Image super-resolution using generative adversarial network

Adithya Chowdary Boppana
Indiana University
aboppana@iu.edu

Satyaraja Dasara
Indiana University
sdasara@iu.edu

Siva Charan Mangavalli
Indiana University
simang@iu.edu

Abstract

Image super-resolution (SR) can be defined as the process of recovering a high resolution (HR) image from a low resolution (LR) image. Traditional methods include upscaling and interpolation techniques, however, recent advances in the field of deep learning have led to novel approaches. Traditional techniques lack finer texture details due to excessive smoothing of the image. Standard metrics to evaluate performance of super-resolution models such as PSNR and SSIM don't correlate well with the perceptual quality which is judged with mean opinion scores.

This issue can be handled in a better fashion by using the generative adversarial network (GAN) architecture. The discriminator can act like a person classifying whether the image is natural looking or not. GANs combined with appropriate loss functions restore more details and output more realistic images compared to other super-resolution techniques. Using the mean squared loss of the feature representations of intermediate layers from pre-trained models like VGG19 can aid in generation of perceptually good reconstructions compared to the traditional Mean Squared Error. which can result in averaging of pixels leading to blurred reconstructions. Incorporating the adversarial loss based on cross entropy of the discriminator probabilities to train our generator helps in pushing the reconstructions into the natural image manifold.

Our results show a model which can reconstruct perceptually good quality images of 4x super resolution from low resolution images.

1. Introduction

There have been a lot of advances in the field of image resolution enhancement over the years. However, the issues with the traditional algorithms is that they are not generalizable for images of different sizes, aspect ratios, scales and noise. A deep learning-based approach can provide us with a more general solution. The main issue with image resolution enhancement lies with the upscaling approach where we don't know how to fill the additional pixels without losing the original information present in the image. This is the main drawback of interpolation techniques as they leave a lot to be desired in terms of visual quality. Deep learning based approaches like SR-CNN and SR-ResNet [2] have shown good results, however, the metric commonly used i.e PSNR (Peak Signal to Noise Ratio) is considered to be a not so reliable metric for estimating resolution quality and those architectures may be suboptimal for the resolution task as they are more suited towards classification and object detection tasks. This project seeks to explore a new deep learning-based approach based on Generative Adversarial Network architecture using perceptual loss function for super resolution of images. This is an implementation project of the research paper, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”.

The traditional metrics to evaluate the performance for super-resolution tasks are PSNR and SSIM. However, these are full reference-based metrics which require a pair of images to evaluate the performance. This is useful for super-resolution of medical images and images requiring retention of structure of objects in it like text characters after upscaling. However, this logic is not necessarily true if the goal is to generate perceptually good-looking images. This can be attributed to the fact that the perceptual quality of an image is independent of its ground truth. Since most no-reference quality measures can't be generalized, the best approach is to have a discriminator differentiate between the images as real or fake i.e. natural or unnatural looking images in our case. The discriminator is like a human being quantifying the naturalness of the images.

The super-resolution generative network (SRGAN) as described in Ledig[1] has been implemented which contains residual blocks with skip connections. Our implementation includes a minor change of replacing the convolution and pixel-shuffler layers after the residual blocks with a transposed convolution layer.

2. Previous Work

The most well-known implementation of GAN for image super-resolution is by Christian Ledig [1]. A custom perceptual loss function was used which was a weighted sum of content loss and adversarial loss to train the generator network. Mean opinion score was used as the metric for evaluating the methods and GAN based approach was shown to be better at reconstructing more photo-realistic images.

The loss functions for super-resolution applications have been explored in Justin Johnson [6]. The two loss functions explained in detail are Feature Reconstruction loss and Style Reconstruction loss. The experiment results showed that though the feature reconstruction loss function resulted in a lower PSNR value, it increased the visual performance of the models by outputting more natural looking higher resolution images.

The SR-CNN method presented in Jiwon Kim [2] used a CNN for end to end learning for the image reconstruction. It used residual learning and high learning rates for the deep networks. It was proposed that this single model could be used for multiple scale factors for SR.

