Comparison of Convolutional Neural Network Models for Image Super Resolution

*Chen Jian, **Songhyun Yu, ***Jechang Jeong

Department of Electronics and Computer Engineering, Hanyang University

*chenjian105@naver.com, **3069song@naver.com, ***jeong@hanyang.ac.kr

Abstract

Recently, a convolutional neural network (CNN) models at single image super-resolution have been very successful. Residual learning improves training stability and network performance in CNN. In this paper, we compare four convolutional neural network models for super-resolution (SR) to learn nonlinear mapping from low-resolution (LR) input image to high-resolution (HR) target image. Four models include general CNN model, global residual learning CNN model, local residual learning CNN model, and the CNN model with global and local residual learning. Experiment results show that the results are greatly affected by how skip connections are connected at the basic CNN network, and network trained with only global residual learning generates highest performance among four models at objective and subjective evaluations.

1. Introduction

Single image super-resolution (SISR) is a classic computer vision problem that aims to reconstruct a high-resolution (HR) image from one single low-resolution (LR) input image. SISR can be widely used in many areas such as medical imaging [1, 2], and face recognition in surveillance videos [3, 4], where more image details are required. Over the past few years, many SISR methods have been exploited in the computer vision community, such as the sparse-coding-based SR [5, 6] which is the representative approach. In recent years, due to the powerful learning ability of the deep learning technology, a convolutional neural network (CNN) model which learns a mapping from LR to HR image patches with large image datasets has been widely used in vision tasks ranging to super-resolution (SR). Super-resolution using convolutional neural network (SRCNN) [7] interpolates the input LR images using bicubic interpolation as desired image size, and predicts a nonlinear LR-to-HR mapping using CNN. The SRCNN method is one of the representative SR methods which uses 3-layers CNN: patch extraction/representation, non-linear mapping and reconstruction. Image super-resolution using very deep convolutional networks (VDSR) [8] accelerates the convergence speed and enhances performance of very deep networks utilizing residual learning [9]. The VDSR stacks 20 weight layers (3×3 for each layer), in order to speed up the convergence, and also uses the global residual learning method. Similarly, the enhanced deep residual networks for single image super-resolution (EDSR) [10] utilizes local residual learning and optimizes the conventional network by removing unnecessary modules such as batch normalization layers, therefore it has significant performance improvement. In this paper, we compare the results obtained from four different networks, which are composed of the general network model, global residual learning network model, local residual learning network model, and a network with both global and local residual learning.

The organization of this paper is as follow: in Section 2, we describe the structure of the four CNNs and training methods, Section 3 represents experiment results with analysis, and Section 4 concludes this paper.

2. Proposed Method

2.1. Proposed Network

Generally, training a CNN model involves two steps: network architecture design and learning from training data. In SR, the network is designed to predict HR images from a LR input image. In this section, we compare four CNN models for SR. We construct 5-layers CNN structures for each network. Fig. 1, shows structure of four convolutional neural network models. The first model is general network (Fig. 1, General Network), the layers have the same number of filters, so layers except last are of the same type: 64 filters reconstruction and has 1 filter of size 3×3×64. In order to
solve the problem of the vanishing/exploding gradient, the residual learning [9] method is used. Based on the above general network, we adopt the global residual learning method in the second CNN (Fig. 1, Global Network), the residual image is estimated from the input and output of the network. The SR output is vastly similar to the input, therefore the global residual learning method eases the difficulty of training network. In addition, for comparison, we use an additional residual unit structure, termed as local residual learning in third network (Fig. 1, Local Network), where the residual unit not only in order to prevent the loss of image details, but also helps gradient flow. The residual unit structure is formulated as:

\[ H^u = F(H^{u-1}, W^u) + H^{u-1} \tag{1} \]

where \( u = 1, 2, \cdots, U \) and \( U \) is the number convolution layers, \( H^{u-1} \) and \( H^u \) are the input and output of the convolution layer \( u \). Activation function is denoted as \( F(\cdot) \), and \( W \) means weights of filters. However, because the number of feature maps are different at first and last layers, local skip connection is not connected at the first and last layers. Lastly, the fourth CNN (Fig. 1, Both Global and Local Network) utilizes both global and local residual learning. In this network, both global and local residual learning are utilized at single model.

