Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 139 (2018) 144–150

Procedia Computer Science

www.elsevier.com/locate/procedia

The International Academy of Information Technology and Quantitative Management, the Peter Kiewit Institute, University of Nebraska

Lightweight Convolutional Neural Network with SE Module for Image Super-Resolution

Yuwen Wu[a,b], Xiaofei Zhou*[a,b], Ping Liu[a,b], Jianlong Tan[a,b], Li Guo[a,b]

[a]Institute of Information Engineering, CAS, China
[b]School of Cyber Security, University of Chinese Academy of Sciences, China

Abstract

In recent years, research on single image super resolution has progressed with the development of deep convolutional neural networks(DCNNs). Among current techniques, models based on residual learning demonstrated great progress. Despite their great performances, the depth and width of the super-resolution models has increased a lot, which brought the challenges of computational complexity and memory consumption. In order to solve the above questions, attention has been paid to improving model efficiency. In this work, we address this issue by proposing a novel model with new residual block and new training method. By introducing the squeeze and excitation(SE) module and depthwise separable convolution, we can get a slimmer model with more efficiency. In addition, we apply a cascade training approach in training our model. Experiments on benchmark datasets show that our proposed image super resolution model achieves the state-of-the-art performance with fewer parameters and less time cost.

Keywords: super resolution; convolutional neural network; model compression

1. Introduction

Image super resolution is a classical problem in low-level problem in computer vision, especiall single image super resolution(SISR) problem, which aims to restore a high resolution(HR) image from a given low resolution(LR) image. It's obvious that one low resolution image can be downsampled from various high resolution images, therefore SR problem is ill-posed with many-to-one mapping. A lot of SR algorithms have been proposed, such as interpolations, image priors and sparse coding. They can be simply divided into three categories, including interpolation-based methods, reconstruction-based methods and example-based methods.

*Corresponding author. Tel: +86-13552951803.
E-mail address: zhouxiaofei@iie.ac.cn.

10.1016/j.procs.2018.10.230

Recently, deep convolution neural networks(DCNNs) have proven to be effective models in various computer vision tasks. It also shows outstanding performance in SISR tasks. SRCNN [1] is the pioneering work to solve SR problem with CNN-based models. Although SRCNN only contains three convolution layers and a small receptive field, it still outperforms most SR models at that time. To improve its accuracy and larger receptive field, a number of improved models are proposed. VDSR[2], a 20-layer CNN model with skip connections and adaptive gradient clipping is proposed. Deep recursive structures are adopted in DRCN[3] and DRRN[4], which reduces the enormous parameters brought by deep networks and reduce redundant parameters. The models mentioned above take bicubic interpolation as a predefined upsampling operator to upscale the input images before input to the network, which leads to unnecessary computational complexity, especially in large-scale tasks such as 8x. To deal with these issues, transposed convolution and sub-pixel convolution[5] are proposed and widely in FSRCNN[6], ESPCN[5], LapSRN[7], SRGAN[8]. By stacking a series of convolution layers, these model are capable of capturing features for image restoration.

Despite current prominent performance, there still remains some drawbacks. To gain better performance, super resolution models tends to be designed deeper and wider, which leads to growth in computational complexity and memory consumption. In devices with limited computing power such as mobile, inference speed is of great importance in user-experience perspective. Thus these models are less practical.

To handle these drawbacks, we proposed a Cascading Trained Squeeze and Excitation Convolutional Network (CT-SECN) model as solution. We build squeeze and excitation(SE) convolution block and apply it to origin model. SE[9] module helps to reweight the features and select information among channels, therefore improving the efficiency of the model. Then we apply depthwise separable convolution[10] and extend our model to CT-SECN. We find the performance of CT-SECN is competitive compared with the state-of-the-art SR models and only 25% of the parameters are used compared with SECN. In addition, we explore a better approach to train CT-SECN model. Instead of training all layers initialized with random weights, we gradually replace the convolutional block in our model. We show that with the cascading training method, our model can achieve better performance and get a good trade-off between the performance and heaviness of the model.

Our main contribution can be summarized in the following points: 1)We propose a novel neural network model for super resolution by introducing squeeze and excitation convolution block. Our model is proved to achieve the state-of-the-art performance. 2)We introduce depthwise separable convolution in convolutional block and find it works well with good computational efficiency and significantly smaller size. 3) A cascade training approach is utilized, which proves to perform better than training from scratch.

