Machine Learning based Signal Processing for Video Quality Enhancement using GANs

Devansh Raj
Electrical and Electronics dept. (Student)
BITS Pilani K.K. Birla Goa campus
Goa, India
f20180531@goa.bits-pilani.ac.in

Nitin Sharma
Electrical and Electronics dept. (Professor)
BITS Pilani K.K. Birla Goa campus
Goa, India
nitinn@goa.bits-pilani.ac.in

***Abstract*—Recent studies on Super-resolution in videos has shown that the use of loss function like mean squared error(MSE) can be improved by defining perceptual loss as the sum of MSE and adversarial loss. Research seeks to solve the problem of enhancing low resolution videos using Generative Adversarial Networks (GANs), applying image super resolution in the frames of the video for video quality enhancement, so that we can transfer limited data and can enhance the output in the receiving end. Our model uses Super resolution generative adversarial network architecture combined with a perceptual loss function to enhance the quality of the input video and reduce the flickering effects on the frames. Since the quality of the video is a subjective thing depending upon applications and individuals, we used mean-opinion score(MOS) for testing and rating the output video frames.**

***Index Terms*—Generative Adversarial Network, Video Enhancement, Super Resolution**

I. Introduction

Current high-resolution camera systems require more expensive equipment and a large amount of digital storage to operate, reducing their efficiency and effectiveness. For exchanging and storing large chunks of data our devices use different types of compression techniques for optimizing the data storage space. In case of visual data compression, video compression algorithms result in reduction of image quality in the frames of the video, this affects commercial streaming services such as Netflix, Amazon Prime etc., as well as many applications which provide video conferencing services like Google meet. Recording and saving security camera footage also requires video compression as this allows you to store more data in the hard drive, but which will lead to a significant increment in the cost and complexity of such kind of systems [1], [2]. So, in all these above-mentioned cases it's very important to improve the quality of the video when it is processed after the compression, for human view and automatic video analysis.

Super Resolution Generative Adversarial Network (SRGAN) is used for single image super-resolution which we're using in the frames of the video so as to finally get an enhanced video. The basic idea of how GANs work is quite simple and interesting, it is basically a system where two competing neural networks compete with each other to create or generate variations in the data. GANs models are used for unsupervised learning. In this report we've mentioned about several objectives of the project and the important information regarding them, based on the research.

II. Related Works

A. Video enhancement and SRGANs

Most of the papers which include video quality enhancement use only spatial filtering techniques like toning, contrasting, etc. to form their network, such as in [3]. They enhance the video quality in a trade-off with the specifications of the frame, though the pixels above a certain threshold will be differentiated from the pixels below the threshold, but if the histogram of pixels is nearly equalised then important details from the frame will be lost which is undesired in most of the cases.

Researchers also tried to make different algorithms using various computer vision libraries like Open Source Computer Vision(OpenCV) for different streams like medical field [4], visualising data for mobile multimedia satellites services [5] etc. but a major problem in using these libraries was the applications were particularly subjected to the supervised learning, which resulted in limiting the domain of the application.

Some of the amazing researchers actually figured out the problems based on the source of images/videos and explained reasons which results in lower quality of the output videos. They also proposed different possible solutions depending upon the application, one of the solution was the use of Super-resolution GANs in the images for enhancing the quality, but most of the researches were particularly focused on image enhancement rather than video quality enhancement, such as in [6]. The advantage with SRGANs models when implemented with loss function is that they can be used for unsupervised learning while making sure that the required details are retained in comparison with simple neural networks.

B. Contribution

SRGANs provide an excellent framework to produce realistic frames similar to High resolution frames with pretty good perceptual quality. In this paper we mainly focused on improving the perceptual loss function with optimizing the processing code for a faster and more accurate output with better MOS value. Our main contributions are:

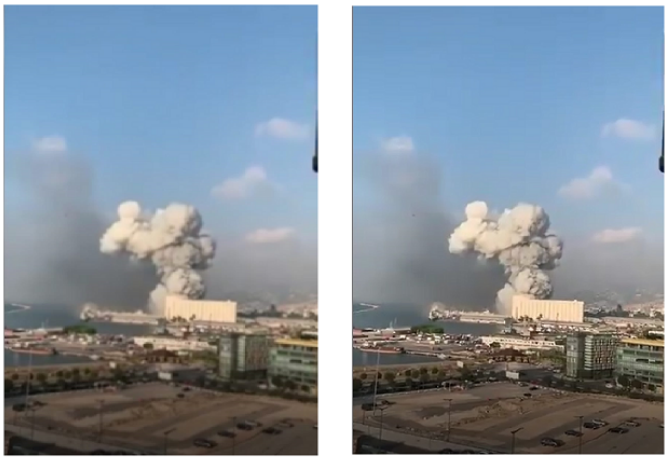


Fig. 1. Comparison between LR input frame and Generator output.

