Optimizing SR-GAN for Resource-Efficient Single-Image Super-Resolution via Knowledge Distillation

Sajid Hussain[1,2], Jung-Hun Shin[1,2], Kum-Won Cho[1,2]
1. Korea Institute of Science and Technology Information.
2. University of Science & Technology,
Daejeon, South Korea

Sajid@kisti.re.kr, shandy77@kisti.re.kr, ckw@kisti.re.kr

Abstract

Generative Adversarial Networks (GANs) have facilitated substantial improvement in single-image super-resolution (SR) by enabling the generation of photo-realistic images. However, the high memory requirements of GAN-based SRs (mainly generators) lead to reduced performance and increased energy consumption, making it difficult to implement them onto resource-constricted devices. In this study, we propose an efficient and compressed architecture for the SR-GAN (generator) model using the model compression technique Knowledge Distillation. Our approach involves the transmission of knowledge from a heavy network to a lightweight one, which reduces the storage requirement of the model by 58% with also an increase in their performance. Experimental results on various benchmarks indicate that our proposed compressed model enhances performance with an increase in PSNR, SSIM, and image quality respectively for x4 super-resolution tasks.

1. INTRODUCTION

Generative Adversarial Networks (GANs) have proven to achieve success throughout several computer vision applications, along with the generation of realistic images, image-to-image translation, 3D object creation, and super-resolution [1]. These models comprise two neural networks, specifically the generator and the discriminator, which work together to produce realistic images. The generator network is responsible for synthesizing images, while the discriminator network is designed to classify how well an image is real or not.

Super-resolution is a fundamental problem in the domain of computer vision, and its significance has grown most recently due to the rising insistence on high-quality images in various industries such as medical imaging, satellite imaging, and surveillance. GAN-based super-resolution methods have achieved remarkable success in generating high-quality images, but GAN generators require significant amounts of memory and computing resources, making them difficult to deploy on resource-constrained platforms. This challenge has led to the development of various model compression techniques that aim to decrease the size of models while maintaining their performance.

One approach to model compression is knowledge distillation [2], which involves transferring knowledge from a larger and more complex model of the teacher to a smaller and more efficient model of the learner, resulting in similar performance but with significantly fewer parameters. Pruning is another technique that involves removing unimportant weights from a neural network to reduce its size while maintaining performance. Finally, network quantization is a compression technique that involves reducing the precision of network weights to reduce memory and computation requirements. These techniques enable the development of smaller, more efficient models that are better suited to resource-constrained environments.

In this paper, we present a potent strategy to reduce the size of the SR-GAN generator model [1] while maintaining its performance. Our approach uses knowledge distillation to train a smaller generator model. Specifically, we calculate several loss functions between the output of the smaller generator model, the teacher generator model, the discriminator model, and the ground truth images. Through this approach, we were able to compress the SR-GAN [1] generator model by 58% without sacrificing performance.

To assess the convincingness of our approach, we conducted extensive experiments on various benchmarks. The outcomes indicated substantial improvements in peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and image quality for x4 high-resolution tasks.

2. Methodology

2.1 Overview of the Proposed Approach

In our research, we present a compact architecture for enhancing the resolution of single images through the use of adversarial training and knowledge transfer. The proposed method has three models: generator, student generator, and discriminator. The generator acts as a mentor to the student generator, which is a more compact version of the model. We train our compressed model using knowledge distillation, adversarial training, and several loss functions that encourage high performance.

2.2 Architecture Details

We adopt the framework of SR-GAN [1] for our models.

A generator model is a deep neural network that molds a low-resolution image into a high-resolution one. It contains 16 residual blocks (N=16) and 2 up-sampling blocks with skip connections, which we used as a pre-trained teacher model. Student generator, which is similar to the generator model but has only 4 residual blocks (N=4). Discriminator models include convolutional layers, and fully connected layers. It takes both super-resolution and high-resolution images and generates a probability value indicating the authenticity of the image.

