IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

Soham Sheth
Dr Mrs M P Atre, Prof M P Potadar, Prof Mrs T S Khatavkar

Department of Electronics and Telecommunications,
Pune Vidhyarthi Griha's College of Engineering and Technology, India

Abstract

This paper implement the Convolution Neural Network Architecture with an efficient Sub-pixel convolution layer as proposed by wes of the paper,"Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network" authored by Wenzhe Shi, Jose Caballero et.al. With the use of this novel architecture the author replaced the handcrafted Bi-cubic Interpolation with more complex up-scaling filters specifically trained for each feature map. The proposed algorithm also re- duces the computational complexity as compared to a hand-crafted Bi-cubic Interpolation.

1 Introduction

In most electronic imaging applications, we desire images with High-Resolution, which means that the pixel density within an image is high and consequently a High Resolution Image can offer more details that may be critical to certain applications. For example, High Resolution medical images are very crucial for doctors to make an accurate diagnosis.

Due to limitations of sensors and optics manufacturing technology, the cur- rent resolution level and consumer price will not satisfy the future demand. One promising approach is to use Signal Processing to obtain a High-Resolution (HR) Image from a Low Resolution (LR) image. A key assumption that underlies many Super Resolution (SR) techniques is that much of the high- frequency data is redundant and thus can be accurately reconstructed from the low frequency components. This way we can save a lot of storage and reduce computational complexity by storing and processing on LR images and reconstructing the HR image as and when the need arises.

2 Motivation

The problem of Super Resolution was motivated partly by the Digital Image Processing course undertaken by one of the members, who pitched it as an extension of the Restoration problem faced commonly. The problem of Super-Resolution specifically was one of the most intriguing because of its numerous and diverse usage. This project was also carefully thought of in the context of Applications and contribution to Atmanirbhar Bharat. Super- Resolution can help cut out immense hardware and cost requisites, both of which are a rarity in Rural Areas. As the population of Rural Areas comprises about 65% of India's population, this application affects the heart of India directly. Application use cases are explained further towards the end of the report.

3 Background

3.1 Problem of Super Resolution

The problem of Super Resolution has been tackled by researchers in several ways. The global SR problem assumes the Low Resolution data to be a blurred (low pass filtered in frequency domain), down-sampled and a noisy version of the High Resolution image. The major hurdle in the reconstruc- tion of the HR image is that we lose high frequency data during LPF and Sub-sampling operations, which are non-invertible.

In addition to this, the mapping of a HR Image to a LR image is many to one which makes solution to this problem statement non-trivial. Possibil- ity of many solutions make this an inference problem and the solution relies on how well can the model predict the human perceptible high resolution image.

3.2 Data Processing Inequality

Data Processing Inequality states that the *information content of a signal cannot be increased via a local physical operation.* In other words, post- processing of the signal/data (image in our case) cannot increase the infor- mation present inside the data.

This may suggest that super resolution is impossible at first because you can't add additional details/pixel values while Up-sampling but just fill the missing pixel values based on the neighbouring values of the LR image given. This is what exactly happens in basic interpolation techniques to up-scale the image which creates no additional information. Hence, the results are not that up to the mark but similar or worse in quality.

But if the author uses a pre-trained model which takes an LR image and outputs a HR image based on the training samples, the author can add more details. This is because the trained model acts as an additional source of information. A neural network can learn to hallucinate details based on some prior informa- tion it gains from large sets of images. Pixel values added this way won't violate the data processing inequality because information is there somewhere in the training sets even if it isn't present in the input image. This is what the author is going to implement to get better quality HR images compared to other methods.

3.3 Nearest Neighbour Interpolation

A Naive solution to up-scale the image is to use predetermined fInterpolation techniques. In Nearest neighbour interpolation, as the name suggests, new pixels added are assigned the value of the pixel nearest to it where nearest neighbour is decided based on certain basic rules. As there is no extra infor- mation the author makes use of, this method is not useful for applications where we require a higher resolution.

