Images Super Resolution

1st Hà Bảo Anh
Dept of Information Technology
Industrial University of Ho Chi Minh City
Ho Chi Minh, Viet Nam
baoanhdsk15@gmail.com

2nd Lưu Đức Anh
Dept of Information Technology
Industrial University of Ho Chi Minh City
Ho Chi Minh, Viet Nam
lanhluu8@gmail.com

***Abstract*—Image Super Resolution is one of the image preprocessing steps to enhance image quality. Namely going from a low resolution (LR) image to an enhanced or improved version. Some implementation methods include taking a low-resolution image as input and upscaling to a high-resolution (LR) image using a single filter, commonly bicubic interpolation, before reconstruction. This means that the super-resolution (SR) operation is performed in the HR space but proved to be suboptimal and highly complex. There are many modern methods to help improve image quality through learning. Still, in this part of the study, we approach the convolutional neural networks (CNN) method and learn about the author's new proposal, an efficient sub-pixel convolution layer that learns an array of upscaling filters to upscale the final LR feature maps into the HR output.**

I. Introduction

Image Super Resolution refers to the task of enhancing the resolution of an image from low-resolution (LR) to high (HR). Super Resolution received substantial attention from within the computer vision research community and has a wide range of applications many areas such as Surveillance, to detect, identify, and perform facial recognition on low-resolution images obtained from security cameras [1]. Medical imaging [2], [3] and face recognition [4]. Deep learning techniques have been fairly successful in solving the problem of image and video super-resolution. This article's approach to a method is not new, but this is the foundation for further in-depth research.

II. Methodology

A. Pre-Upsampling Super Resolution

The method used traditional techniques–like bicubic interpolation and deep learning–to refine an upsampled image. Interpolation is estimating unknown values from known sample values and continuous samples from discrete samples. Image processing applications of interpolation include image magnification or reduction, sub-pixel image registration to correct spatial distortions, image decompression, and image super-resolution reconstruction. There are many interpolation techniques few famous them include nearest neighbor interpolation, bilinear, and cubic spline interpolation. Since interpolation does not give good results within an image processing environment, and image data is generally acquired at a much lower sampling rate, the mapping between the unknown high-resolution image and the low-resolution image is not invertible, and thus a unique solution to the inverse problem cannot be computed.

B. Post-Upsampling Super-Resolution

Since the feature extraction process in pre-upsampling SR occurs in the high-resolution space, the computational power required is also on the higher end. Post-upsampling SR tries to solve this by doing feature extraction in the lower resolution space, then upsampling only at the end, significantly reducing computation. Also, instead of using simple bicubic interpolation for upsampling, a learned upsampling in the form of deconvolution/sub-pixel convolution is used, thus making the network trainable end-to-end. Due to the reduced input resolution, they can effectively use a smaller filter size to integrate the same information while maintaining a given contextual area. The task of single image super resolution is to estimate a HR image $\mathbf{I}^{SR}$ given a LR image $\mathbf{I}^{LR}$ downscaled from the corresponding original HR image $\mathbf{I}^{HR}$. The downsampling operation is deterministic and known: to produce $\mathbf{I}^{HR}$ from $\mathbf{I}^{HR}$, they first convolve $\mathbf{I}^{HR}$ using a Gaussian filter - thus simulating the camera's point spread function - then downsample the image by a factor of r. They will refer to r as the upscaling ratio. In general, both $\mathbf{I}^{LR}$ and $\mathbf{I}^{HR}$ can have C colour channels, thus they are represented as real-valued tensors of size $H \times W \times C$ and $rH \times rW \times C$, respectively. For a network composed of L layers, the first $L-1$ layers can be described as follows:

$$f^1(\mathbf{I}^{LR}; W_1, b_1) = \phi(W_1 * \mathbf{I}^{LR} + b_1) \tag{1}$$

$$f^l(\mathbf{I}^{LR}; W_{1:l}, b_{1:l}) = \phi(W_l * f^{l-1}(\mathbf{I}^{LR})_l) \tag{2}$$

Where $W_i, b_i, l \in (1, L-1)$ are learnable network weights and biases respectively. W_l is a 2D convolution tensor of size $n_{l-1} \times n_1 \times k_l \times k_l$, where n_l is the number of features at layer $l, n_0 = C$, and k_l is the filter size at layer l. The biases b_l are vectors of lenght n_l. the nonlinearity function (or activation function) ϕ is applied element-wise and is fixed. The last layer f^L has to convert the LR feature maps to a HR image $\mathbf{I}^{SR}$. W. Shi [5] proposes a sub-pixel convolution neural network instead of using a deconvolution layer for upsampling. With two convolutional layers to extract feature maps and a sub-pixel convolution layer that aggregates feature maps from low-resolution space and builds a super-resolution image. Sub-pixel convolution works by converting depth to space, as shown in Fig 1. Pixels from multiple channels in a low-resolution image

are rearranged into a single channel in a high-resolution image $(*, C \times r^2, H, W)$ into a tensor of the form $(*, C, H \times r, W \times r)$ where r is the upscale factor. For example, if the input image has dimensions (C_{in}, H_{in}, W_{in}), then the output will be resized:

$$C_{out} = C_{in} \div r^2$$

$$H_{out} = H_{in} \times r$$

$$W_{out} = W_{in} \times r$$

with upscaling factor r = 2, for example, an image of size $(4 \times 5 \times 5)$ can rearrange the pixels resulting in an output image $(1 \times 10 \times 10)$