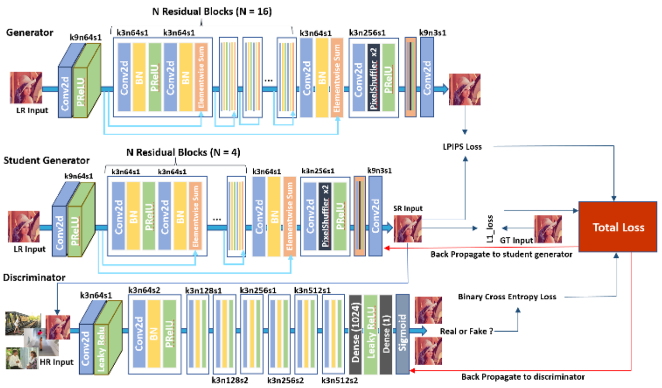


Figure 1. Student Generator (N=4) model training process.

2.3 Training of Student Generator

Our compact model is trained using a variety of loss functions, including learned perceptual image patch similarity (LPIPS) loss, content loss (L1 loss), perceptual loss, and regressive loss. The entire loss (1) for our model is calculated as follows:

$$L_{Generator} = L_{percep} + \delta_{lpips}\, L_{Distillation} \qquad (1)$$

Where δ_{lpips} is the coefficient to balance loss term.

We utilize adversarial training to enhance the quality of the images generated by our model by training it to deceive the discriminator model. The discriminator network D is optimized alongside G to address the adversarial min-max problem (2):

$$\min_{\theta_G}\max_{\theta_D} E_{I^{HR}\sim P_{train}(I^{HR})}\left[\log D_{\theta_D}(I^{HR})\right] + E_{I^{LR}\sim P_{G}(I^{LR})}\left[\log\left(1 - D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right)\right)\right] \qquad (2)$$

2.4 Perceptual Loss

Perceptual loss (3) is computed as a weighted aggregate of content loss and an adversarial loss component, which prompts the generator to create realistic images. Specifically, the perceptual loss is defined in the following manner:

$$L_{percep} = L_{content} + 10^{-3} L_{Gen}^{SR} \qquad (3)$$

2.4.1 Content Loss

To enforce the congruence between the generated images and the real images in terms of their underlying content, we use two loss functions for content loss.

We utilize the VGG-based MSE loss (4) as explicated in the SR-GAN paper [1], which computes the mean square error between feature representations of generated and ground truth images derived from a pre-trained VGG-19 network. This approach has been shown to improve the preservation of image structures and textures in the generated images.

$$L_{VGG/ij}^{SR} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \left(\varphi_{i,j}(I^{HR})_{x,y} - \varphi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}\right)^2 \qquad (4)$$

In addition, we also use L1_loss (5), which evaluates the 1-norm distance between the pixel values of the super resolute and real images. Compared to other loss functions, the L1 loss is more robust to outliers and tends to preserve sharp edges and details in the generated images.

$$L1 = N_{xi}\,|G_{xi} - GT| \qquad (5)$$

To combine the L1 loss and the VGG-based MSE loss, we map the content loss (6) as the weighted score of these two losses:

$$L_{content} = L_1 + \delta_c\, L_{VGG/i.j}^{SR} \qquad (6)$$

Where δ_c the weight-assigned VGG-based MSE loss.

2.4.2 Adversarial Loss

The Adversarial loss (7) component prompts the generator to produce real images that can deceive the discriminator. The generative loss based on the discriminating probabilities for all training samples is defined as follows:

$$L_{GEN}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right) \qquad (7)$$

Here, $D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right)$ is the possibility that the reconstructed image $G_{\theta_G}(I^{LR})$ is an original HR image.

2.5 Distillation Loss

In our proposed approach, we exploit the learned Perceptual image Patch Similarity (LPIPS) loss [3] as distillation loss (8). LPIPS loss is a perceptual metric that measures the perceptual similarity among images by comparing their feature representations in a deep neural network trained specifically for perceptual similarity. The LPIPS loss has been shown to better correlate with human perception than other commonly used perceptual metrics.

$$L_{lpips} = \sum_N M^k\left(\alpha^k(I^{Gen}) - \varphi^k(I^{GT})\right) \qquad (8)$$

where α is the feature extractor, *M* transforms the deep embedding into a scalar LPIPS score, and the rating is computed and averaged over *N* layers.

3. Experiments

Dataset For training our model, we utilized the DIV2K dataset, containing 900 2K-resolution images designed specifically for image restoration tasks. We randomly extracted 96 x 96 patches from the training set and generated the corresponding 24 x 24 input patches. To examine our model's performance, we ran experiments on four widely manipulated datasets: Set5, Set14, BSD100, and Urban100.