3.4 Bi-Linear and Bi-Cubic Interpolation

Interpolation technique can be further improved by using Bi-Linear and Bi-Cubic Interpolation techniques. Unlike assigning the neighbouring pixel

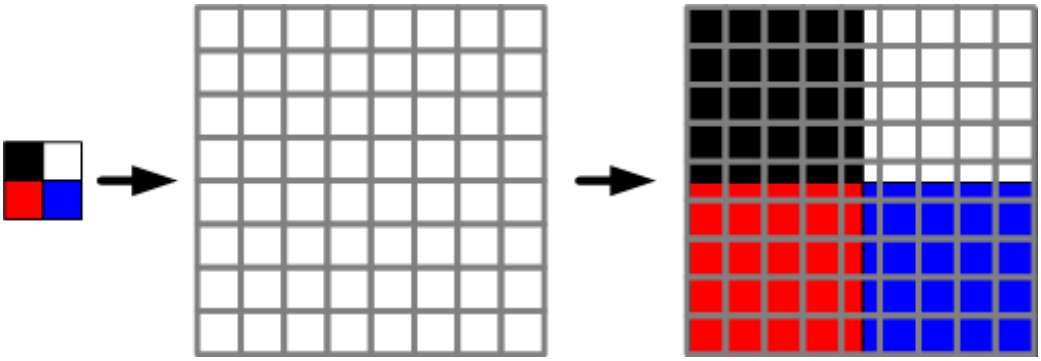


Figure 1: Nearest Neighbour Interpolation

value to the new pixels padded, in Bi-Linear Interpolation, weighted average of the neighbouring four pixel values is assigned to the new pixel based on it's distance from neighbours.

Bi-Cubic interpolation takes this one step further. Intuitively, it tries to in- corporate not just the neighbouring pixel values but the derivative at those places too. Now the value of the new pixel depends on the neighbouring val- ues of sixteen pixels. In general, this results in better quality scaled images compared to previous methods.