2. Related Works

2.1. Classical super resolution methods

Numerous research has been taken on image super resolution problem. Early SR algorithms are based on interpolation,such as nearest, bicubic and Lanczos. Although these algorithms are widely used nowadays due to their fast speed, their outputs are usually too smooth and lack high-frequency information, which is not satisfying. Other approaches are based on reconstruction, which assume mapping between LR space and HR space, such as A+[11] based on anchored neighborhood regression, SelfExSR[12] based on transformed self-exemplars.

2.2. Image Super-Resolution Based on Deep Learning

Since the success of Alexnet[13] in Imagenet classification task[14], deep convolutional neural networks are widely used in various computer vision tasks. Dong et al. proposed SRCNN[1], which is the first deep learning model to solve super-resolution task. With a three-layer convolutional network, SRCNN outperformed most of traditional algorithms. Following work improves image super-solution model based on deep learning from two kinds of approaches. On one approach, some researchers propose

to use more complicated networks. Kim et al. proposed VDSR[2], a 20-layer CNN model with small filters, skip connections and adaptive gradient clipping. In order to make the model deeper without extra parameters, DRCN[3] and DRRN[4] are proposed. They are designed with recursive structure and skip connections, which reduces redundant parameters. Christian Ledig, et al. proposed SRResNet [8], a very deep residual network similar with ResNet and further employed generative adversarial network to generate images with better visual feelings. Lim et al. proposed EDSR[15], which removed unnecessary modules in convolutional structure and proved its excellence by winning the NTIRE2017 Super-Resolution Challenge[16]. On the other approach, researchers focused on optimizing the efficiency of convolutional networks. Dong et al. proposed FSRCNN[6], which upsamples images at the end of the networks, while SRCNN and the algorithms above upsample images before input. This leads to the reduction in computational complexity. In ESPCN[5], Shi et al. introduced an efficient sub-pixel convolutional layer to take place of simple transposed-convolutional layer and achieved better performance. Lai et al. proposed LapSRN[7] to progressively reconstruct residuals of high-resolution images from different scale factors.

2.3. Efficient CNN structure

For some computer vision tasks, deeper model does not perform much better than the shallower one, which means explosive growth of the size of network brings limited improvement in the final performance. Thus rising attention has been paid to build small and efficient neural network. Landola et al. proposed SqueezeNet[17] with comparable performance and 50x fewer parameter compared with AlexNet. MobileNet[18] and MobileNetV2[19] build an efficient network with depthwise separable convolution, inverted residual and linear bottlenecks. Han et al.[20] proposed deep compressing techniques to reduce the size of pretrained network, including pruning, vector quantization and Huffman coding.

3. Proposed Methods

3.1. Network Structure

The overall structure of CT-SECN is shown in Fig. 1. The model can be simply divided into three parts: feature extraction, non-linear mapping and reconstruction. Feature extraction includes one convolutional layer to map image channels to various feature maps. Non-linear mapping consists of several squeeze and excitation residual blocks and we?ll discuss about it in the next section. Finally, the output entered reconstruction part. Instead of transposed-convolution, we use pixel shuffle method mentioned in ESPCN, which reduces computational complexity. To better train our CT-SECN model, we introduce cascade training approach, which will be discussed in 3.4. Considering the balance of performance and model size, we set the number of block B=16 and the number of feature channel F=64.

We use the L1 loss as our loss function instead of L2 loss. The latter is widely used in various tasks due to its relationship with the peak signal-to-noise ratio (PSNR). However, Lim et al.[15] demonstrate that training with L2 loss does not guarantee better performance in terms of PSNR and SSIM. In our experiment, L1 loss performs better than L2 loss.

3.2. Squeeze and Excitation Residual Block

We design our convolutional unit based on residual block. Follow EDSR, we removed batch normalization(BN) layer and it turns out to perform better. Each residual block contains two convolutional layer, one ReLU layer, squeeze and excitation(SE) module.

Inspired by SENet, we introduced SE module compared in widely-used ResNet-based residual block, as shown in Fig. 2:

Squeeze and excitation module manages to calculate weights of the output convolutional channels, which makes the network more efficient by emphasizing important features and suppressing useless features among channels. We first use global average pooling to generate channel-wise vector. The squeeze function is shown as below:

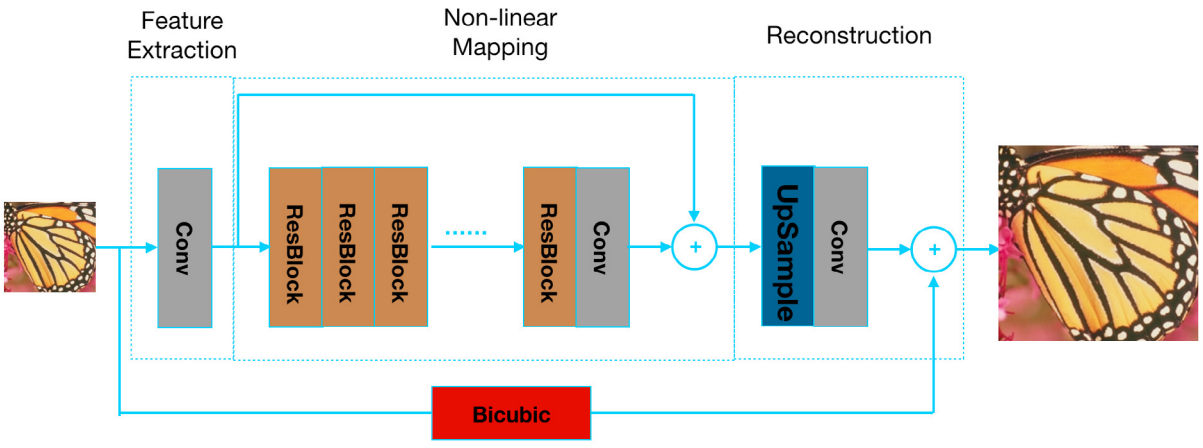

Fig. 1. Overview of model.

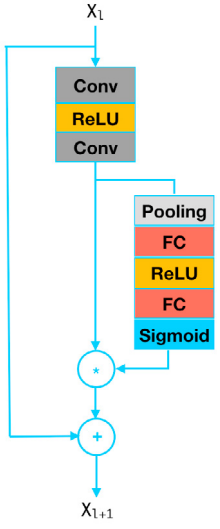

Fig. 2. Squeeze and Excitation Residual Block.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

H and W denotes height and width of image, while c denotes the c-th element of squeezed channel. For the input feature map size $H \times W \times C$, the output size will be $1 \times 1 \times C$.

The excitation function is shown as below:

$$s = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

W_1 and W_2 denotes two FC layers, σ denotes Sigmoid function and δ denotes ReLU function.

3.3. Depthwise Separable Convolution

To reduce computational complexity and memory consumption, we introduce depthwise separable convolution. Each convolution operation in residual block is split into channel-wise spatial convolution and pointwise convolution, without non-linear function between them. As shown in figure Fig. 3:

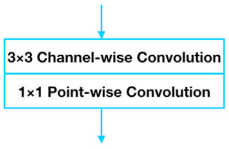

Fig. 3. Depthwise Separable Convolution.

By introducing depthwise separable convolution, we can achieve comparable performance with much fewer parameters and fewer compute operations compared with custom convolution operation.

3.4. Cascade training approach

Inspired by CT-SRCNN[21], we apply cascade training approach in training our neural network. Instead of training CT-SECN model from scratch, we first train SECN model, which uses custom convolution operation in each residual block. In each stage, we substitute convolution operation in two residual block with depthwise separable convolution, then fine-tune the whole network until convergence. No weights will be frozen in the training process. Convolutional layer in feature extraction part and reconstruction part will not be replaced. When all residual blocks are substituted, the fine-tune process ends.

4. Experiment

4.1. Datasets

In this work we use DIV2K[22] to generate the training LR and HR patches. DIV2K is a newly proposed high-quality image dataset for SR tasks, which consists of 800 training images, 100 validation images and 100 test images. The model is evaluated on four standard benchmark datasets, Set5[23], Set14[24], BSD100[25] and Urban100[12]. These datasets contains natural scenes and urban scenes images.

4.2. Implementation Details

For training, we randomly sample patches of size 192 192 for training dataset as RGB input patches. To make full use of training data, for each patch, we randomly flip images horizontally and rotate the images with degrees of 90°, 180° and 270°. Following EDSR[15], we pre-process all the images by subtracting the mean RGB value of the DIV2K dataset. For testing, we downscale the ground truth images with bicubic interpolation to generate LR/HR image pairs. RGB images of the four benchmark datasets were converted to YCbCr color space and only Y channel is evaluated. We set upsample scale at 2.

We train our model with ADAM[26] optimizer. Learning rate begins at 1e-4 and is halved every 100 epochs. Batch size is set to 16. As mentioned above, we use L1 as our loss function.

In experiment, we used four NVIDIA Titan XP for training our models. The model is built with PyTorch[27].

4.3. Experiment results

As mentioned above, our CT-SECN model is fine-tuned from SECN model, so we first train SECN model and compare its performance with current state-of-the-art architectures, as shown in Table 1.