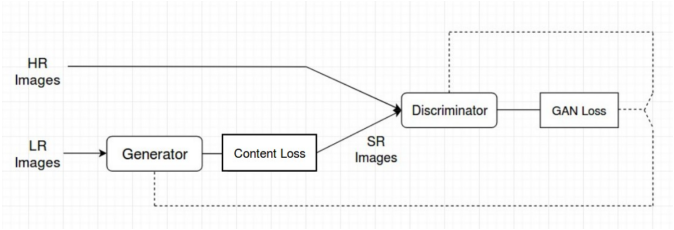


Fig. 2. Structure of the SRGAN with loss function.

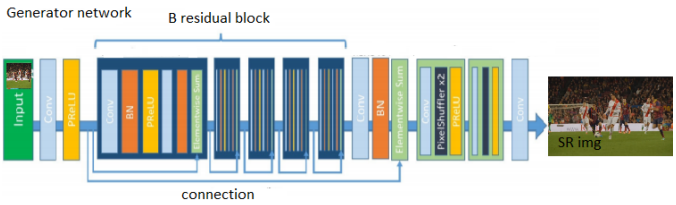


Fig. 3. Architecture of Generator network.

- Replaced sigmoid function which was used in most of the research papers to form the neural network such as in [9], with rectified linear activation function which is:

$$ReLU(a) = max(0, a) \tag{1}$$

  ReLU function is basically a simplification of sigmoid function with an advantage that it's easier to train with less time requirement.
- Replaced old school way of Mean squared error loss with a new perceptual loss function after getting inspired from [10], where they used a similar concept for image enhancement.
- Improved processing time of the generator network, after taking the help of [11] for building the basic generator network. We made sure that mean opinion score does not suffer with the high factor(4x) up-scaling of low resolution frames fed to the generator, which can be seen in Fig. 1.

III. Proposed Solution

The major driving force behind the research was to find the best performing Super-resolution Generative Adversarial Network which can be combined with a loss function to super resolve the input video by 4x and reduce the flickering frames problem. To perform the mentioned objective we built a stable model which can capture the perceptual difference between generator's output and the original high resolution frames during the training of the model. Fig. 2 shows the architecture of SRGAN, with perceptual loss function which consists of content and adversarial loss.

A. Adversarial network architecture

- The main idea behind the formation of network is to train the generator part G such that it fools the discriminator D which is trained to differentiate super-resolved frames from real frames(i.e. HR frames during the training). With the mentioned idea of SRGANs the generator will eventually learn to create output that are very similar to High-resolution frames and thus making it difficult to distinguish by Discriminator.
- Generator: At the base of our complex Generator network are B residual blocks with similar layouts. We used two convolutional layers with 3x3 kernels and 64 feature maps which is followed by batch-normalization layers and rectified linear activation function.Resolution of input frames are increased by sub-pixel convolutional layers which is proposed by Shi el at [7]. Architecture of the described generator network is shown in Fig. 3.
- Discriminator: The main job of the discriminator is to discriminate the super-resolved frames generated by the generator from the real high resolution frames, architecture of the network is shown in Fig. 4. The network consists of 8 convolutional layers with an increasing number of 3x3 filtering kernels, which is increasing by a factor of two from 64 to 512 kernels similar to the VGG network as in [8].

B. Perceptual loss function

While the perceptual loss is generally calculated by summing all the squared errors between all the pixels and taking the mean such as in [7], [12]. We, in our model calculated the perceptual loss as the weighted summation of content and adversarial loss, taking the idea from [11]. In contrast to per pixel loss function where we sum all the absolute errors between pixels, we made a comparison between the mentioned methods, perceptual loss calculation is not only

Fig. 4. Architecture of Discriminator network.

accurate, but faster than per pixel loss when optimized. For the three types of losses i.e. pixel wise mean error loss, loss defined on feature map from pre-trained VGG-network using ImageNet and self-feature where loss is defined on features taken from our network, we made comparison based on two datasets, visual results of which are shown in section IV. Table of comparison is shown here:

Set-1	***Pixel***	***VGG-feature***	***Self-feature***
PSNR	29.31dB	26.93dB	29.85dB
SSIM	0.8849	0.81	0.8871

Set-2	***Pixel***	***VGG-feature***	***Self-feature***
PSNR	26.22dB	24.6dB	29.85dB
SSIM	0.8054	0.7322	0.82

Self feature loss function id defined as:

$$L_1 = \frac{1}{2}||F1 - \hat{F}1||^2 \tag{2}$$

$$L_2 = \frac{1}{2}||F2 - \hat{F}2||^2 \tag{3}$$

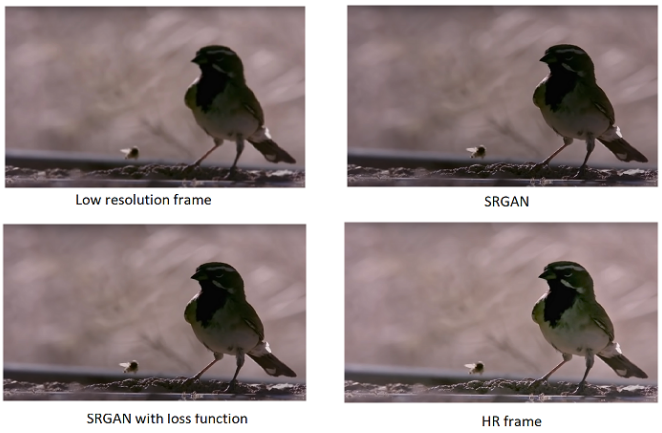
$$L_p = \frac{1}{2}||y - \hat{y}||^2 \tag{4}$$

$$L = L_1 + L_2 + L_p \tag{5}$$

In the above equations, L1 and L2 represents the feature loss of two consecutive layers and Lp represents the pixel loss, the above basic functions are taken from [10] ,though for the weighted sum of the losses the weights are adjusted according to the need.

IV. Visual Results

Above result shows the low resolution frame(top left), generated output using only SRGAN structure(top right), generated output using SRGAN with loss function(bottom left), and the original High resolution frame(bottom right), with improved frame quality, loss function also helps in reducing the flickering effect in the generated output.



V. Conclusion

In our model we have presented simple and efficient framework for video quality enhancement using SRGAN architecture, we compared different loss functions and calculated their mean opinion score to compare them on various grounds. A major objective behind the research was to formulate a SRGAN model with an optimized loss function to decrease the processing time once the generator is trained.

VI. Acknowledgment

I would like to show my gratitude to Professor Nitin Sharma for giving me this opportunity to enhance technical, interpersonal and research based skills through practice while working on this project. It helped me gain proficiency to solve various problems, understanding neural networks and it's implementations in real world.

References

[1] Z. Ashani, "Architectural considerations for video content analysis in urban surveillance", 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 289-289, Sep. 2009.

[2] Y. Zheng, C. Ye, S. Velipasalar and M. C. Gursoy, "Energy efficient image transmission using wireless embedded smart cameras", 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 62-67, Aug 2014.

[3] D. BakkiyaLakshmi, R. Kanchana and V. Nagarajan, "Video enhancement using tone adjustment," 2012 International Conference on Communication and Signal Processing, 2012, pp. 124-127, doi: 10.1109/ICCSP.2012.6208407.

[4] L. Lévêque et al., "On the Subjective Assessment of the Perceived Quality of Medical Images and Videos," 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), 2018, pp. 1-6, doi: 10.1109/QoMEX.2018.8463297.

[5] M. Nakayama, H. Suda and H. Mizuno, "Quality enhancement scheme for mobile multimedia satellite services," Vehicular Technology Conference. IEEE 55th Vehicular Technology Conference. VTC Spring 2002 (Cat. No.02CH37367), 2002, pp. 1859-1863 vol.4, doi: 10.1109/VTC.2002.1002943.

[6] H. Wang, W. Wu, Y. Su, Y. Duan and P. Wang, "Image Super-Resolution using a Improved Generative Adversarial Network," 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2019, pp. 312-315, doi: 10.1109/ICEIEC.2019.8784610.

[7] W. Shi, J. Caballero, F. Huszar, 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. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1874–1883, 2016.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR), 2015.

[9] Go Tanaka, Noriaki Suetake and Eiji Uchino, "Image enhancement based on multiple parametric sigmoid functions," 2007 International Symposium on Intelligent Signal Processing and Communication Systems, 2007, pp. 108-111, doi: 10.1109/ISPACS.2007.4445835.

[10] Z. Gao, E. Edirisinghe and S. Chesnokov, "Image Super-Resolution Using CNN Optimised By Self-Feature Loss," 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 2816-2820, doi: 10.1109/ICIP.2019.8803279.

[11] A. Lucas, A. K. Katsaggelos, S. Lopez-Tapuia and R. Molina, "Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution," 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 51-55, doi: 10.1109/ICIP.2018.8451714.

[12] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016.