A Comparison Study of Deep Learning Techniques to Increase the Spatial Resolution of Photo-realistic Images

Andrew M Shackleton[1] and Abdulrahman Altahhan[2]

[1] Leeds Beckett University, Leeds, UK
Andrew.shackleton1@ntlworld.com
[2] Leeds Beckett University, Leeds, UK
A.Altahhan@leedsbeckett.ac.uk

Abstract. In this paper we present a perceptual and error-based comparison study of the efficacy of four different deep-learned super-resolution architectures, namely ESPCN, SRResNet, ProGanSR and LapSRN, all performed on photo-realistic images by a factor of 4x; adapting some of the current state-of-the-art architectures using Convolutional Neural Networks (CNNs). The resultant application and the implemented CNNs are tested with objective (Peak-Signal-to-Noise ratio and Structural Similarity Index) and perceptual metrics (Mean Opinion Score testing), to study their relative quality and implementation within the program. The results of these tests demonstrate the effectiveness of super-resolution, showing that most network implementations give an average gain of +1 to +2dB (in PSNR), and an average gain of +0.05 to +0.1 (in SSIM) over traditional Bicubic scaling. The results of the perception test also show that participants almost always prefer the images scaled using each CNN model compared to traditional Bicubic scaling. These findings also present a look into new diverging paths in super-resolution research; where the focus is now shifting from solely error-reduction, objective-based models to perceptually focused models that satisfy human perception of a high-resolution image.

1 Introduction

Traditional image scaling techniques such as nearest-neighbour, bilinear, and bicubic interpolation offer computationally quick methods of increasing the size of an image, but they do not provide any benefit to quality as they cannot construct or infer new data; able to only increase the scale of what is already present in the original image.

Nearest neighbour interpolation works by first enlarging the image by the desired factor and spreading the already available pixels within the newly defined space. The original pixels are surrounded by a 'grid' of blank space in which there are no original pixels from the image; the blank spaces are then filled by copying the 'nearest-neighbour' pixels to the blank space, turning one pixel to four identical pixels (for 4x scale). To perform bilinear interpolation, pixels are sampled in two directions. This type of scaling takes the closest 4 pixels

located diagonally into account (2x2) and takes a weighted average, as opposed to nearest-neighbours singular sample. Bicubic interpolation further considers the weighted average of the nearest 16 pixels (in a grid of 4x4), which produces an overall smoother image and reduces artefacts. Because the region of sampling is greater for this algorithm compared to others, pixels closer to the chosen interpolated pixel are given a greater weighting in the calculation.

Whilst such image resampling techniques increase the actual 'resolution' of the image when upscaling, they do not present any added detail that contributes to the increase in spatial resolution of the final image. This results in an equal or less-than equally detailed output image, such that one might refer to the output as 'blurry' when compared to a similar image of native resolution. This issue has led to the research and development of machine learned models to improve upon traditional methods of image upscaling; a method known as super-resolution.

1.1 Motivation and Rationale

Super-Resolution can have applications in surveillance, medical imaging, astronomical observation, and so on (Yue et al., 2016)[5]. Super-Resolution also has novel uses; a popular application of such techniques is upscaling textures from older video games to bring them into the modern era, as well as enhancing old low-resolution photographs, or enhancing complex drawings and diagrams. Image super-resolution by nature is an ill-posed problem as there is no true output to an image that does not have a corresponding high-resolution parent. There are a number of different approaches that have been taken using machine learning and convolutional neural networks (CNNs) for image super-resolution; such as SRCNN, SSResNet, Deep Image Prior and ESPCN. These all attempt, using different architectures, to up-scale an image while retaining/reconstructing fine image detail that is not found within the original low-resolution image (such as sharp edges on geometric shapes, or texture detail on small scale objects). Many of these methods for super resolution exist in a primitive form however; the majority being simply proposals that offer independent python command line implementations based on Linux, or working models built using and running within MATLAB.

1.2 Related Literature

ESPCN The following method by Shi et al., (2016) [4] ESPCN, uses a shallow 3-layer convolutional neural network and avoids upscaling the low-resolution input like in (Dong et al., 2015)[3]. A convolutional layer is applied directly on the low-resolution input to extract the feature maps, followed by a sub-pixel convolutional layer to upscale these feature maps to produce the super resolution output. This method differs from Dong et al, (2015)[3] in that it uses an efficient sub-pixel convolution layer instead of deconvolution layer (which recovers resolution from the max pooling layer, also known as backwards convolution). This pixel shuffle layer is faster than methods that use a deconvolution layer specifically in training, as well as being faster than methods performing

upscaling or pre-processing before convolution is applied. In Shi et al (2016)[4], ESPCN with ReLU activation trained with ImageNet data achieved significantly better performance compared to SRCNN models. Training the ESPCN model with more images saw a greater gain in PSNR than the values found with SRCNN. Interestingly, performance on this architecture is found to be high enough that it is capable of running on video without severe performance degradation.

SRResNet Another architecture by Ledig et al, (2017) [6] presents a method of Super-resolution combining error reduction focused architectures with a GAN architecture. The authors pose that while performance and accuracy of current super-resolution models are a benefit, recovering fine-detail in the image has not yet been tackled successfully. Most methods (A+, SRCNN, ESPCN, and LapSRN for example) are based on Mean Squared Error (MSE) reduction during reconstruction. While the resultant PSNR values for these techniques are high, high-frequency details are missing and the images do not give the visual perception of being high-resolution to the human eye. By combining a CNN optimised for Mean Squared Error (SRResNet) with a Generative Adversarial Network-based model (SRGAN), this problem can be overcome. This architecture sees greater gains in PSNR and SSIM over both ESPCN (Shi et al. 2016) [4] and SRCNN (Dong et al. 2015)[3], however as the authors rightfully state, that these values are not representative of the fine detail reconstruction that SRGAN provides. The authors therefore take an extra step and use Mean Opinion Score testing to quantify the super resolution capabilities of each of these models.

