Super Resolution of Videos using GANs

Kevin Joseph
31004593
kevjoseph@umass.edu

Tejasvi Ravi
31225442
travi@umass.edu

Abstract

The task of super resolving an image (increasing the spatial dimensions of an image given a low-resolution image) is an ill-posed problem and is an active research topic within computer-vision. Recent work by Ledig et al. *[8] on photo realistic super-resolution of an image using a generative adversarial network(GAN) showed promising results. It could recover finer texture details of an image even when super-resolved with large scaling factors, more specifically, it was able to infer photo-realistic natural images for 4X upscaling factors. As part of our Computer Vision project we explore the usage of this technique on videos with some modifications that help output better results. Modifications to the standard implementation of SRGAN's [8] on videos was necessary for without which the resulting super-resolved video contained several "flickering" frames, i.e. frames with varying brightness. Our network is able to super-resolve variety of videos with gains in perceptual quality.*

1. Introduction

The idea of single image super resolution (SISR) of an image is to take a low resolution image and upscale the image to a higher resolution. There are many techniques that have been researched and showcased over the past years. Earliest techniques either used bicubic or were prediction based like Lanczos [5]. Though these were fast, they resulted in overly smoothed textures. More recently convolutional neural network based super resolution techniques have shown promising results. Works by Johnson *et al*. [7] and Bruna *et al*. [2], use deep networks with a loss function that is closer to perceptual similarity which recovers visually more convincing HR images.

Based on these works and with the aim to retain textures in super resolved images, Ledig *et al*. [8] proposed the use of GAN's which provided upto 4X up-scaling, generating sharp textures. The success of this technique on images, raises questions over the feasibility of SRGANs on videos as well. In this project we experiment with SRGANs on different videos and present our findings. [1]

2. Related Work

There has been substantial work on super-resolution of videos. The motivation to do this is fueled by the numerous implications this might pose. If successful, areas like video compression, transfer of video over network, video processing etc might have scope for improvement with these new techniques. Osama *et al*. [9] proposed an end-to-end video super-resolution network that included the estimation of optical flow in the overall network architecture. This paper showed that processing of whole images are responsible for a large increase in accuracy than using independent patches. Cabellero *et al*.[3] showcased a spatio-temporal sub-pixel convolution network that effectively exploited the temporal redundancies and improved reconstruction accuracy while maintaining real-time speed. All these works use different information from the frames preceding and succeeding the frame being super-resolved. For example, Osama et al [9] uses optical flow across the frames. In this project we see how photo-realistic the super-resolved video would look that has been super-resolved with SRGANs that just looks at a single frame at a time. With respect to single image super resolution, learning the upscaling filters become pivotal as they improve accuracy along with speed over models like the ones mentioned in Dong *et al*.[4] for example, which uses bicubic interpolation to upscale the image before feeding to the CNN. We also see that for videos unlike that in images, preserving average coherent brightness across multiple frames is essential.

3. Method

We have designed a framework that inputs a low resolution video, super resolves it by 4X. To achieve this, we separate the video into frames and super resolve each of these frames as if they were a single image, like in single image super resolution.

In single image super resolution, the aim is to estimate a

[1] Code is available at: https://github.com/ravisvi/super-resolution-videos.