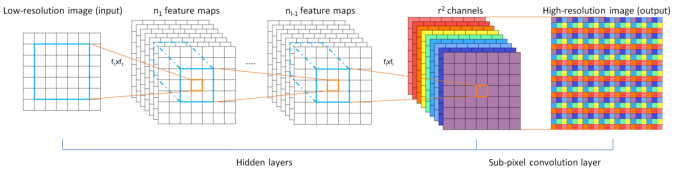

Fig. 1. Efficient Sub-pixel Convolutional Neural Network

III. Experiments Results

A. Dataset

We used the dataset DIV2K dataset was 1000 2K resolution images divided into 800 images for training, 100 for validation, and 100 for testing. In addition, using the Celeb-HQ dataset, the face image dataset was 30000 high-resolution images selected from the Celeb-A dataset split with the ratio (8:2) to do prerequisite for enhancing the task quality of low-resolution face images. For our final models, we use 50,000 randomly selected images from ImageNet [6] for the trainin

B. Training details and parameters

The model is trained with upscaling factor of 3, the loss function Mean Squared Error Loss, and the hyper-parameters were selected based on try-and-error and previous publications. The batch size, learning rate, momentum, and epochs were 64, 0.001, 0.9, and 100, respectively. We use the peak signal-to-noise ratio (PSNR) as the performance metric to evaluate our models.

$$PSNR = 20\log10(\frac{MAX_f}{\sqrt{MSE}})$$

f: represents the matrix data of our original image.
g: represents the matrix data of our degraded image in question.
m: represents the numbers of rows of pixels of the images and i represents the index of that row.
n: represents the number of columns of pixels of the image and j represents the index of that column.
MAX_f: is the maximum signal value that exists in our original “known to be good” image.

$$MSE = \frac{1}{m\,n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ||f(i,j) - g(i,j)||^2$$

The mean squared error (MSE) for our practical purposes allows us to compare our original image’s "true" pixel values to our degraded image. The MSE represents the average of the squares of the "errors" between our actual and noisy images. The error is the amount by which the values of the original image differ from the degraded image. The proposal is that the higher the PSNR, the better the degraded image has been reconstructed to match the original image, and the better the reconstructive algorithm. This would occur because we wish to minimize the MSE between images concerning the maximum signal value of the image. Typical values for the PSNR in lossy image and video compression are between 30 and 50 dB, provided the bit depth is 8 bits, where higher is better. The processing quality of 12-bit images is considered high when the PSNR value is 60 dB or higher [7], [8] For 16-bit data typical values for the PSNR are between 60 and 80 dB [9], [10] Acceptable values for wireless transmission quality loss are considered to be about 20 dB to 25 dB [11], [12] When you try to compute the MSE between two identical images, the value will be zero; hence, the PSNR will be undefined (division by zero). The main limitation of this metric is that it relies strictly on numeric comparison and does not consider any level of biological factors of the human visual system, such as the structural similarity index(SSIM).

C. Training and Evaluation Process Results

We trained and tested the results with some illustrations in Figure 2,3. In addition, Figure 4 shows the change in the loss function over each epoch. The value of the loss function converges very early in the first few epochs when the loss on the validator stops decreasing. However, Figure 5 shows that the PSNR value can still increase in successive epochs. Typical values for the PSNR in lossy image compression are between 30 and 50 dB, provided the bit depth is 8 bits, where higher is better. The processing quality of 12-bit images is considered high when the PSNR value is 60 dB or higher. For 16-bit data typical values for the PSNR are between 60 and 80 dB. Acceptable values for wireless transmission quality loss are considered to be about 20 dB to 25 dB

Fig. 2. Result of images tested with the weights of the data set DIV2K

Fig. 3. Result of images tested with the weights of the data set ImageNet

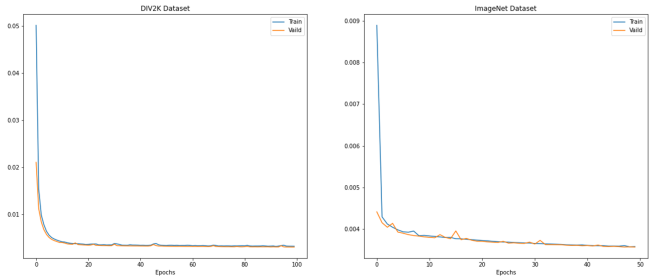

Fig. 4. The result of the loss function

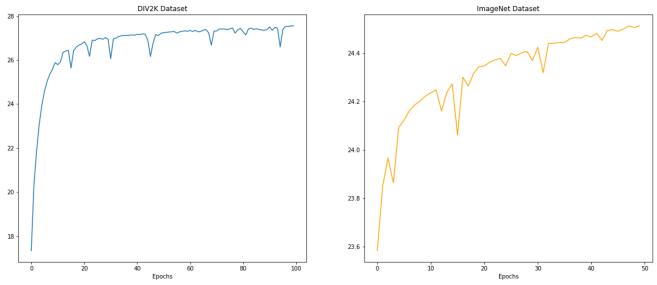

Fig. 5. The result of the PSNR Value

TABLE I
THE MEAN PSNR (DB) FOR DIFFERENT MODELS DURING TRAINING

Dataset	PSNR (Validation)
DIV2K	26.71
ImageNet	24.32

TABLE II
THE MEAN PSNR (DB) FOR DIFFERENT MODELS FOR EVALUATION

Dataset	PSNR(DIV2K)	PSNR(Celeb)	PSNR(ImageNet)
Set5	29.56	30.24	**30.31**
Set14	26.36	26.76	**26.95**
BSD100	26.33	26.54	**26.75**

IV. CONCLUSION

Through the process of researching, we have learned some more methods to improve image quality. It is, namely, going from a low-resolution (LR) image to an enhanced or improved version. We approach the convolutional neural networks (CNN) method and learn about the author's new proposal, an efficient sub-pixel convolution layer that learns an array of upscaling filters to upscale the final LR feature maps into the HR output. In the future, we will continue to explore some other more improved methods, such as applying generative models to super-resolution problems.

REFERENCES

[1] L. Zhang, H. Zhang, H. Shen, and P. Li. A super-resolution reconstruction algorithm for surveillance images. Signal Processing, 90(3):848–859, 2010.

[2] S. Peled and Y. Yeshurun. Superresolution in MRI: application to human white matter fiber tract visualization by diffusion tensor imaging. Magnetic resonance in medicine : official journal of theSociety of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine, 45(1):29–35, 2001.

[3] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. Marvao, T. Dawes, D. ORegan, and D. Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, Medical Image Computing and Computer Assisted Intervention (MICCAI), volume 8151 of LNCS, pages 9–16. 2013.

[4] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M.Mersereau. Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing, 12(5):597–606, 2003.

[5] Wenzhe Shi et al. Real-Time Single Image and Video Super-Resolution Us- ing an Efficient Sub-Pixel Convolutional Neural Network. 2016. DOI: 10. 48550 / ARXIV . 1609 . 05158. URL: https : / / arxiv . org / abs / 1609 .05158

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, pages 1–42, 2014.

[7] Faragallah, Osama S.; El-Hoseny, Heba; El-Shafai, Walid; El-Rahman, Wael Abd; El-Sayed, Hala S.; El-Rabaie, El-Sayed M.; El-Samie, Fathi E. Abd; Geweid, Gamal G. N. (2021). "A Comprehensive Survey Analysis for Present Solutions of Medical Image Fusion and Future Directions". IEEE Access. 9: 11358–11371. doi:10.1109/ACCESS.2020.3048315. ISSN 2169-3536.

[8] Chervyakov, Nikolay; Lyakhov, Pavel; Nagornov, Nikolay (2020-02-11). "Analysis of the Quantization Noise in Discrete Wavelet Transform Filters for 3D Medical Imaging". Applied Sciences. 10 (4): 1223. doi:10.3390/app10041223. ISSN 2076-3417.

[9] Welstead, Stephen T. (1999). Fractal and wavelet image compression techniques. SPIE Publication. pp. 155–156. ISBN 978-0-8194-3503-3.

[10] Raouf Hamzaoui, Dietmar Saupe (May 2006). Barni, Mauro (ed.). Fractal Image Compression. Document and Image Compression. Vol. 968. CRC Press. pp. 168–169. ISBN 9780849335563. Retrieved 5 April 2011.

[11] Thomos, N., Boulgouris, N. V., Strintzis, M. G. (2006, January). Optimized Transmission of JPEG2000 Streams Over Wireless Channels. IEEE Transactions on Image Processing , 15 (1).

[12] Xiangjun, L., Jianfei, C. Robust transmission of JPEG2000 encoded images over packet loss channels. ICME 2007 (pp. 947-950). School of Computer Engineering, Nanyang Technological University.