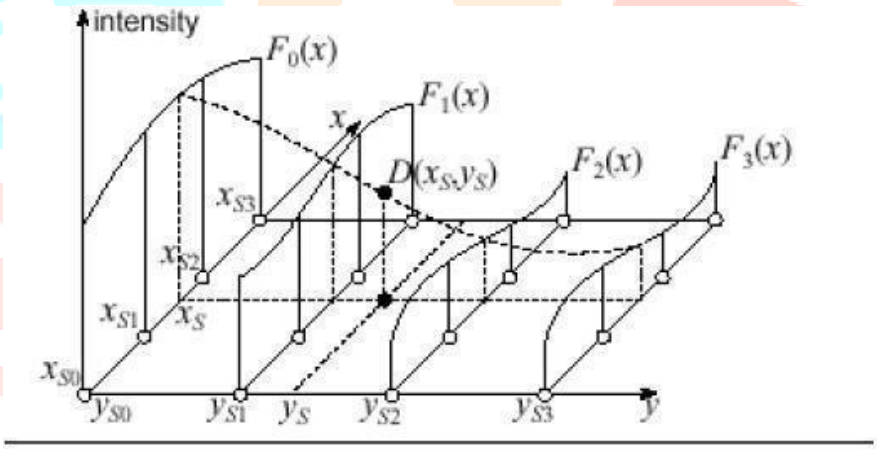


Figure 2: Bi-Cubic Interpolation

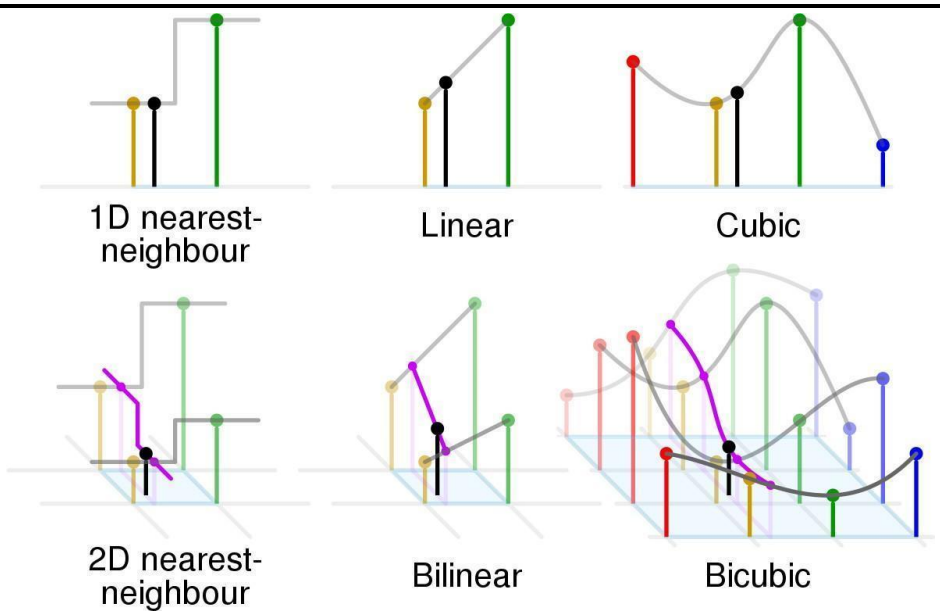


Figure 3: Various Interpolation schemes

3.5 Multi-Image Super Resolution

Another method of constructing HR images is using multiple LR images of the same scene with different perspectives. These multi-image SR methods exploit explicit redundancy by constraining the ill-posed problem with ad- ditional information and attempting to invert the down-sampling process. However, these methods usually require computationally complex image reg- istration and fusion stages, the accuracy of which directly affects the quality of the image constructed.

3.6 Convolutional Neural Networks

Finally we have the Single Image Super Resolution (SISR) method that we wish to implement. These techniques seek to learn implicit redundancy that is present in natural data to recover missing HR information from a single LR instance. This usually arises in the form of local spatial correlations for images and additional temporal correlations in videos.

In this case, prior information in the form of reconstruction constraints is needed to restrict the solution space of the reconstruction. This is done by training a Convolutional Neural Network on the sets of pair of HR and SR images to learn the best mapping from SR domain to HR domain. Since many such mappings are possible, best mapping optimised over the chosen loss function is taken as our super resolution model.

CNNs are able to learn this mapping better than normal Feed-forward Neu- ral Networks because instead of learning a mapping in which each pixel value contributes independently, CNNs first obtain the high level features that are responsible for human perception. These are obtained by making use of 2- dimensional Convolution layers which are extensions of convolution operation in one dimension.

The objective of the Convolution Operation is to extract the high-level fea- tures such as edges, from the input image. CNNs need not be limited to only one Convolutional Layer. Conventionally, the first CNN Layer is responsible for capturing the Low-Level features such as edges, color, gradient orien- tation, etc. With added layers, the architecture adapts to the High-Level features as well, giving us a network which has the wholesome understanding of images in the dataset, similar to how humans would.

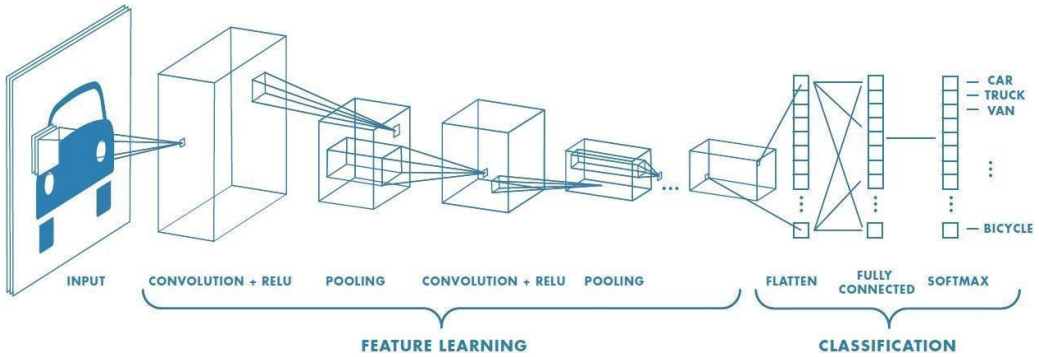
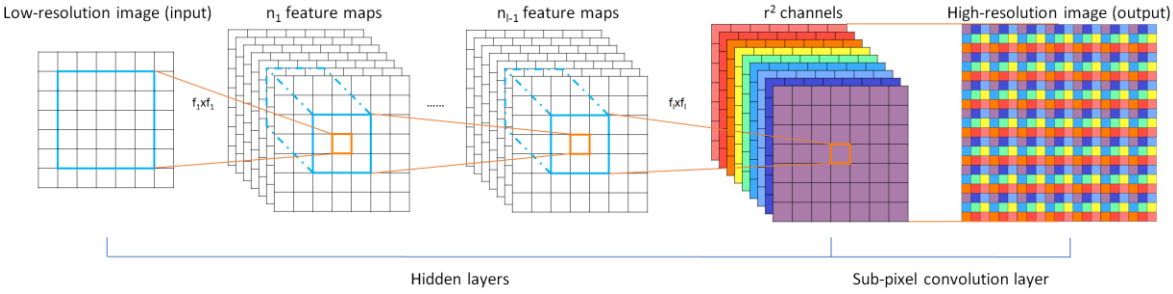


