Iran University of Science and Technology
Deep Learning (Spring 2020)

Single Image Super Resolution

Taha Samavati
Department of Computer Engineering
Iran University of Science
and Technology
taha_samavati@comp.iust.ac.ir

Abstract

Single Image Super Resolution (SISR) is an important task of computer vision, in which we seek to improve resolution of images. In recent years, due to the increase in computing power, significant progress has been made in the field of deep learning. Convolutional neural networks form the basis of many machine vision algorithms including SISR. Generating realistic high resolution images not only requires deployment of very deep convolutional architectures, but also requires adversarial training approach. In this project we use a very deep residual generator network to generate realistic high resolution images and train it in an adversarial manner. We propose a new fully convolutional discriminator architecture which has fewer parameters than regular CNN discriminators and is more efficient to train. We show that FCN discriminators are capable of outperforming traditional discriminators with much lower parameter count. We also add DCT loss function to the perceptual loss to enhance high frequency reconstruction ability of the generator.

1 Introduction

Given a low resolution image (LR), Image Super Resolution (ISR) estimates a High Resolution (HR) image. This challenging task is an important technique in computer vision and image processing. It has applications in a wide range of fields such as medical imaging, surveillance and security. It is also used as a pre-processing step for some vision task to improve the results[1].

With emerge of convolutional neural networks, rapid improvement in various fields of computer vision research happened. In 2014, the first CNN based SISR algorithm was introduced [2]. In recent years, The CNN based methods have been often achieved state-of-the-art results. more recently, GANs form the basis for this task.

The ill-posed problem of ISR has a number of challenges including which type of loss function to use in order to get a perceptually acceptable result, which type of network architecture to use, which learning strategy to select or other types of challenges which are application specific such as inference time. Also it is worth mentioning that up sampling is always required to enhance resolution of the image. Therefore, having mentioned this, the up sampling method is also a game changing factor.

2 Related work/Background

In 2014, Dong et al [2] first used Convolutional neural networks for the task of image super resolution. That research changed and paved the way for other researchers which were using conventional

(LateX template borrowed from NeurIPS 2019)

algorithms for this problem. the aforementioned research (SRCNN) achieved acceptable results but it's inference time was quite high. to mitigate this issue, Dong et al [3], proposed Fast SRCNN. The proposed model eliminated the bicubic interpolation step which was done in first stage of the SRCNN to . Indeed, using bicubic interpolation caused the convolutions to apply on images with larger spatial dimensions and as a result inference time grows very large. In FSRCNN, This step was removed. Instead in the final layers of the model, after convolutional layers had been applied to the input, a number of deconvolutional layers were used for the up-sampling stage.

The two former methods use mean squared error as the loss function to train the model. Models trained with this loss function fail to enhance high frequency details of the image hence the output would be blurry, lacking details. To solve this issue Johnson et al [4] proposed a new loss function called "*Perceptual Loss*" which computes mean squared error on activation maps of a specific intermediate layer of a pretrained VGG19 model. They demonstrated that a network trained with this type of loss function is capable of reconstructing and enhancing high frequency details. A year later, along with increasing popularity of generative adversarial networks, Ledig et al [5] proposed a new loss function which combines adversarial loss function with "*Perceptual Loss*". They reported that using adversarial training procedure highly improves network performance to generate realistic high resolution images. They also changed the up sampling method and used the pixel shuffle up sampling proposed by Shi [7]. This up sampling method is capable of producing better quality up sampled images by eliminating the de-convolution operation in which some unhelpful and unmeaning zeros are added between pixels. This periodic up sampling method is shown in figure 2.

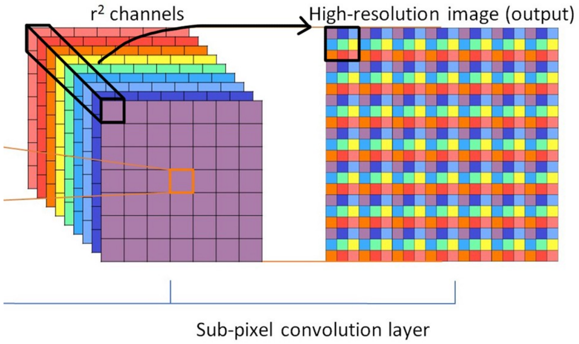

Figure 1: Pixel shuffle up sampling [7]

3 Proposed method

3.1 Preprocessing

In the first stage, we preprocess the data. For a faster training process, We train the model with fixed sized smaller pair of images. We input a random patch of size 32×32 from LR images to the model and get 128 × 128 SR images as output. To avoid over fitting we apply random rotation and flipping.