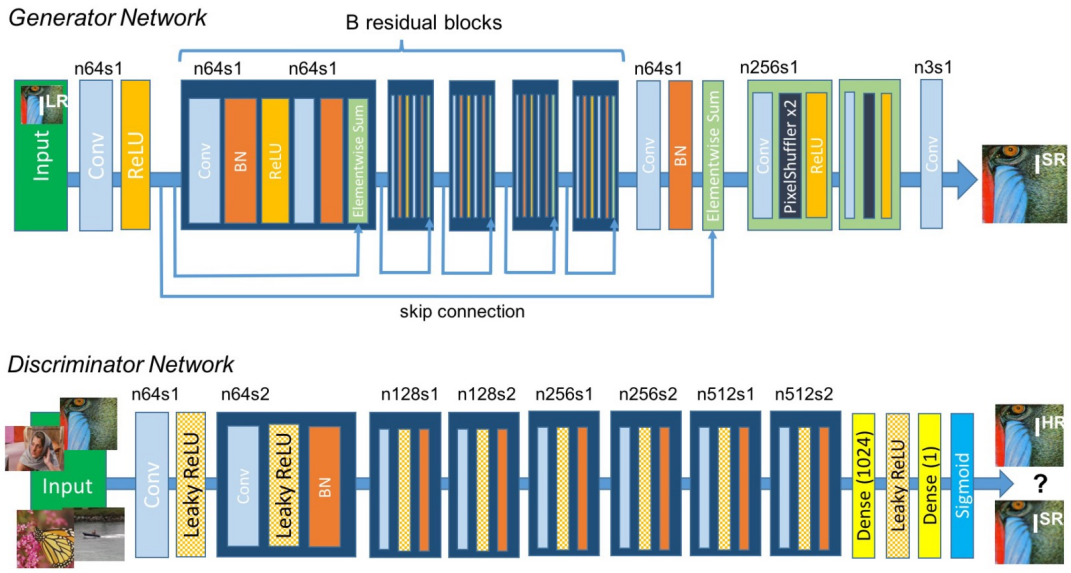

Figure 1. Architecture of Generator and Discriminator Network [8]

super-resolved image I^{SR} from a low-resolution input image I^{LR}. Here I^{LR} is the low-resolution version of its high-resolution counterpart I^{HR}. The high-resolution images are only available during training. For training, I^{HR} is obtained from the DIV2K dataset [1]. And the I^{LR} is obtained by downsampling the respective I^{HR}

We then use these to train a generator network as a feed-forward convolutional neural network G_{θ_G} parametrized by θ_G as in the Ledig *et al.* [8]. Here $\theta_G = \{W_{1:L}; b_{1:L}\}$ denotes the weights and biases of a L-layer deep network and is obtained by optimizing a SR-specific loss function l^{SR}. For training images I_n^{HR} , $n = 1, \ldots, N$ with corresponding I_n^{LR} , $n = 1, \ldots, N$, we solve:

$$\hat{\theta}_G = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR}) \qquad (1)$$

3.1. Adversarial network architecture

In our project, like in the Ledig *et al.* [8] paper there is a discriminator network, D_{θ_D} which is optimized in an alternating manner along with G_{θ_G} to solve the adversarial min-max problem:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^{HR} \sim p_{\text{train}}(I^{HR})}[\log D_{\theta_D}(I^{HR})] + \mathbb{E}_{I^{LR} \sim p_G(I^{LR})}[\log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))] \qquad (2)$$

The idea behind this is that it allows one to train a generative model G with the goal of fooling a differentiable discriminator D that is trained to distinguish super-resolved images from real images thus encouraging perceptually superior solutions.

3.1.1 Sub-pixel Convolution Layer

In the generator architecture all the activation maps that are computed are of the same spatial dimension as the low-resolution image. There are repeated units of {Conv, Batch-norm.ReLU and Element-wise Sums}. The spatial dimension of the activation maps are increased by a multiplicative factor using a sub-pixel convolution layer as proposed by Shi *et al.* [10]. This layer essentially uses regular strided convolutional layers followed by a specific type of image reshaping called a phase shift. In other words, instead of putting zeros in between pixels and having to do extra computation, they calculate more convolutions in lower resolution and resize the resulting map into an upscaled image. This way, no meaningless zeros are necessary.

3.2. Loss function

The perceptual loss function l^{SR} is pivotal for the performance of the generator network. Perceptual loss is the weighted sum of a content loss (l_{X}^{SR}) and an adversarial loss component:

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

In the following we describe possible choices for the content loss l_{X}^{SR} and the adversarial loss l_{Gen}^{SR}.

3.2.1 Content loss

The first part of the content loss is the VGG loss based on the ReLU activation layers of the pre-trained 19 layer VGG network described in Simonyan and Zisserman [11] as the euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR}:

$$l^{SR}_{VGG/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \quad (4)$$

Here $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network.