Training Details According to SR-GAN [4], All experiments are run at a scale aspect ratio of 4 between low-resolution (LR) and high-resolution (HR) images. The Student Generator is trained using Adversarial training and the Teacher Generator. We empirically set δ_{lpips} =1e-6, δ_c = 6e-3. The learning rate is set to begin with 2 x 10^{-4} and decreases by a factor of 0.1 every 2000 epoch. Adam is used for optimization with β = 0.9. We trained our model with the PyTorch framework on the NVIDIA TESLA V100 GPU.

4. Results

Table 1. provides an Overview of Teacher and Student network in terms of Residual Blocks, GFLOPS, and Model Size, where GFLOPs are calculated as processing a 256×256 image with a scale factor of 4. we achieved a substantial 58% reduction in model size compared to the teacher network.

Table 1. Overview of Teacher and Student network.

	Residual blocks	GFLOPS	Model Size
Teacher	16	56.58G	5.92MB
Student	4	5.29G	2.48MB

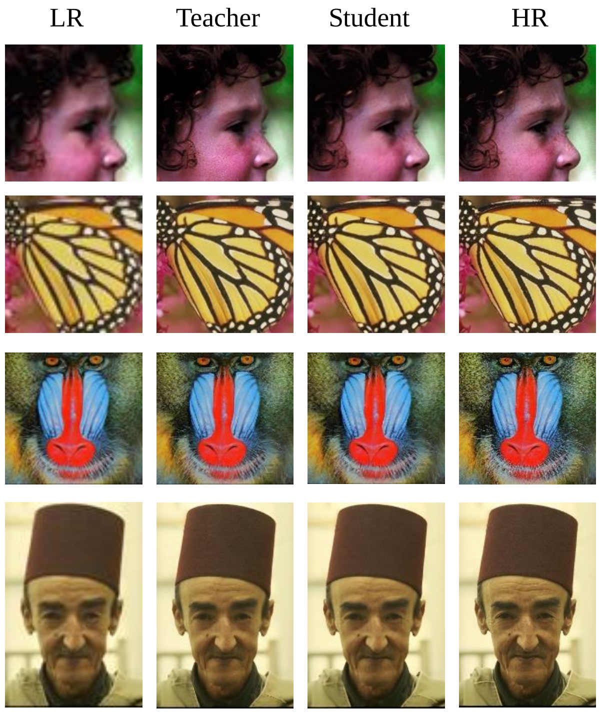
Our study evaluated the effectiveness of our knowledge distillation method in improving the performance of a student network using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics on multiple test datasets. The results are shown in Table 2. show that the student network outperformed the teacher network in the provision of PSNR and SSIM scores, with statistically significant differences highlighted in bold.

Table 2. PSNR and SSIM results.

Method	Set5	Set14	Urban100	BSDS100
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Teacher	29.93/0.869	26.95/0.745	**24.67/0.769**	26.10/0.695
Student	**30.41/0.891**	**27.25/0.770**	24.37/0.764	**26.39/0.720**

We also conducted a visual assessment to assess the performance of both networks, evaluating the quality of the upscaled images generated by each network across multiple test images. Our visual analysis demonstrated that the student network consistently generated high-quality images that closely resembled the output of the Teacher Network.

This improvement in perceptual quality underscores the effectiveness of our approach in enhancing the performance of the network.



5. Conclusion & Future Work

Overall, our study has demonstrated that utilizing the knowledge distillation method for SR-GAN training can produce a compressed generator without loss of performance. By doing so, we have reduced the storage requirement of the model by 58%.

Our future work will focus on exploring additional model compression techniques to further optimize the efficiency of the SR-GAN architecture. Additionally, we aim to replace the discriminator network in the SR-GAN with a new network that provides per-pixel assessment to the generator, which could lead to more visually pleasing and realistic results.

Acknowledgment

This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. CRC21011) and the National Supercomputing Center with supercomputing resources including technical support (TS-2023-RE-0001).

References

[1] Ledig Christian et al., Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition, (Honolulu, HI, USA), 2017, pp. 105-114.

[2] Hinton Geoffrey et al., Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[3] Richard Zhang et al., The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, (Salt Lake City, Utah), 2018, pp. 586-595.