3.1. Implementation

Given a Low Resolution (LR) image our aim is to generate the corresponding super resolution (SR) image. The HR images are obtained by randomly cropping, flipping, adding gaussian noise and changing the brightness of the images in the data. The corresponding LR images were obtained by bicubic interpolation. The HR images are present during the training, during inference we only have the LR images.

The main goal is to find a Generator function that takes the LR image as input and produces the SR/HR image. To find such generator image, a discriminator function is trained to identify if the generated image belongs to the original HR image distribution. The objective is to train the generator so that the discriminator no longer classifies whether it is a real or a fake image.

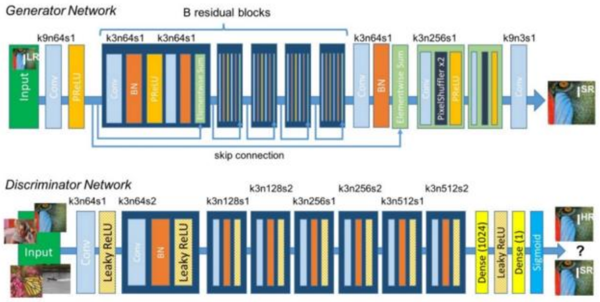
The model is built and trained using Keras framework. We deployed the code on Google Cloud Instance with NVIDIA TESLA P100 GPU and 32gb RAM configuration. A batch size of 8 images is used to prevent GPU memory allocation error.

3.2. Architecture

The Generator has 16 residual blocks with skip connections followed by a two up-sampling blocks to get 4x SR. The weights in the convolution layers are initialized with He Normal. Batch normalization is applied at each layer to increase the speed of training. Furthermore, Parametric ReLU activation is used at the end of all the residual layers.

The discriminator has stacked convolutional layers with batch normalization and Leaky ReLU with alpha=0.25 applied at end of them. There are two dense layers with outputs as 1024 and 1 at the end. Sigmoid activation is used to get the probability of real or fake for each image.

We have implemented the same architecture proposed by Ledig[1], except for the up sampling block in the Generator. Ledig[1] have used a pixel shuffler layer; we instead have used a transposed convolution for the same operation.



Architecture proposed in Ledig[1]

3.3. Loss Function

A perceptual loss function is proposed by Ledig[1] for SR. It has two main components, content loss and adversarial loss.

$$l^{SR} = \underbrace{\underbrace{l^{SR}_{X}}_{\text{content loss}} + \underbrace{10^{-3} l^{SR}_{Gen}}_{\text{adversarial loss}}}_{\text{perceptual loss (for VGG based content losses)}}$$

Loss proposed in Ledig[1]

The content loss ensures that we retain perceptual quality instead of pixel to pixel similarity. Using the feature maps from a convolutional layer prior to activation can suffice this. The hypothesis is that it would help extract photo realistic textures from the LR image. The loss is just the Euclidian distance of the reconstructed images feature maps and reference image feature maps.

$$l^{SR}_{VGG/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$$

Content Loss proposed in Ledig[1]

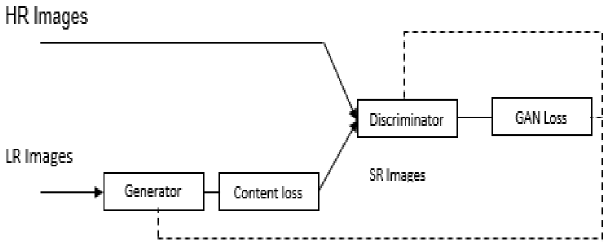
The adversarial loss is binary cross entropy which ensures that the discriminator network distinguishes between the real and generated images.

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR}))$$

Adversarial Loss proposed in Ledig[1]

3.4. Training steps

- Read the RGB images and convert them to Numpy arrays. Normalize them to [-1,1]
- Define the architecture for the generator and the discriminator in a class object
- Create a class object for the generator and discriminator initializing it with an input shape which returns the Keras model when its method is called
- Declare the custom loss function by importing pre trained VGG-19 model and freezing the weights of all layers and loss function
- Define the GAN network by feeding the output of the generator to the discriminator and add the corresponding losses with their weights
- First step is to generate the SR images using generator and feed it into the discriminator with a low probability labels and original HR images with high probability labels
- Freeze the discriminator weights and then feed the LR images to the GAN network with high probability labels. This trains the generator to generate images that are like HR images
- Repeat the training process in batches for all the data
- During inference the saved generator model can be imported, and the weights can be loaded. The model can generate the SR image for any given LR image