For model EDSR and SECN, *BxFy* denotes the model contains *x* residual blocks and *y* feature maps. Compared with former CNN-based model, our SECN model performs better. Compared with EDSR with the same number of residual blocks and feature maps, SECN performs better than EDSR, which means squeeze and excitation module assists the non-linear mapping of the SR model effectively.

We use SECN_B16F64 to train our CT-SECN model. *n* in CT-SECN_n denotes the number of residual blocks using depthwise separable convolution. The experiment results are shown in Table 2.

Table 1. PSNR evaluation of models.

Dataset	Bicubic	A+	SRCNN	VDSR	EDSR_B16F64	SECN_B16F64	EDSR_B32F256	SECN_B32F256
Set5	33.66	36.54	36.66	37.53	37.95	38.02	38.11	38.20
Set14	30.24	32.28	32.42	33.03	33.57	33.59	33.92	33.96
BSD100	29.56	31.21	31.36	31.90	32.16	32.18	32.32	32.34
Urban100	26.88	29.20	29.50	30.76	31.98	32.15	32.93	32.95

Table 2. PSNR evaluation of models.

	CT-SECN_0(SECN)	CT-SECN_4	CT-SECN_8	CT-SECN_12	CT-SECN_16	CT-SECN(From Scratch)
PSNR(Set5)	38.02	38.01	38.00	37.94	37.86	37.46
Para.	1379K	1122K	865K	608K	351K	351K

Note that residual blocks are replaced from back to front, we find that there is no distinct drop in PSNR when blocks in the deep layer are replaced. For our final model CT-SECN_16, we find it still shows competitive result with only 1/4 parameters. Also, we tried to train CT-SECN from scratch instead trained from SECN, results shows that there is an obvious performance decay.

5. Conclusion

In this paper, we proposed a cascading trained squeeze and excitation convolutional network for super resolution(CT-SECN). The model achieves both high performance and efficiency. SE module improves network efficiency by weighting features among channels. Depthwise separable convolution reduces the parameters of model without markable decay in performance. Cascade training approach helps to better train our model without any change in model structure. Experimental results on benchmark datasets show that our model achieves competitive performance compared to current state-of-the-art models with less parameters and faster speed.

6. Acknowledgements

This work is supported by National Key R&D Program 2016 (No.2016YFB0801300), and the National Natural Science Foundation of China (No.61202226),We thank all anonymous reviewers for their constructive comments.

References

[1] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE transactions on pattern analysis and machine intelligence 38 (2) (2016) 295–307.
[2] J. Kim, J. Kwon Lee, K. Mu Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.
[3] J. Kim, J. Kwon Lee, K. Mu Lee, Deeply-recursive convolutional network for image super-resolution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637–1645.
[4] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 5.
[5] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
[6] C. Dong, C. C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: European Conference on Computer Vision, Springer, 2016, pp. 391–407.
[7] W.-S. Lai, J.-B. Huang, N. Ahuja, M.-H. Yang, Deep laplacian pyramid networks for fast and accurate superresolution, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2017, p. 5.
[8] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, arXiv preprint.

[9] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, arXiv preprint arXiv:1709.01507 7.
[10] F. Chollet, Xception: Deep learning with depthwise separable convolutions, arXiv preprint (2017) 1610–02357.
[11] R. Timofte, V. De Smet, L. Van Gool, A+: Adjusted anchored neighborhood regression for fast super-resolution, in: Asian Conference on Computer Vision, Springer, 2014, pp. 111–126.
[12] J.-B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197–5206.
[13] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.
[15] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in: The IEEE conference on computer vision and pattern recognition (CVPR) workshops, Vol. 1, 2017, p. 4.
[16] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, et al., Ntire 2017 challenge on single image super-resolution: Methods and results, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE, 2017, pp. 1110–1121.
[17] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size, arXiv:1602.07360.
[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861.
[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[20] S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, arXiv preprint arXiv:1510.00149.
[21] H. Ren, M. El-Khamy, J. Lee, Ct-srcnn: Cascade trained and trimmed deep convolutional neural networks for image super resolution, arXiv preprint arXiv:1711.04048.
[22] E. Agustsson, R. Timofte, Ntire 2017 challenge on single image super-resolution: Dataset and study, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.
[23] M. Bevilacqua, A. Roumy, C. Guillemot, M. L. Alberi-Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding.
[24] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International conference on curves and surfaces, Springer, 2010, pp. 711–730.
[25] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, Vol. 2, IEEE, 2001, pp. 416–423.
[26] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[27] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch.