L)^2$, for a RGB image, $L = 255$. The interval of SSIM is [-1, 1], when SSIM=1, it means the two images are the same. Both PSNR and SSIM are evaluated in dB unit.

4 Experiments & Results

4.1 Experiment Implementation Details

The experiment concrete steps can be expressed below:

ImageNet dataset is used as our training set. Original oversized image is cut into 96x96 non-overlapping image block as groundtruth set {Yi}, then downsample each image block by 1/2, 1/3, 1/4 ratio of {Yi} respectively, and use bicubic interpolation to recover them to the original size, thus, low resolution image set {Xi} is obtained. We utilize the classic Caffe library to train our DWSR network. The input image is in RGB color space, its batch size is set to 160, and Adam gradient descent algorithm is applied here to optimize the network, calculate losses of each stage's output and back propagate as the input's feedback, therefore, in the process of training our network, there are three losses needed to be calculated in total, their relevant loss weights are set to 1.

Then we come to the step of training the three-stage CNN network. Its initial learning rate of the network is set to 0.0001, when the loss is no longer dropped, it is reduced at a rate of 0.5 and stops changing until the loss is converged.

Experimental results show that the image quality of our network's first stage output can be as good as SRCNN, the second stage's output can reach VDSR level, and the recovery effect of the third can exceed VDSR. Except on PSNR standard, the output images of our network also perform well when evaluated by SSIM.

4.2 Results & Analysis

From the red box in Fig 2, we can clearly see that VDSR cannot reconstruct the exactly detailed image as the original one, the lines are blurred in certain level, while our method DWSR eliminates such a phenomenon, clarifies how amazing the learning based SR can do.

Furthermore, we compare four frequently used SR algorithms by computing the PSNR & SSIM, there is no question that our method outperforms the other three. To be specific, on the smallest set set5 (only 5 tested images) and upscaling by the factor of 2, the PSNR value and the SSIM value are improved by 11.91% and 3.12% respectively compared to bicubic interpolation, 2.81% and 0.57% compared to SRCNN, 0.43% and 0.09% compared to VDSR; when the upscaling factor increases, the PSNR and SSIM of each algorithm unavoidably decrease, but for parallel comparison, our DWSR still has the best performance. And there is one fact that is worth pointing out, when the tested set becomes larger, the recovery results will get worse, which can be explained by that bsd100 set (100 tested images) corresponds to the lowest PSNR & SSIM.

Also, by observing the output of three stages' loss in our training network, we find that the quality improvement from the first stage to the second stage is significantly higher than that from the second to the third. We try to increase the network depth to four and five stages in the training process, and find it more difficult to get significant improvement while the increase of the training & testing time and complexity is huge. So there is no need to deepen the current network anymore, the modification should be found from other aspects.

(a) Original HR Image (b) VDSR Recovered Image

(c) DWSR Recovered Image (ours)

Fig. 2. Comparison Between the Original HR Image, VDSR Recovered Image and DWSR Recovered Image

Table 1. Comparison Between the 4 Algorithms' Performance In Terms of PSNR & SSIM

		Bicubic		SRCNN		VDSR		**DWSR(Ours)**	
		psnr	ssim	psnr	ssim	psnr	ssim	psnr	ssim
set5	x2	33.68	0.9306	36.66	0.9542	37.53	0.9587	**37.69**	**0.9596**
	x3	30.41	0.8688	32.75	0.9090	33.66	0.9213	**34.01**	**0.9241**
	x4	28.43	0.8103	30.49	0.8628	31.35	0.8838	**31.68**	**0.8874**
set14	x2	30.24	0.8691	32.45	0.9067	33.03	0.9124	**33.48**	**0.9164**
	x3	27.54	0.7738	29.30	0.8215	29.77	0.8314	**30.14**	**0.8361**
	x4	26.00	0.7015	27.50	0.7513	28.01	0.7674	**28.37**	**0.7724**
bsd100	x2	29.56	0.8437	31.36	0.8879	31.90	0.8960	**32.01**	**0.8976**
	x3	27.21	0.7391	28.41	0.7863	28.82	0.7976	**28.95**	**0.8008**
	x4	25.96	0.6680	26.90	0.7101	27.29	0.7251	**27.41**	**0.7292**

5 Conclusion & Future Works

In conclusion, we propose a fast, easy implemented learning based SR method by using a 3-layer deep CNN network with wider convolutional kernel size, name it DWSR, and it is proven by simulation this method can outperform other current SR method to some extent. Even though it reduces the network depth in VDSR, it can still complete a higher quality SR reconstruction, and the better performance is due to the help of wider kernel size. Additionally, the worst performance of traditional bicubic interpolation proves the necessity of adding a deep learning network.

However, there is still some limitation of our proposed method. The weak recovery on larger set should be paid more attention, and whether the time cost of the algorithm can be shortened while still maintain the results' quality so that the technique is more applicable.