Figure 4: Convolutional Neural Network

3.7 ESPCN (Efficient Sub-Pixel Convolutional Neural Network)

FSPCN Network Architectu



re

Suppose there are *L* layers for the network,

1. **For the first *L*-1 layers, the input LR image goes through *fl*×*fl* convolution and obtains *nl*-1 feature maps.**
2. **At the last layer, an efficient sub-pixel convolution is performed to get back the HR image at the output.**

Specifically, *L*=3 which means it is a shallow network.

And the parameters for each layer are: (f1,n1)=(5,64), (f2,n2)=(3,32) and f3=3.

- **1st layer: There are 64 filters with the filter size of 5×5.**
- **2nd layer: There are 32 filters with the filter size of 3×3.**
- **3rd layer: There is only 1 filter with filter size of 3×3. This is because for a YUV image, only Y is considered**

as human eyes are more sensitive to luminance than chrominance.

3.7 Loss Function

Loss functions are used to measure the difference between the generated HR image and the ground truth HR image. This difference (error) is then used to optimize the supervised learning model (ESPCN in our case).

During the training process, the original HR images will be ground truth data. The mean squared error (MSE) is used to measure the difference be- tween the generated SR images and the ground truth HR images.

$$\text{MSE} = \frac{1}{\square} \sum_{\square=1}^{\square} (\square_{\square} - \hat{\square}_{\square})^{2}$$

Here I^{HR} represents each original image in the dataset; I^{LR} represents each down-sampled LR image; $f^{L}(I^{LR})$ represents the predicted HR image (model function that maps LR image to HR image); r represents the upscaling factor; H represents the image's height value; W represents the image's width value, W (1 : L) represents all the learnable network weights and b(1 : L) represents all the learnable biases.

4 Method

The author will be employing CNN based deep learning in the TensorFlow framework to accomplish the task of SISR (Single Image Super Resolution). Our model will be trained on the set of downsampled images as input and original high quality images as output. *YUV domain* of images was used instead of RGB to make the model computationally fast. For the upsampling in the model, an *Efficient Sub-Pixel Convolution Layer* instead of commonly used Bi-cubic interpolation was used after the initial convolution layers. After training the model, he predicted 5 up-sampled images, converted each of them back to their RGB versions and displayed them side by side with their low resolution versions and the original HR images with the Peak-to-Signal Noise Ratio (PSNR) obtained.

4.1 Similarity Metric

Since PSNR is highly correlated to MSE (via logarithmic function) we used PSNR as a reliable metric to compare performance. So a high PSNR (Signal-to-Noise Ratio) meant Lower MSE i.e. predicted image matrix values more closer to original HR image matrix values.

$$PSNR = 20 \log_{10}(MAX_I) - 10 \log_{10}(MSE)$$

A more realistic metric close to human visual perception is Structural Similarity (SSIM) Metric. It is a subjective metric used for measuring the structural similarity between images, based on three relatively independent comparisons, namely luminance, contrast, and structure. As evident from Figure 5, MSE isn't the best loss function. But due to complexity of SSIM, we used PSNR as the reliable metric