3.2 Model Architecture

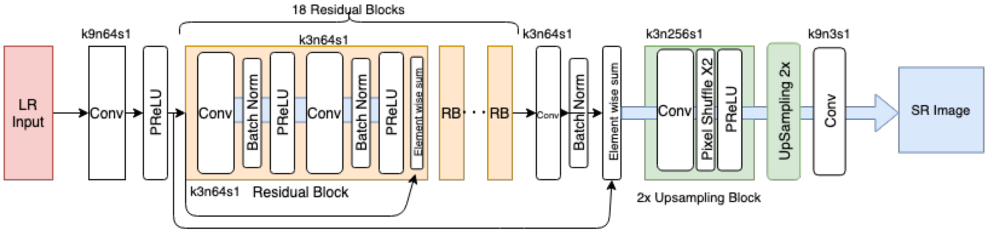

Figure 2: Architecture of Generator

Figure 2 shows the generator network architecture. We improve over SRGAN [5] by using 18 residual blocks in order to make the generator deeper by 2 blocks. We also made use of instance normalization layer instead of batch normalization in the residual blocks. Instance normalization is similar to layer normalization but goes one step further, It computes the mean/standard deviation and normalize across each channel in each training example. In recent years, instance normalization has also been used as a replacement for batch normalization in GANs [6]. But that didn't improve the results so we turned back to using batch normalization. Figure 3 shows the proposed discriminator architecture.

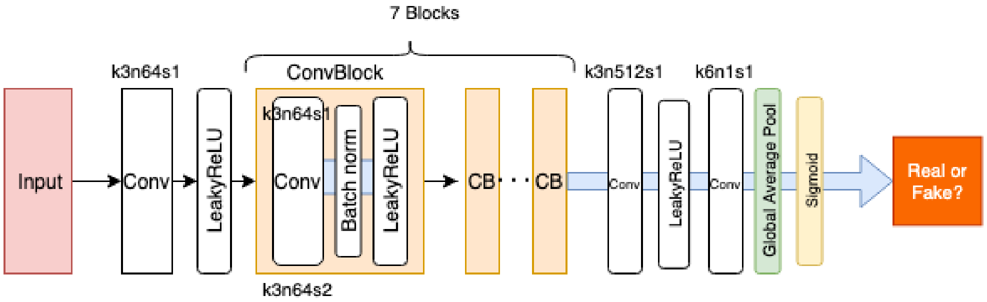

Figure 3: Architecture of Discriminator

The proposed model for generator has fully convolutional structure. The same network with fully connected classifier would have a large number of parameters which was harder to optimize. By using a FCN based architecture the number of parameters reduced to 4M compared to the 40M in SRGAN [5].

3.3 Loss Function

The loss function is comprised of three components including adversarial loss, content loss and DCT loss. equation 1 shows the full objective function to be minimized.

$$Loss = L_{content} + \beta L_{adv} + \gamma L_{DCT} \tag{1}$$

Where, in this setup, β and γ are equal to 1e-3 and 1e-3 respectively. The adversarial loss is defined as equation 2. intuitively, it makes the generator to produce better images in order to fool the discriminator. On the other hand at the same time forces discriminator to better distinguish between real and fake images.

$$L_{adv} = min_{\theta_G} max_{\theta_D} E_{I^H R}[\log D_{\theta_D}(I^H R)] + E_{I^L R}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^L R)))] \tag{2}$$

The content loss is measured by inputting both real and fake images to a pretrained VGG19 model and computing mean squared error on resulting feature maps of intermediate convolutional layer, in this case before 5th max-pooling layer. Finally, the equation for DCT loss is shown in equation 3. This kind of loss function helps to reconstruct high frequency details of the image easier by taking the discrete cosine transform of images and computing L^2 loss on resulting DCT coefficients.

$$L_{DCT} = ||\ DCT_{real} - DCT_{fake}\ ||_2 \tag{3}$$

4 Results

We train two models, one is exactly same as SRGAN[5] and second is the our improved model which was explained in detail before. To train the SR model we use the first 800 images of DIV2K dataset for image super resolution. As mentioned earlier this data goes through pre-processing stage and then goes trough the network. The next 100 images are for the validation set but since the maintainer has not granted access to the test set for the public, we use the validation part for as our test set. To further strengthen analysis of model performance, we use benchmark datasets such as Set5, BSD100, Urban100 for evaluation and report detailed results. Since we had limited resources to train our proposed model, in order to have a fair justification between performances of original SRGAN model and our improved one, we also train SRGAN model with same number of steps as our improved model.

Table 1: Performance of models on benchmark datasets. both SRGAN and improved models were trained for 50000 steps.

Dataset	Metrics / Model	SRGAN	Improved(ours)	SRGAN Paper [5]
Set5[8]	MSE	440	108	—
	SSIM	0.67	0.793	0.84
	PSNR	23.29	27.79	29.4
BSD100[9]	MSE	661	221.4	—
	SSIM	0.59	**0.742**	0.66
	PSNR	20.97	24.68	25.16
URBAN100[10]	MSE	1272	1087	—
	SSIM	0.58	0.62	—
	PSNR	18.07	17.66	—

Table 1 shows the performance of the two aforementioned models on evaluation datasets. from the table above we can see that the improved model performs better than implemented original model in terms of metrics on different evaluation datasets. It should be noted that due to the limitation in computational resources we were not able to fully train the two models as the researchers did in SRGAN. So there is a small gap between results of improved model and reported results of paper which we believe can be further improved to outperform paper results. So far, With this setup we were able to outperform SRGAN[5] results on BSD100 dataset. the SSIM of our proposed model was 0.742 compared to 0.66 reported in the SRGAN[5] paper. Since the test set for DIV2K was not available. we evaluate performance of the two models on validation part of DIV2K dataset. The results are listed in the table 2.