However, using this alone as content loss will not capture the pixel intensities. This was not the case in [8] as it was more concerned with just a photo-realistic super resolution of an image. But when it comes to super resolving a video, the pixel intensities become important as well. Because in videos, the pixel intensity differences from one frame to the next don't change much. Hence there is a need to maintain consistency of pixel intensities. Thus we added a L1 loss to the content loss to enforce pixel intensity constancy. We see a remarkable improvement in the video quality that is shown in the experiment section.

$$\underbrace{l^{SR}_{\text{X}}}_{\text{content loss}} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \|(I^{HR})_{x,y} - (G_{\theta_G}(I^{LR})_{x,y}\|_1 + \alpha l^{SR}_{VGG} \quad (5)$$

In our experiments we chose α to be 2×10^{-6}.

3.2.2 Adversarial loss

To favor solutions that reside on the manifold of natural images, [8] also adds an adversarial loss. This generative loss l^{SR}_{Gen} is defined based on the probabilities of the discriminator $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ over all training samples as:

$$l^{SR}_{Gen} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \quad (6)$$

Here, $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ is the probability that the reconstructed image $G_{\theta_G}(I^{LR})$ is a natural HR image. For better gradient behavior they minimize $-\log D_{\theta_D}(G_{\theta_G}(I^{LR}))$ instead of $\log[1 - D_{\theta_D}(G_{\theta_G}(I^{LR}))]$ [6].

3.3. Training Procedure

It is to be noted that it is possible for the generator network to learn an up-sampling function by minimizing just the *Content Loss* ($l^{SR}_{\text{content loss}}$) on pairs of low-resolution and high resolution images. The addition of the Adversarial and the GAN training procedure incentivizes the generator to produce more realistic high resolution images with finer details.

Figure 2. SRGAN's (trained on MNIST) super-resolving by a factor of 2.

4. Experiments

4.1. Testing on MNIST dataset

To test the model, the first step was to test it on a domain specific dataset. We used the MNIST dataset which comprised of handwritten grayscale images of numerals. Figure 2 shows the results on images that were not used during the training process. The SRGAN super-resolved images (scaling factor of 2) shows significantly better perceptual quality than compared to the ones upscaled using bi-cubic interpolation.

4.2. Testing on DIV2K dataset

The next step was to train the model on a dataset with more variation. We made use of the DIVerse 2K [1] dataset. This data-set consisted of 800 high resolution images (not specific to any class) and for the corresponding low-resolution images we downsampled the HR images by a factor of 4 using bi-cubic interpolation.

Figure 3 shows the SRGAN results on a test image where the input LR image was upscaled by a factor of 4. We see that the fine details have been reconstructed.

4.3. Testing on videos

We applied the SRGAN model to several videos and what was observed initially was that even though the model was able to recover details in the upscaled frames, there was a significant change in the overall brightness of the output frames.[2]

[2]Links to videos-https://www.youtube.com/watch?v=aQelFUwvXlI https://www.youtube.com/watch?v=8OY8HFGsbKM https://youtu.be/arO-yovw_9A

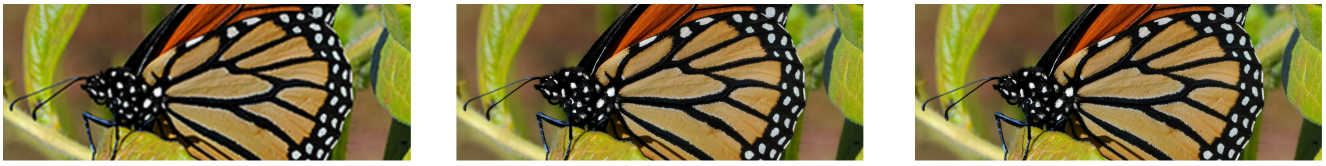

Figure 3. SRGAN results compared to traditional bi-cubic interpolation

Figure 4. Comparison of SRGAN's (with and without L1 loss) when applied to video frames.