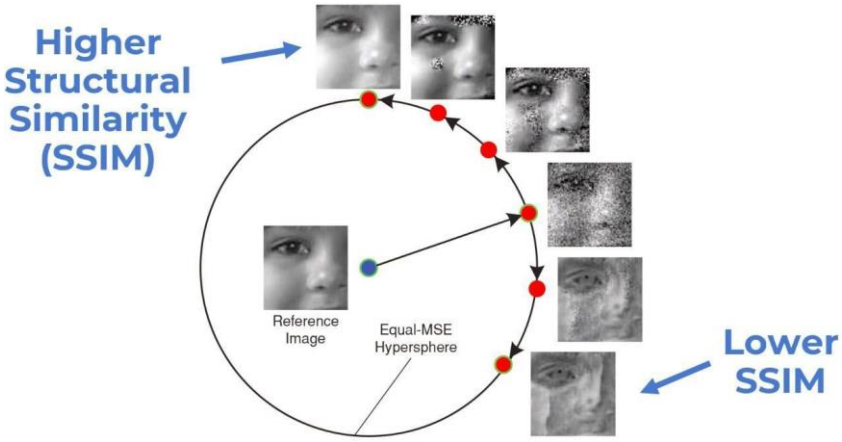


Figure 5: SSIM vs MSE

4.2 Architectural Details

The model follows an architecture quite similar to that of SRCNN with a difference that 4 convolution layers were used instead of 3 and at the last step for upscaling, ESPCN interpolation was used contrary to bicubic inter- polation deployed in SRCNN.
For the backpropagation mean squared error between the Y component orig- inal HR images and the corresponding Y component predicted SR images was minimised.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, None, None, 1)]	0
conv2d (Conv2D)	(None, None, None, 64)	1664
conv2d_1 (Conv2D)	(None, None, None, 64)	36928
conv2d_2 (Conv2D)	(None, None, None, 32)	18464
conv2d_3 (Conv2D)	(None, None, None, 9)	2601
tf_op_layer_DepthToSpace (Te	[(None, None, None, 1)]	0

Total params: 59,657
Trainable params: 59,657
Non-trainable params: 0

Figure 6: Model Architecture when coded in Tensorflow

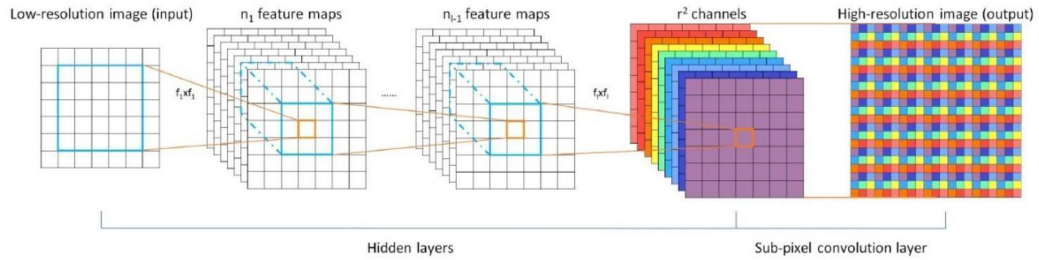


Figure 7: High Level View of Model

(The proposed efficient sub-pixel convolutional neural network (ESPCN), with two convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from LR space and builds the SR image in a single step)

Training Details

With the model architecture given in the previous subsection the author trained 2 versions of the same model with sight modifications.

1. Model1 with **'tanh'** as activation function in convolutional layers trained for **100 epochs**
2. Model2 with **'relu'** as activation function in convolutional layers trained for **200 epochs**

- The **learning rate** was taken as **.001** and **optimizer** as **Adam**.

- While the learning rate had already been experimented with in re- search paper, the choice of optimizer was rather the same as the research paper and some other code references the author found online. Also a rationale was that *Adam optimizer* works particularly well for *Computer Vision* tasks.

- To reduce the training time we converted RGB images to YUV domain and took only the Y dimension. For the human eye and the sophisti- cation of our project, it suffices to take only the Y component.

(We couldn't experiment ourselves with learning rate because it is a hit and trial process and the training time of the model was large. So we just stuck to the better value already known)

4.3 Dataset Used

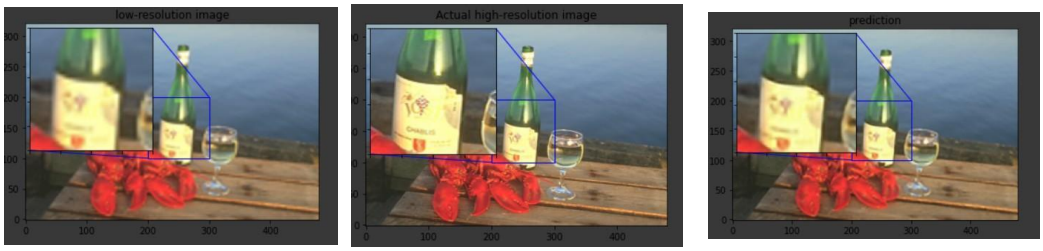
The Dataset used for Training and Testing purpose was the Berkeley Seg- mentation DataSet and Benchmarks 500 (BSDS500)