Table 2: Performance of models on DIV2K validation (100 images)

Metrics / Model	SRGAN	Improved(ours)
MSE	297.83	**210.55**
SSIM	0.71	**0.76**
PSNR	27.63	**29.50**

Figure 4 shows the 4x-up-scaled images produced by our proposed setup. The results are almost acceptable. Model has a good ability to enhance image details when up scaling the image. However, in some images with complicated pattern, the output is a little noisy. Which we believe is due to the DCT loss function.

5 Discussion

In this project, we implemented original SRGAN model and proposed an improved method for the task of SISR. As we mentioned earlier due to the limited resources we were not able to train the proposed model further. But since we trained original model and improved one for same steps, there is enough evidence that using fully convolutional structure for the discriminator not only reduces parameter count by a large factor, But also can further improve the convergence of the generator network. performing visual analysis on model outputs shows that adding DCT loss function to the perceptual loss increases model ability to reconstruct high frequency details. However, the some of generated SR images especially the ones with lots of patterns are noisy. This is a draw back of using DCT loss function and should be further investigated.

References

[1] M. Haris, G. Shakhnarovich, and N. Ukita, M.C. (2018) Task-driven super resolution: Object detection in low-resolution images, *Arxiv:1803.11316*

[2] Dong, C. L. (2014). Learning a Deep Convolutional Network for Image Super-Resolution. *Springer International Publishing*, 184-199.

[3] Dong, C. L. (2016). Accelerating the super-resolution convolutional neural network. . In European conference on computer vision (pp. 391-407). *Springer*.

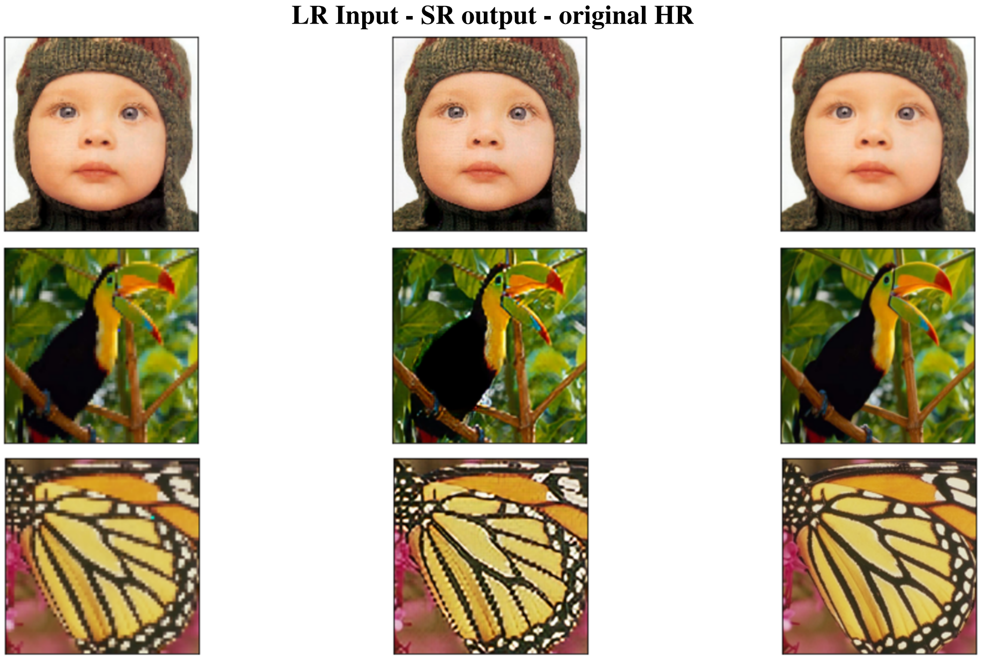

Figure 4: Visual comparison of the results. The Left column shows the model LR input. The center shows the model output(4x up scaled). The right column shows the original HR target.

[4] Johnson, J. A.-F. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision. Springer*, Cham, 694-711.

[5] Ledig, C. T. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4681-4690).

[6] Jun-Yan Zhu*, Taesung Park*, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", in IEEE International Conference on Computer Vision (ICCV), 2017.

[7] Shi, W. C. (2016). 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*, (pp. 1874-1883).

[8] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel.Low-complexity single-image super-resolution based on nonnegative neighbor embedding. BMVC, 2012

[9] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In Curves and Surfaces, pages 711–730.Springer, 2012

[10] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In IEEE International Conference on Computer Vision (ICCV), volume 2,pages 416–423, 2001