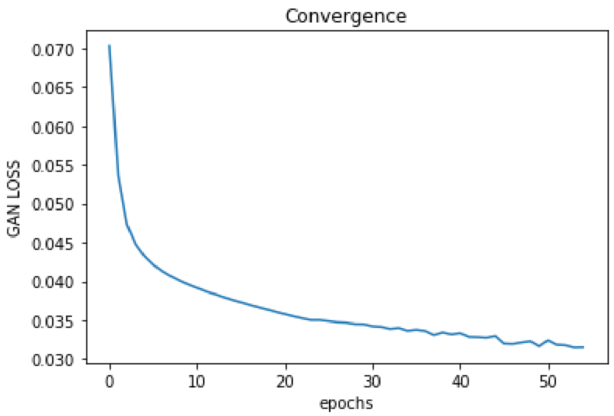
Training Process

3.5. Data

We have used DIV 2k [7] bicubic data-set for the training. For inference we have also explored the other data sets provided by DIV 2k. These include mild, difficult and wild LR images in the order of difficulty, these datasets are classified as hard level difficulty for SR process. After augmentation of the DIV 2k Bicubic dataset, we had a total of 4400 training, 100 validation and 1000 test images.

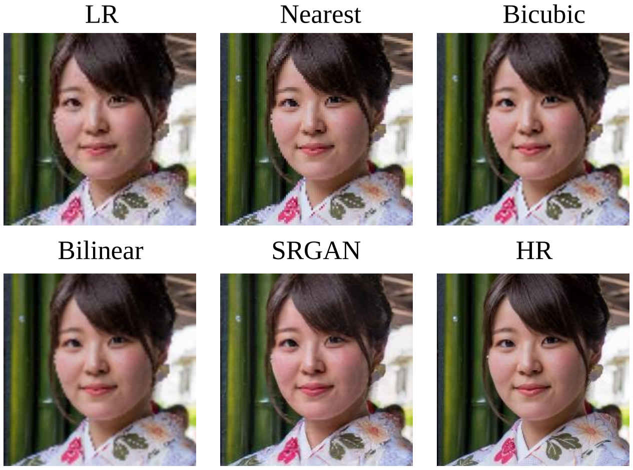
4.1. Results

We have trained the model for 55 epochs for 19 hours and the loss convergence graph below.

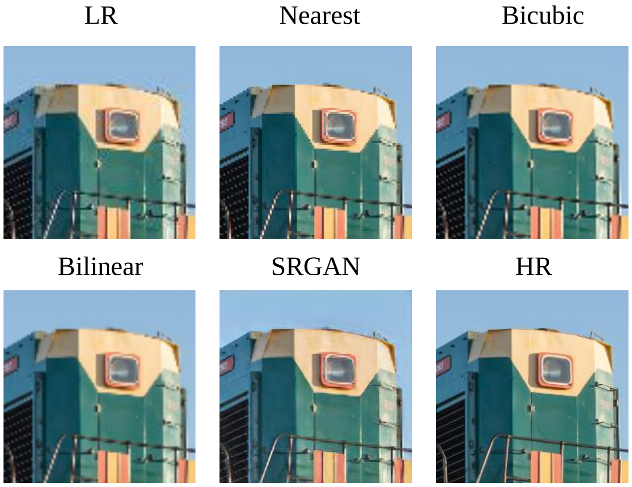


Convergence plot

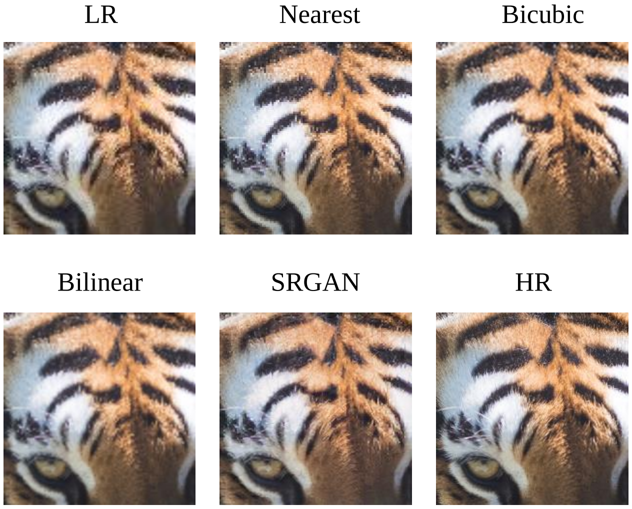
After 30 epochs the model tends to saturate in terms of training loss. The faster drop from initial loss can be attributed to batch normalization of the data and He normal initialization of the weights. We were unable to track the validation loss, due to out of memory issue.



Example – 1



Example – 2



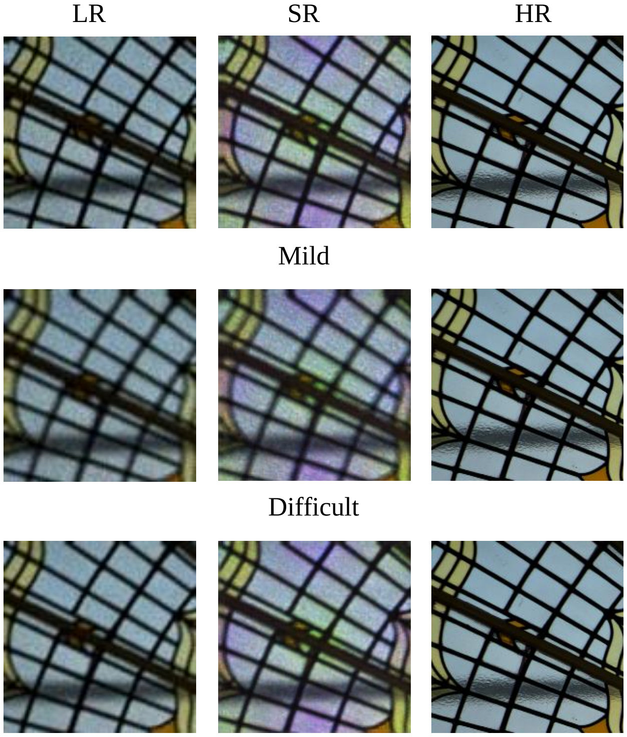
Example – 3

Example -1, Example -2 and Example – 3 are from theDiv2K Bicubic test data set which have been downscaled using bicubic interpolation. The plots show the corresponding SR images generated by using traditional and SRGAN. We compare it with the original HR image.

The SR image appears better in terms of perceptual quality compared to the traditional SR techniques like Nearest, Bicubic and Bilinear interpolation. The edges are preserved and the SRGAN output doesn't have pixelated patches in the image. The SRGAN work best for images with large similar color patches with minimal texture details inside it, whereas it tends to lose some of the fine texture details in case of image patches with granular details. In the Example - 1 the edges are preserved in the case of SRGAN. The traditional interpolation techniques appear to smoothen the SR image but the SRGAN appears to generate sharper images in the case of Example - 2.

In the case of Example – 3 where there are significant texture details, the traditional techniques appear to fail to extrapolate the texture details as the images appear blurry and pixelated. SRGAN appears sharp and some of the texture details are estimated in a better way. SRGAN output is not close to the original HR image in this case but it generates the best output among the lot.

Upon trying the inference on LR images that have down sampled with unknown realistic downgrading, the output seems to aggravate the noise in the images. These images have been obtained from the DIV 2K website, which have the mild, difficult and wild degree of degradation in the LR images.

