Analysis SRCNN & ESRGAN for Image Super Resolution

Changlin Su
University of Toronto
sheldon.su@mail.utoronto.ca

Guanglei Zhu
University of Toronto
guanglei.zhu@mail.utoronto.ca

Xinyi Liu
University of Toronto
helenxinyi.liu@mail.utoronto.ca

Abstract

In this project, we are interested in the application of deep neural networks in image super-resolution. We re-implement one discriminative method (SRCNN [1]) and one generative method (ESRGAN [2]). Moreover, We apply these two existing algorithms to several new image data sets and perform in-depth analysis both quantitatively and qualitatively.

1 Introduction

Single image super-resolution is a problem in computer vision that attracted a lot of attention in recent years. SR problem is hard to solve because it is underdetermined. For any given image at a lower resolution, the solution might not be unique and depends heavily on the content of the image. In recent years, various methods were developed to address this problem. In this paper, we will discuss the performance of two models that are very different, SRCNN and ESRGAN. The SRCNN model was introduced in 2015 which uses two convolution layers to recover a low-resolution image. On the other hand, ESRGAN was introduced in 2018 and uses the GAN architecture. ESRGAN is a much larger model than SRCNN, it has over 240 convolutional layers in our implementation. We will re-implement the two models discussed above and compare their performance on several different data sets and compare their performance. Due to limited computing resources, our implementation of ESRGAN hasn't reached training convergence, hence we will also include a pre-trained ESRGAN when comparing the performances of a different model.

2 Related Work

Deep neural network techniques have mostly overcome the image super-resolution challenge in recent years. Dong et al. [1] propose SRCNN as a pioneering work that learns the mapping from low resolution to high-resolution images in an end-to-end way, outperforming earlier research. Since then, new network topologies have emerged, including the deeper network with residual learning [3], and the Laplacian pyramid structure [4].

Using adversarial training, GAN [5] is used to increase photo-realism. Recently, a number of papers have been published that focus on the development of more effective GAN frameworks. SRGAN [7] is a single image super-resolution generative adversarial network. It employs a perceptual loss function that combines adversarial and content losses. Using a discriminator network that is trained to differentiate between super-resolved images and original photo-realistic images, the adversarial loss drives the solution to the natural image manifold. ESRGAN enhanced SRGAN by employing a more effective relativistic average GAN.

GitHub link: https://github.com/Sheldonsu28/CSC413-Project

3 Models

3.1 SRCNN

SRCNN is one of the first deep learning methods for single image super-resolution, it utilizes 2 convolutional layers to achieve relatively good performance on single image super-resolution tasks. The SRCNN model we used in our work consists of three convolutional layers with two activation layers.

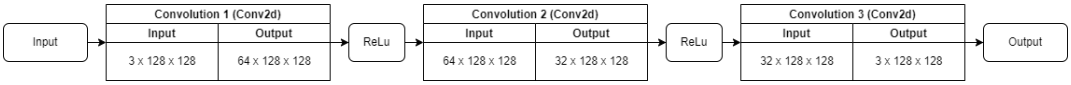

Figure 1: An algorithm box of SRCNN

The first convolutional layer of the SRCNN extracts a set of feature maps from a low-resolution image. The second layer performs a nonlinear conversion of these feature maps to high-resolution patch representations. The final layer, which generates the final high-resolution image, combines the predictions into a spatial neighbourhood.

3.2 ESRGAN

ESRGAN is one of the recent approaches for solving single image super-resolution. It consists of two different networks: the Generator and a Discriminator. The generator is responsible for receiving a low-resolution image and trying to recover a high-resolution version of it while the discriminator judges the quality of the output of the image generated by the generator.

The structure of ESRGAN differs from that of SRGAN's generator G in two ways: 1) Remove all BN layers, as shown in Figure 2; 2) Replace the original basic block with the proposed Residual-in-Residual Dense Block (RRDB), which combines a multi-level residual network with dense connections as shown in Figure 3.

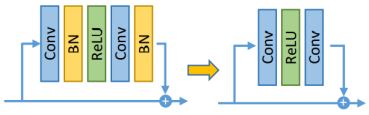

Figure 2: Removal of BN layers within residual blocks [2]

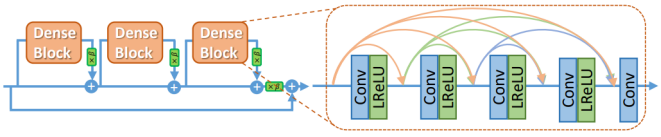

Figure 3: Proposed Residual-in-Residual Dense Block (RRDB) [2]

The Discriminator used in ESRGAN also differs from that of SR's discriminator. Instead of using a conventional discriminator, it uses a relativistic discriminator for training. The structure of the discriminator remains the same as what SR models have, but instead of classifying the image as a fake of real image, it tries to predict the probability of an image being more real than the other image or not based on the following loss function:

$$L_D = -\mathbf{E}_{x_r}[\log(D_{R_a}(x_r, x_f))] - \mathbf{E}_{x_f}[\log(1 - D_{r_a}(x_f, x_r)]$$

4 Experiments

4.1 Training Details

In this experiment, both models are trained with the DIV2K [8] train data set. We centre crop the images in the data set to obtain the 128×128 ground truth high-resolution image. Then we use the transform function provided in PyTorch to shrink the image down to 64×64 as low-resolution input for training. Our implementation of ESRGAN uses a shallower architecture proposed by Wang's team which only contains 16 residual blocks instead of the 23 blocks in the original architecture to reduce model complexity. We also use different loss functions when training the models, we use MSE loss for SRCNN while ESRGAN uses the original loss functions proposed in the paper. In our experiment, we trained ESRGAN with $\lambda = 0.001$, $\eta = 1.0$ and residual scaling parameter $\beta = 0.2$. During training, both models' learning rate was set to 1×10^{-4}. For optimization, we uses Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To accelerate our training, we trained both models with mixed precision. Unfortunately, our ESRGAN was only trained for 6000 iterations due to limited computational resources, hence it might not have reached its full capabilities. On the other hand, our SRCNN model was trained for 400 iteration and converges at around 300 iterations.

4.2 Datasets

As mentioned above, we use cropped DIV2K for training and an additional set of 100 images for testing. Flickr-Faces-HQ Dataset (FFHQ) [10] is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN). The dataset consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity and image background. We choose a subset of 1000 images in thumbnail resolution (128×128) for testing. Selfie2Anime [9] is a dataset of female anime face images transformed from real female face images. The test set consists of 100 face images of size 256×256.

4.3 Quantitative Results

In this section, we demonstrate the quantitative results using two popular metrics, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), on DIV2K (we use a 256 $\times$ 256 crop for testing to reduce computational cost), FFHQ and Selfie2Anime datasets. We have two versions of ESRGAN, one trained by us for 6k iterations and one pre-trained for about 50k iterations. Our ESRGAN performs the best across all three datasets under the PSNR metric with SRCNN performing very close. Since SRCNN is PSNR-oriented [1], using MSE as the loss function often favours a higher PSNR score. However, PSNR does not indicate visual quality, since PSNR metric fundamentally disagrees with the subjective evaluation of human observers [12]. When using SSIM, pre-trained ESRGAN performs the best across all datasets. It shows that ESRGAN is better at preserving the structure of reconstructed images than SRCNN.

Models	DIV2k	FFHQ	Anime
SRCNN	29.1	29.2	27.9
ESRGAN	**29.5**	**30.2**	**30.3**
ESRGAN (pre-train)	26.5	26.4	23.6

Table 1: Average PSNR on various datasets

Models	DIV2k	FFHQ	Anime
SRCNN	86.7	90.7	90.9
ESRGAN	86.9	91.4	93.3
ESRGAN (pre-train)	**95.9**	**96.9**	**97.1**

Table 2: Average SSIM on various datasets

4.4 Qualitative Results

We compare our final models: SRCNN, ESRGAN and pre-trained ESRGAN on several public benchmark datasets: DIV2K[8], selfie2anime [9] and FFHQ [10].

Figure 4: Image from DIV2K test data set [8]

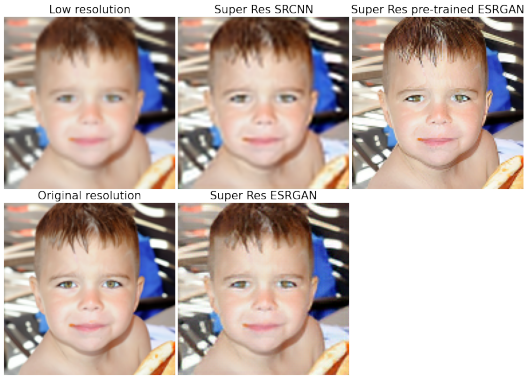

Figure 5: Image from FFHQ data set [10]

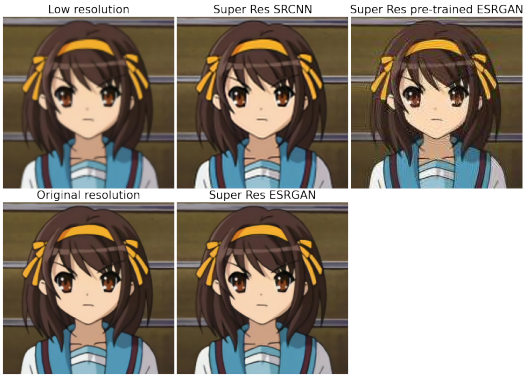

Figure 6: Image from selfie2anime data set [9]

For qualitative analysis, we choose three images to represent natural, human faces and anime images. Pre-trained ESRGAN outperforms SRCNN for both natural and human face images, in terms of sharpness and details. For example, in Figure 5, pre-trained ESRGAN produces a sharper baby face with more natural hair. Since the pre-trained model also uses Flickr2K dataset [11] during training, it certainly benefits from it when predicting human faces. Notice that, due to insufficient training, our ESRGAN suffers from unnatural noise in the image.

As for SRCNN, the output images are often over-smoothed and lack high-frequency details. Because of this, SRCNN performs well on anime images which often contain less detail than real-world images. This is also the reason why our insufficient-trained ESRGAN outperforms the pre-trained ESRGAN on anime images.

5 Conclusion

In general, we observed that the pre-trained ESRGAN generalises well across a wide range of real-world image classes and produces the highest overall quality results. On the selfie2anime data set, however, the insufficiently trained ESRGAN and SRCNN outperform the pre-trained ESRGAN.

SRCNN can recover some detailed texture of the low-resolution image. In addition, ESRGAN produces sharper edges compared to images generated by SRCNN. However, since the pre-trained model for ESRGAN was trained on real-world image data sets, it does not perform as well as SRCNN visually on anime. Anime in general tends to have less texture compared to real-world images and faces, hence favours the undertrained ESRGAN and SRCNN. Furthermore, even in the anime data set, under-trained ESRGAN beat the SRCNN with sharper edges and smoother colour patches.

References

[1] Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV. (2014)

[2] Xintao W., Ke Y., Shixiang W., Jinjin G., Yihao L., Chao D., Yu Q., and Loy, C.C.. Esrgan: Enhanced super-resolution generative adversarial networks. In ECCV. (2018)

[3] Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR. (2016)

[4] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR. (2017)

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014)

[6] Yang, C.Y., Ma, C., Yang, M.H.: Single-image super-resolution: A benchmark. In: *European Conference on Computer Vision*, pp. 372–386 (2014)

[7] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image superresolution using a generative adversarial network. In: CVPR. (2017)

[8] Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW. (2017)

[9] Junho K., Minjae K., Hyeonwoo K., & Kwang H. L.: U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In ICLR. (2020)

[10] T. Karras, S. Laine, and T. Aila.: A style-based generator architecture for generative adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410. (2019)

[11] Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L., Lim, B., Son,S., Kim, H., Nah, S., Lee, K.M., et al.: Ntire 2017 challenge on single imagesuper-resolution: Methods and results. In: CVPRW. (2017)

[12] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image superresolution using a generative adversarial network. In: CVPR. (2017)

A Appendix

A.1 Contribution

Changlin Su: Implemented utility functions for data pre-processing, implemented and trained both SRCNN and ESRGAN. Responsible for introduction, ESRGAN, training details and conclusion section in the report.

Guanglei Zhu: Implemented code for testing DIV2K, FFHQ and Selfie2Anime datasets using our SRCNN, ESRGAN and pre-trained ESRGAN models. Produced tables for quantitative analysis (Table 1, 2). Responsible for Abstract, Datasets, Quantitative and part of Qualitative Results in the report.

Xinyi Liu: Collected test datasets, implemented code to predict these datasets using trained models (including our SRCNN, ESRGAN and pre-trained ESRGAN), and produced images for qualitative analysis (Figure 4, 5, 6). Responsible for Related Work, SRCNN, part of Qualitative Results and References sections in the report. Merged, formatted, and cleaned all of the files and codes at the end.

All authors reviewed the final report.