Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 5, No. 2, 2023

Research on Image Super-Resolution Using Attention Mechanisms based on Super-Resolution Generative Adversarial Network

Zhouli Wu *

Software Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang, 310023, China
*** Corresponding author Email:** luckywzli@163.com

Abstract: With the continuous advancement of technology, Super-Resolution Generative Adversarial Networks (SRGAN) have played a significant role in the field of image super-resolution, significantly enhancing the resolution of images. However, while SRGAN excels in generating details, sometimes the restored details do not always meet people's expectations. To further enhance the quality of images and make image details clearer, this paper introduces improvements to the architecture and loss functions of the SRGAN network. Specifically, this research draws inspiration from the architecture of ESRGAN, removing the original Batch Normalization layers and introducing a newly designed Residual Block. Leveraging insights from attention mechanisms, we incorporate three layers of convolutional operations and introduce attention mechanisms into these new Residual Blocks. Furthermore, to simplify the computational complexity of the model, this paper simplifies the original loss functions, consolidating the previous four losses into two. These enhancements result in a significantly improved model in capturing visual elements, making key objects in the images more prominent compared to SRGAN. Detailed experimental results demonstrate that this model, while maintaining the clarity of details, provides higher visual quality. These achievements provide valuable insights and inspiration for further research and applications in the field of image super-resolution.

Keywords: SRGAN; Attention Mechanism; Residual Blocks.

1. Introduction

Super-resolution reconstruction technology (Super-resolution, SR) refers to the technique of increasing the resolution of an original image through hardware upgrades or improvements in software. It involves processing one or multiple blurred but similar low-resolution images (LR) using corresponding algorithms to reconstruct one or multiple clear high-resolution images (HR) [1]. Single Image Super-Resolution (SISR) is an important application area for this technology, and it has garnered significant attention from the research community and artificial intelligence companies, especially in the past decade with advancements in machine learning and deep learning.

SRGAN is a deep learning model that aims to convert low-resolution images into high-resolution ones through adversarial training. It consists of two components: a generator that maps low-resolution images to high-resolution ones and a discriminator that evaluates the authenticity of generated images, both engaged in a competitive training process. However, SRGAN has drawbacks, including complex training, a need for extensive data, susceptibility to artifacts, and image degradation, which require further refinement.

Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), is a deep learning model aimed at enhancing the quality of high-resolution image generation from low-resolution inputs. It utilizes a generator and discriminator network in adversarial training, striving to produce high-quality, realistic high-resolution images. However, like SRGAN, it encounters challenges such as complex training, data requirements, and potential image artifacts that require ongoing improvement.

The attention mechanism dynamically computes the importance of different regions within an image, enabling deep learning models to focus more on crucial areas to enhance task performance. Its typical process involves extracting features from the input image, then calculating attention weights for each feature location, and ultimately synthesizing weighted features to generate the output. However, the limited interpretability of why the model selects certain regions is a challenge.

Considering the aforementioned issues, the new model builds upon SRGAN with several improvements. In terms of loss functions, complexity is reduced by simplifying the original four loss functions into two. Regarding network architecture, this study draws inspiration from ESRGAN, eliminating the use of Batch Normalization layers and introducing a completely redesigned residual block to replace the previous one. Additionally, the new residual block incorporates an attention mechanism [2], which is integrated on top of the three convolutional layers.

Through these improvements, our enhanced model has shown significant improvements in capturing visual elements compared to SRGAN, making key objects in images stand out. Experimental results indicate that our model provides higher visual quality while maintaining detail clarity in image super-resolution tasks. These achievements serve as important references and inspirations for further research and applications in the field of image super-resolution.

2. Related Work

In 2014, Chao Dong, Chen Change Loy, and Xiaoou Tang introduced the Super-Resolution Convolutional Neural Network (SRCNN) model [3], which not only improved image quality but also inspired subsequent research. However, it still had drawbacks such as fixed upscaling factors and high computational complexity.

Figure 1. The super-resolution results of ×4 for SRGAN, the proposed model. This new model outperforms SRGAN in sharpness and details

In 2016, Shi et al. proposed the Efficient Sub-Pixel CNN (ESPCN) model [4], which is based on pixel rearrangement. It extracts feature maps by performing convolution operations on low-resolution images and then feeds these feature maps into sub-pixel convolutional layers to generate high-resolution images by rearranging and combining channels. Although it reduced computational complexity, it did not address the fixed super-resolution factor problem.

Therefore, in 2017, SRGAN (Super-Resolution Generative Adversarial Network) was the first attempt by Ledig et al. to use Generative Adversarial Networks (GANs) in the field of image super-resolution [5]. This model consists of two opposing modules: a generator that synthesizes high-resolution images by adding random noise to the original image, and a discriminator that distinguishes whether an input image is generated by the generator or a real image. However, there were still issues with unstable training and difficulty in preserving details.

To address these issues, in 2018, the ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) model made an important attempt to introduce GANs into the image super-resolution field [6]. This model has a similar overall structure to SRGAN but introduces a new Residual-in-Residual Dense Block (RRDB) network unit, which removes the Batch Normalization (BN) layer and uses Group Normalization to replace BN. Additionally, it improves the GAN network by using Relativistic average GAN (RaGAN). These enhancements not only improved performance but also increased the model's generalization ability, achieving excellent results in image detail and texture. However, there were still challenges related to capturing fine details and slightly lower resolution.

3. Proposed Methods

The main objective of our study is to further enhance the visual quality of images based on SRGAN. In this section, we first describe the network architecture of the new model and then introduce some modifications to the loss functions.

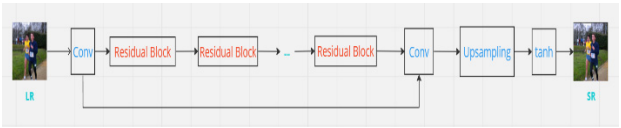

Figure 2. This study adopts the basic model of SRGAN and improves upon it by modifying the residual blocks, thereby achieving enhanced performance

3.1. Network Architecture

Unlike SRGAN, the novel model proposed in this study introduces several improvements to the residual blocks. It not only removes the original BN layers but also transforms the original two convolutional layers into three. Additionally, it incorporates an attention mechanism after the convolution, replacing the old residual blocks with these new ones. Finally, Group Normalization was employed to replace BN. The specific details are illustrated in Figure 3.

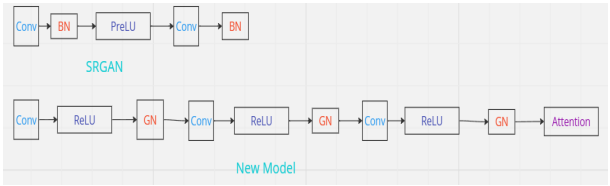

Figure 3. On the left is the original residual block of SRGAN, while on the right is the improved residual block.

While Batch Normalization (BN) layers offer several advantages in deep learning, such as accelerating training convergence [7], mitigating gradient vanishing issues, and enhancing model generalization, recent research has uncovered some limitations in specific scenarios. Firstly, BN layers require batch data mean and variance for feature normalization during both training and inference. This implies that during the testing phase, mean and variance calculations need to be performed over the entire training dataset. Moreover, if there is a significant distribution difference between the test data and the training data, it can lead to a decrease in performance, a problem known as the "batch effect." Secondly, BN layers introduce additional learnable parameters (estimates of mean and variance), increasing the model's storage and computational complexity. Particularly in deep networks, this can result in artifacts in the generated images.

Therefore, this study replaced BN layers with Group Normalization (GN) layers while building upon the SRGAN

model [8]. GN layers partition features into several groups and normalize each group's features, instead of normalizing the entire batch. This approach offers several advantages: it is more robust to small batch data with unstable distributions and is less susceptible to batch effects. GN layers do not introduce additional learnable parameters, leading to smaller computational and storage overheads, making them especially suitable for deep networks. In this experiment, a multi-level residual network was constructed by applying normalization through three layers of convolution and GN layers, thereby improving image quality. Specific details will be discussed in Section 4.

3.2. Loss Function Design

In terms of the specific structure of the loss function, this work has not made significant changes. The computation formula for the loss function remains as in

$$g\_loss = l1\_loss + 0.001 * adversarial\_loss \quad (1)$$

However, in our new model, this paper has adopted a relatively simpler calculation approach. Specifically, the loss function consists of two main components: Adversarial Loss and L1 Loss.

Adversarial Loss is computed as torch.mean(1 - out_labels). Its purpose is to encourage the generated images (out_images) to be classified as real images by the discriminator (or discriminator), meaning that the probability should be close to 1. L1 Loss is computed using nn.L1Loss() and is employed to measure the pixel-level differences between the generated images and the target images.

The initial version of the loss function is relatively straightforward, including only Adversarial Loss and L1 Loss. This simplicity enhances the stability and efficiency of the training process. The reason for opting for this straightforward structure is that L1 Loss helps generate images that closely resemble the target images at the pixel level. Furthermore, due to the modifications in the architecture of the generator, the original model has become relatively complex. To avoid introducing training instability and to make the model more amenable to optimization and debugging, this paper has chosen to follow a more classical approach.

Overall, this design of the loss function, although simple, serves the purpose of encouraging high-quality image generation, and its simplicity aids in training stability, optimization, and debugging.

4. Experiments

4.1. Training Details

This experiment, which draws inspiration from SRGAN, was configured with the following specific parameters: a scaling factor of 4 between LR (Low-Resolution) and HR (High-Resolution) images, a batch size of 16, and HR image blocks cropped to a spatial size of 128 × 128 pixels.

The experiment consisted of 100 iterations, and its overall process can be summarized as follows:

4.1.1. For Initialization: Initially, instances of the generator (netG) and discriminator (netD) were created.

4.1.2. Loss and Optimization Setup: The generator's loss function (generator_criterion) and optimizers (optimizerG and optimizerD) were defined. These components were used for training the generator and discriminator.

4.1.3. Data Storage: An empty dictionary named "results" was established. This dictionary was employed to store various metrics during the training process, including losses, scores, PSNR, SSIM.

4.1.4. Training Loop: The main training loop began. In each iteration, the following steps were executed:

Training Mode: *The generator and discriminator were set to training mode.*

Data Loading: *The code iterated through the training dataset, loading batches of image data.*

Discriminator Update: *For each batch, the discriminator (D) network was updated. This process involved computing the discriminator outputs for both real and generated images and calculating the discriminator loss. Subsequently, backpropagation and optimization were performed.*

Generator Update: *The generator (G) network was then updated. This step included computing the discriminator output for the generated images and calculating the generator loss. Again, backpropagation and optimization were carried out.*

Loss and Score Updates: *The losses and scores for the current batch were computed and updated.*

4.1.5. Evaluation Mode: After each iteration, the generator was set to evaluation mode, and the model's performance was assessed using a validation dataset. Metrics such as PSNR and SSIM were computed to evaluate the model's quality.

4.1.6. Results and Model Saving: The model parameters, training results, and evaluation metrics were saved.

4.1.7. Periodic Metric Recording: At regular intervals during training, the experiment recorded metrics such as losses, scores, PSNR, and SSIM into a CSV file. This data was intended for subsequent analysis and visualization

4.2. Data

In the training phase, the dataset used is the DIV2K dataset [9]. DIV2K (Diverse 2K Resolution High-Quality Images) is a commonly used dataset for super-resolution image processing tasks. It consists of 800 diverse high-quality images, typically at a resolution of 2K (usually referring to 2048x1080 pixels), and is used for training and evaluation of super-resolution, image restoration, and other computer vision tasks. In the testing phase, this experiment employed several datasets, including Set5, Set14, BSD100[10], and Urban100, for evaluation.

4.3. Results

This experiment compares the new model with the original SRGAN. Here are the relevant data and performance comparisons.

Table 1. Relevant data

MODEL	Dataset (PSNR/SSIM)			
	Set5	**Set14**	**BSD100**	**Urban100**
SRGAN	28.63/ 0.83	25.49/ 0.73	25.56/ 0.69	25.45/ 0.72
New model	26.73/ 0.84	23.57/ 0.73	23.70/ 0.69	21.33/ 0.72

From this table, it can be observed that the new model exhibits a slight decrease in PSNR compared to SRGAN, but the SSIM values are close to being the same. This phenomenon reflects the distinct performance characteristics of the two models in image reconstruction tasks. Due to the introduction of attention mechanisms, the new model may not particularly emphasize pixel changes in certain areas of an image. However, from a perceptual perspective, the primary concern is the perceived image quality. Therefore, the new model aligns more closely with human perceptual standards.

Figure 4. The comparison between the new model and SRGAN, with SRGAN-restored images on the left and images restored by the new model on the right

From the above images, it can be observed that, compared to SRGAN, even though the new model has a lower PSNR value than SRGAN, it exhibits improvements in both artifact removal and facial image quality. The new model doesn't introduce unnecessary textures, such as wrinkles on the face.

5. Conclusion

Comparison between the new model and SRGAN, with SRGAN-restored images on the left and new model-restored images on the right. In summary, the new model has made changes in the design of the residual blocks and loss functions, and has additionally introduced an attention mechanism, achieving certain improvements in the quality of face restoration and artifact removal. Firstly, by altering the design of the residual blocks, the new model may enhance the model's feature extraction capability and information propagation efficiency. These changes likely contribute to capturing richer image features, especially in the context of face images, potentially resulting in better image quality and higher perceptual quality. Due to the increased complexity of the residual blocks, this study decided to simplify the design of the loss functions to reduce model complexity. This decision helps balance model performance and computational cost, ensuring efficiency during both training and inference.

References

[1] Han, W., Dong, X., Dong, W. (2021). Application research of neural network models in the field of image super-resolution. Journal of Information Engineering University, 22(02), 159-163.

[2] Zhang, Y., Li, K., Li, K., et al. (2018). Image super-resolution using very deep residual channel attention networks. https://arxiv.org/abs/1807.02758.

[3] Dong, C., Loy, C. C., He, K., et al. (2015). Image super-resolution using deep convolutional networks. https://arxiv.org/ abs/1501.00092.

[4] Shi, W., Caballero, J., Huszár, F., et al. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. https://arxiv. org/abs/ 1609. 05158.

[5] Ledig, C., Theis, L., Huszar, F., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. https://arxiv.org/abs/1609.04802.

[6] Wang, X., Yu, K., Wu, S., et al. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. https://arxiv.org/abs/ 1809.00219.

[7] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://arxiv.org/abs/1502.03167.

[8] Wu, Y., He, K. (2018). Group normalization. https://arxiv. org/ abs/ 1803.08494.

[9] Timofte, R., Agustsson, E., Gool, L.V., et al.. (2017). NTIRE 2017 challenge on single image super-resolution: methods and results. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1110-1121.

[10] D. Martin, C. Fowlkes, D. Tal, et al.. (2002). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings Eighth IEEE International Conference on Computer Vision (ICCV 2001), Vancouver, Canada, pp. 416-423.