### 2.2 Training

We analyze the effective patch size and stride in the network model. Finally, by considering both the performance and network complexities, we split input images into \( 41 \times 41 \) patches with the strides of 31. In image classification application [11], performance of the network can be enhanced by training data augmentation. The most commonly used methods include cropping, flipping, and rotation. In SR, rotation and flipping of input images lead to the better results [12]. Therefore, As show in Fig. 2, we apply rotation and flipping on the training image patches. Finally, we use the input patches that is rotated by 90º, 180º, 270º, and flipped for data augmentation. As the network activation function, the LeakyReLU function is used. For the training, given a training dataset \( \{ (x^{(i)}, y^{(i)}) \}_{i=1}^{N} \), where \( N \) is the number of training patches and \( y^{(i)} \) is the ground truth HR patch and the \( x^{(i)} \) is the LR patch, the mean squared error (MSE) is used as the loss function, which is defined as:

\[ L = \frac{1}{N} \sum_{i=1}^{N} \| r - f(x) \| ^2 \tag{2} \]

where \( x \) denotes interpolated LR image and \( y \) means the corresponding HR image. Our goal is to learn a model \( f \) that predicts values \( y = f(x) \). We define the residual image \( r = y - x \). We train all of the CNN models during 50 epochs. Training is carried out by using Adam optimizer, and we set the learning rate to 1e-5.

### 3. Experiment Results

In this section, we evaluate the performance and results on several datasets using the four network types. For the network training, we use 291 images [13] as the training data, and Set5 [14] iamgeset as validation dataset. We train four network models with the upscaling factor 3. The Set5, Set14 [15], and Urban100 [16] are used as test dataset to evaluate the performance of each network.
All images are downsampled using the bicubic kernel. We adopt the peak signal-to-noise ratio (PSNR) to obtain the average quantitative results of the test datasets, and it summarized in the Table I. According to the experimental results, for super-resolution, the results of using global residual learning network model have the highest performance in objective evaluation (PSNR). Compared with the general network model, local network and global network, show the average PSNR improvements by 0.05dB and 0.27dB respectively in dataset Set14. The global network is effective in easing the difficulty of training networks. The local residual network model use the several residual units to learn residual image between the HR and LR images, where the residual unit branch can carriers rich image details, so it outperforms the existing general model. The local learning network model and global residual learning network results have obvious improvement. When we use the both global and local residual learning network model, there is no obvious difference in the results. If both global and local network is trained with more epochs in the deeper network, it may improves results, but in our test with 5-layers and 50 epochs, it does not improves training speed or network performance. With experiment results, the different residual learning methods have significant effects on the super-resolution results. The Fig. 3 and Fig. 4 compare the image results of super-resolution methods using four network models respectively by an upscaling factor 3. As can be observed, all of the network models without severely blurry or distortionary results. The global network model have better result than other models, and the global network model produces relatively sharper edges than other approaches.

4. Conclusion

In this paper, we propose a comparison of four CNN models for SISR. We use the general network structure, global residual learning network structure, local residual learning network structure, and both global and local residual learning network structure for super-resolution, and compare their results. With Experiment results, we can find

<table>
<thead>
<tr>
<th>Dataset</th>
<th>General Network</th>
<th>Global Network</th>
<th>Local Network</th>
<th>Both Global and Local Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set5</td>
<td>32.03</td>
<td>32.39</td>
<td>32.05</td>
<td>32.00</td>
</tr>
<tr>
<td>Set14</td>
<td>28.75</td>
<td>29.02</td>
<td>28.80</td>
<td>28.74</td>
</tr>
<tr>
<td>Urban100</td>
<td>25.54</td>
<td>25.84</td>
<td>25.59</td>
<td>25.54</td>
</tr>
</tbody>
</table>

Table I: The average results of PSNR for scale factor 3 on dataset Set5, Set14, Urban100.

Fig. 3: Single image super-resolution results of “butterfly” from Set5 dataset with upscaling factor 3.

Fig. 4: Single image super-resolution results of “zebra” from Set14 dataset with upscaling factor 3.
different residual learning methods has significant impact on the results of SR. As a result, global residual learning shows highest performance in the test, and most suited for training 5-layers SR.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP.[ 2014-0-00670, Software Platform for ICT Equipment]

References