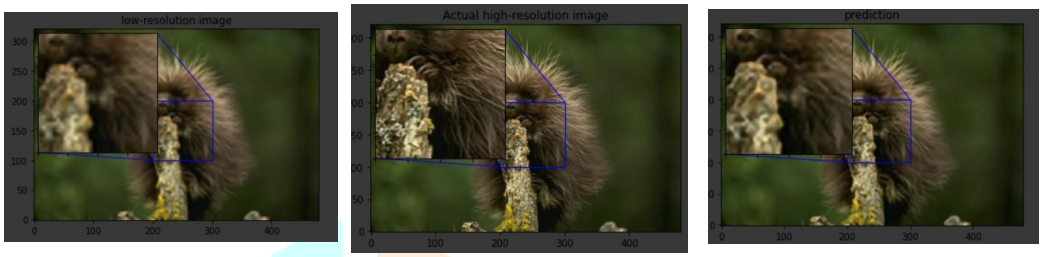
The dataset consists of 500 natural images, ground-truth human annota- tions and benchmarking code. The data is explicitly separated into disjoint train, validation and test subsets. There were 400 training images, 100 vali- dation and 200 test images.

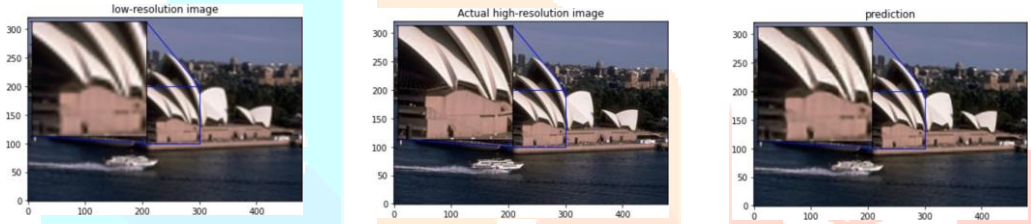
4.4 Output Results

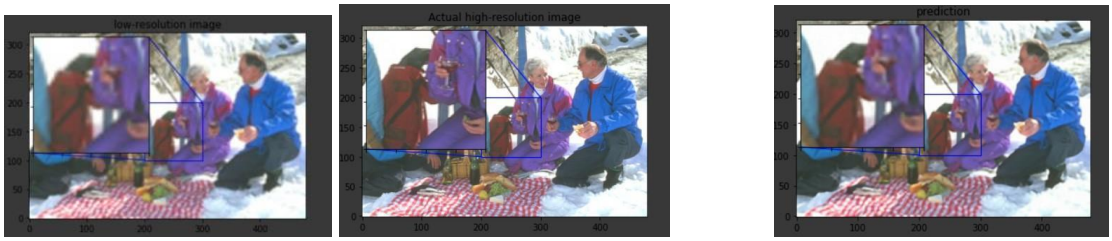
1. The Model1 mentioned in the training details subsection achieved an **average PSNR of 23.022**. However there isn't much visible difference between the reconstructed image and the LR image.
2. The Model2 on the other hand achieved an **average PSNR of 26.305** and these numbers also translate equally well to the predictions as the author can observe noteworthy differences between the reconstructed image and the LR image.

Below are some of the images* the author *reconstructed from Low res- olution images using our model. In all cases LR image, actual HR image and the corresponding predicted image from the model are shown.


low-resolution image
Actual high-resolution image
prediction


low-resolution image
Actual high-resolution image
prediction


low-resolution image
Actual high-resolution image
prediction



5 Remarks

Advantages of this Deep Learning method over other methods:-

- Since the author is using downsampled input images and all the processing (convolution operations and all) are being done on low resolution im- ages, it drastically decreases the computations needed. This is espe- cially true for CNN where image resolution directly dictates the com- plexity.
- The decreased computations further leads to faster training times due to computation intensive nature of backpropagation algorithm which is an added advantage considering the fact that ML is a play of hit and trial of finding right hyper-parameters.