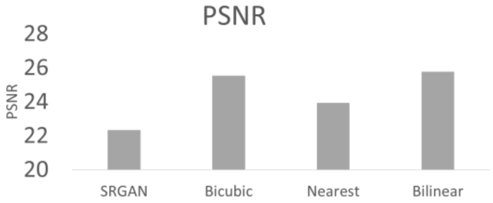


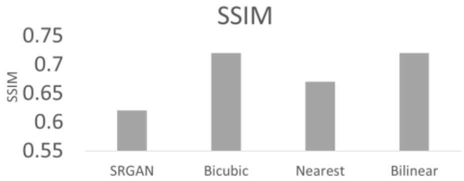
Mild

Difficult

Wild

The popular metrics for SR process are PSNR and SSIM. PSNR is used to measure the quality of lossy transformation. It is based on maximum pixel value and mean square error. Typical value ranges from 20 to 40, higher the value the better. SSIM measures the structural similarity between two images. It is based on comparing the luminance, contrast and structure. Value ranges from 0 to 1 and higher the value the better. We have computed the average PSNR and SSIM on the test images and the plot was shown below. The corresponding traditional interpolation techniques have been compared to the SRGAN output. Both PSNR and SSIM values are less than the corresponding values for traditional SR methods.





4.2. Discussion

The fact that SRGAN performs better than the traditional interpolation methods has been confirmed as claimed by the Ledig [1]. Also, PSNR and SSIM fail to accurately capture the perceptual quality of the reconstructed image. This corroborates the phenomenon given in Blau [8]. The focus of the implementation is to improve the perceptual quality if the SR images and this can be improved by using the content loss function of the feature maps from VGG network. The appropriate choice of the loss function can be critical for different applications were the finer details in the SR image can be of importance or not.

The limitation of this exercise is that the model only learns the SR process for a specific down scaling method. As confirmed by the results, the model doesn't perform well on images that have been down scaled using other techniques.

Also, unavailability of a good metric for measuring the perceptual quality limits the quantitative measure of the generated SR images.

5.1. Conclusion and Future work

We have successfully reimplemented the SRGAN architecture and replicated the results. Our findings substantiate the claims made the original authors that SRGAN offers very good perceptual quality compared to traditional interpolation methods. We have also confirmed the claim that PSNR and SSIM are not good metrics for measuring the perceptual quality of the images. In future, Mean Opinion Score (MOS) can be utilized for gauging the perceptual quality of the images. Also, the network proposed is very deep and it can be further optimized for fewer parameters, low memory and faster training. This could also help in reducing the inference time.

6. Acknowledgement

We would like to take this opportunity to thank all the course staff who supported and helped us in completing this project. Firstly, we thank Prof. David Crandall and IU for making this amazing course available for us to register. Also, we would like to thank Prof. Crandall for making course so informative and for allowing us to choose the project of our liking. We are very grateful for giving us $50 Google cloud credit which has been very instrumental in completing the project in stipulated time frame. Special thanks to TA's (Eriya, Violet and Shyam) for helping us with the course. Finally, we thank all the fellow students who made the discussion sessions fun and informative.

7. Contributions

We started with data collection and literature review, in which all of us are actively involved. Siva Charan tried traditional methods of image super resolution. All three of us are involved in the model building process. Initially we have tried implementing it in Tensorflow, later migrated it to Keras. We further optimized the model and tried few experiments, the optimization part was mainly done by Satya, Adithya, while experiments were carried out by Siva Charan.

References

[1] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham,´ A. Acosta, A. P.Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photorealistic single image super-resolution using a generative adversarial network," in CVPR, 2017.

[2] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super resolution using very deep convolutional networks," in CVPR, 2016.

[3] Boris Kovalenko" Super resolution with Generative Adversarial Networks." X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "Esrgan: Enhanced super-resolution generative adversarial networks," in ECCV Workshop, 2018

[4] M. Cheon, J.-H. Kim, J.-H. Choi, and J.-S. Lee, "Generative adversarial network-based image super-resolution using perceptual content losses," in ECCV Workshop, 2018

[5] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, Christopher Schroers, "A Fully Progressive Approach to Single-Image Super-Resolution"

[6] Justin Johnson, Alexandre Alahi, Li Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution"

[7] Timofte, Radu and Gu, Shuhang and Wu, Jiqing and Van Gool, Luc and Zhang, Lei and Yang, Ming-Hsuan and Haris, Muhammad and others, "NTIRE 2018 Challenge on Single Image Super-Resolution: Methods and Results" https://data.vision.ee.ethz.ch/cvl/DIV2K/

[8] Yochai Blau, Tomer Michaeli, "The Perception-Distortion Tradeoff"