Figure 5. Comparison of SRGAN's (with and without L1 loss) when applied to video frames.

This rapid change in brightness between the frames led to a flickering effect in the super-resolved video. Figures 4 and 5, attempts to depict this phenomena but it is much more prominent when one views the videos provided in the footnote of the previous page. Looking at the "VGG" row in Figure 4 and 5, we can see that between successive frames there is a significant change in brightness. This however was minimized by adding an L1 loss to the *Content Loss* and retraining the SRGAN on the DIV2K dataset.

To evaluate the methodologies, we calculated the mean squared error and the peak signal to noise ratio scores for super-resolved images using bicubic, plain SRGAN's and SRGAN's with L1. The results are tabulated in tables 1 and 2. As expected the MSE scores favors bicubic super-

Figure 6. 1st row: low resolution frame, 2nd row: frame super-resolved with SRGAN and L1 loss, 3rd row: frame super-resolved with SRGAN.

resolved images. But we know that these images cause too many blurry artifacts. For this reason we calculate the PSNR scores as well. Looking at the PSNR scores table, we see that, generally, the PSNR is lower for images super-resolved with SRGANs with L1 than the images that are super-resolved with plain SRGANs. However, when the video results are looked at, the videos super-resolved using the SRGANs plus L1 performs better than the video super-resolved with plain SRGANs. Therefore, for videos, the right metric to evaluate would have to rely more on visual perception rather than the PSNR scores. [3]

[3]The images h1 to h5 and their counterparts can be found in the evaluate folder in the github repository. https://github.com/ravisvi/super-resolution-videos/tree/master/evaluate

Dataset	Bicubic	SRGAN	L1SRGAN
h1	17.84	24.78	42.22
h2	29.60	14.90	24.04
h3	35.18	34.09	40.48
h4	44.01	41.85	55.15
h5	8.83	9.77	17.23

Table 1. The mean squared error of different methods evaluated on few images. Where Bicubic stands for results of images super-resolved with bicubic interpolation, SRGAN stands for results of images super-resolved with plain SRGANs and L1SRGAN stands for results of images super-resolved with SRGANs containing L1 loss.

Dataset	Bicubic	SRGAN	L1SRGAN
h1	35.65	34.22	31.94
h2	33.45	36.49	34.58
h3	32.70	32.84	32.09
h4	31.73	31.95	30.75
h5	38.70	38.26	35.80

Table 2. The peak signal to noise ration (dB) of different methods evaluated on few images. Where Bicubic stands for results of images super-resolved with bicubic interpolation, SRGAN stands for results of images super-resolved with plain SRGAN and L1SRGAN stands for results of images super-resolved with SRGANs containing L1 loss.

5. Conclusion and Future Work

We have successfully trained our network that super resolves a lower resolution image into a higher resolution image. We use this network to then super resolve videos as well. Currently we super resolve each frame at a time, which is one of the main culprit behind the time taken to super resolve the whole video to be almost the same as the video length. We can solve this super resolving the images batch wise, as well as using a multi-threaded approach. Also, comparisons of this technique versus that use information present in the surrounding frames should be made to see which technique suits which video type better. When super resolving live videos or videos of sports, the frames usually contain multiple blurry regions. These frames that go through the SISR generate unsatisfactory super resolved frames. Further investigations on techniques that can remove the blurriness before passing it on to our network can be made.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[2] J. Bruna, P. Sprechmann, and Y. LeCun. Super-resolution with deep convolutional sufficient statistics. *CoRR*, abs/1511.05666, 2015.

[3] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. *arXiv preprint arXiv:1611.05250*, 2016.

[4] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

[5] C. E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.

[8] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[9] O. Makansi, E. Ilg, and T. Brox. End-to-end learning of video super-resolution with motion compensation. In *German Conference on Pattern Recognition*, pages 203–214. Springer, 2017.

[10] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CoRR*, abs/1609.05158, 2016.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.