Further Considerations:-

- Of course, the model can be improved by using all the three YUV channels for training instead of just 1 but there is a trade-off involved that training time goes up.
- One can also experiment further with the batch size and the number of epochs to find an even optimal set of model weights.
- PSNR, which was used as a metric for judging the model's performance may not always be the most accurate representation of human percep- tion. Instead it would be interesting to somehow incorporate SSIM in a modified loss function and minimise that.

6 DSP Techniques encountered

This project was a great learning experience where we were given an oppor- tunity too apply various Digital Signal Processing Techniques we studied. Following is the list of the major DSP Techniques that we encountered and applied during the course of this project:

1. **Down-Sampling using Low Pass Filter**

   HR images given in the dataset were down-sampled via Low Pass filter in order to get LR images which acted as the input to our Convolutional Neural Network.

2. **Interpolation Techniques for Up-Sampling**

   We learnt various Interpolation techniques used to Up-Sample the LR images to get HR images and compared them.

3. **2-Dimensional Convolution**

   Images and outputs of the various Neural Network layers were convul- ved with the 2-D Kernels to obtain higher order features like edges, texture, orientation etc which were further operated to get HR image.

7 Applications

7.1 Medical Imaging

Medical imaging involves exposing patients to radiation in order to obtain the 'image' of the part under examination. In order to get high quality im- ages, intense radiation is required for a long duration which can be harmful to the patient under examination. Super resolution can curb this issue by reconstructing high end images from multiple LR images captured, which can be further used by doctors to pin-point or diagnose the disease (for eg. identify the tumour). The SR technique is useful in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) since the acquisition of mul- tiple images is possible while the resolution quality is limited. These can be very beneficial for the developing nations like India especially for rural areas which are devoid of cutting edge technology, where diagnosis and medical help depends on the family income and how fast they can communicate to the nearest big city. Super Resolution can play a crucial role by obtaining poor resolution images from cheaper instruments and up-scale them for further diagnosis.

7.2 Surveillance and Military

Often, in satellite imaging applications such as remote sensing, large areas are to be scanned and observed by satellites/military in order to gain enemy information and plan the defenses. Not only this, but meteorology depart- ment also often need high resolution images of land, sea floor maps (and many more) and they rely heavily on the satellites for it.

In such cases, several images of the same area are usually provided, and the SR technique is used to improve the resolution of the target in order to get more and finer details.

7.3 Face Recognition and Crime Investigation

Synthetic zooming of regions of interest is another important application in surveillance, forensic, scientific, medical, and satellite imaging. For surveil- lance or forensic purposes, a digital video recorder (eg. CCTV) is often needed to magnify objects in the scene such as the face of a criminal or the licence plate of a car. SR is used to extract crucial information from them.

7.4 Digital and Print Media

Higher quality digital images can be reconstructed from Low resolution im- ages obtained with an inexpensive low resolution camera/camcorder for print- ing or frame freeze purposes. This technique can also be incorporated in mobile phone cameras which are limited by the camera quality where image captured can be enhanced via Super Resolution.

8 References

1.Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network

2.Image Super Resolution using Deep Convolutional Networks

3.BSDS500 Dataset

4.How Super Resolution Works

5.An Introduction to Super Resolution using Deep Learning

REFERENCES :

1. Schulter, Samuel, Christian Leistner, and Horst Bischof. "Fast and accurate image upscaling with super-resolution forests." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
2. Timofte, Radu, Vincent De Smet, and Luc Van Gool. "A+: Adjusted anchored neighborhood regression for fast super-resolution." *Asian conference on computer vision*. Springer, Cham, 2014.
3. Yang, Jianchao, et al. "Image super-resolution as sparse representation of raw image patches." *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008.
4. Aharon, Michal, Michael Elad, and Alfred Bruckstein. "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation." IEEE Transactions on signal processing 54.11 (2006): 4311-4322, 2006.
5. Bevilacqua, Marco, et al. "Low-complexity single-image super-resolution based on nonnegative neighbor embedding." (2012): 135-1 , 2012.
6. Burger, Harold C., Christian J. Schuler, and Stefan Harmeling. "Image denoising: Can plain neural networks compete with BM3D?." *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012.
7. Chang, Hong, Dit-Yan Yeung, and Yimin Xiong. "Super-resolution through neighbor embedding." *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*. Vol. 1. IEEE, 2004.