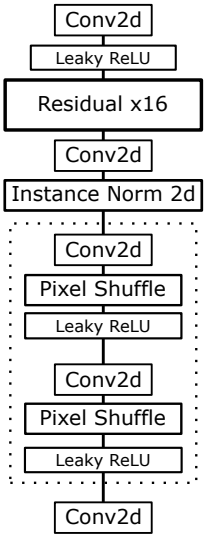

Fig. 1. The architecture of SRResNet

LapSRN The architecture by Lai et al. (2018)[7] referred to as LapSRN provides an alternative process of super-resolution using Laplacian pyramids. The authors highlight drawbacks of using pre-processing methods found in other techniques, in that they increase the computational cost unnecessarily and do not provide any additional high frequency information for a HR output. Many techniques focus around MSE loss, resulting in overly smooth images (the same low-resolution patch may have multiple high-resolution output patches in correspondence). The authors propose a progressive approach which eliminates the single up-sampling step that most other models use (SRCNN, ESPCN use direct reconstruction in a single step), to progressively reconstruct images along the network. The Laplacian Pyramid structure of this network is a key concept; where weights are shared across pyramid levels to reduce network parameters. This subsequently allows for multi-scale training for different levels of super resolution at once (2x, 4x, 8x pyramids). The authors also state the LapSRN can be easily extended to incorporate adversarial training as a part of GAN, as found in Ledig et al (2017)[6] and Wang et al (2018)[8], however this is not provided in the paper.

ProSR Taking the concept of progressive reconstruction a step further, Wang et al (2018)[8] propose an architecture that combines two methods, ProSR; a progressive method to upscale images in intermediate steps, and ProGanSR which follows the same design principle but allows for more photo-realistic results to be generated using a GAN. This diverges from other traditional methods in that it takes a progressive approach with "curriculum learning" as opposed to direct methods which upsample in a single final step. The basis of this is that the network up-samples the image in intermediate steps while the learning process increases in difficulty along with these steps. his approach shares similarity in concept with LapSRN (Lai et al, 2018)[7] due to their progressive approaches, but the authors of ProSR note that the Laplacian pyramid structure increases difficulty of optimisation and reduces performance on levels higher up the pyramid structure. The authors propose Dense Compression Units consisting of both Dense Blocks and Compression.

2 Design and Development

Neural Network development took place using Python 3 with PyTorch 1.0.The GUI was developed using Qt for Python.The four models mentioned above were chosen for implementation; ESPCN[4], SRResNet(w/o GAN)[6], LapSRN[7], ProGanSR[8]. Each has a PyTorch implementation officially provided by the author or independently implemented in Python. Each models code was further adapted to work with the GUI code to produce the resultant application.

2.1 Training

All models are trained for a desired resolution multiplier of 4x. Training was performed locally using an NVIDIA GeForce GTX 1080 Ti. Training was performed

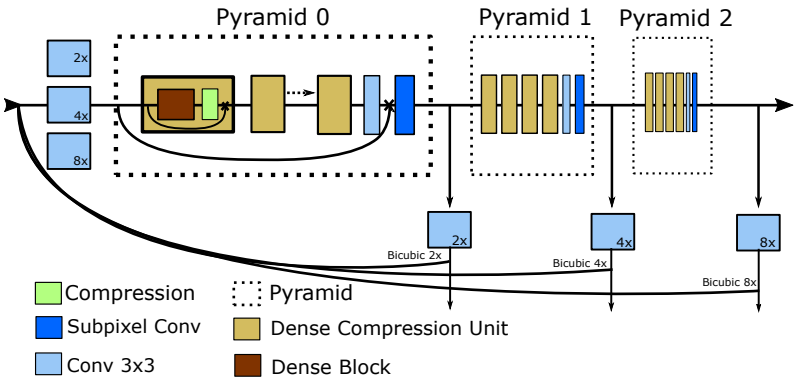

Fig. 2. The architecture of ProSR (without GAN) as found in [8]

using CUDA v9.0 to provide faster execution and training speeds. Datasets used for training include BSDS500 (Arbelaez et al., 2011)[2], DIV2K (Agustsson et al., 2017)[1]. In order to train, the data set images are first downscaled by 4x. An independent implementation of MATLAB's imresize function was used, as this provides the best results for bicubic downscaling compared to other methods found within Python. Training datasets were also augmented with random cropping, flipping, and transposing of each image. Each model was trained individually via said local machine, or via the provided model from the author for 100 Epochs.

2.2 Testing

Two tests performed, a test validating output images from the application using PSNR and SSIM with a python script; and another evaluating human perception on the same set of test images to judge percieved quality via a survey. The Python implementations are not a perfect recreation of the models described in the relevant literature, as such the results for PSNR found within the literature are typically greater than those of the python versions when tested with similar images. The PSNR and SSIM testing for images within relevant literature is performed on the Y channel, and so for this test the image channels are separated, and testing is performed on the Y channel.

2.3 Similar Work

Applications such as Waifu2x and Topaz A.I. Gigapixel perform similar functions to the proposed application; Waifu2x works best on non-photoreal images such as drawings and cartoons at up to 2x factor scaling based around the (no longer state-of-the-art) SRCNN architecture, and Gigapixel is a proprietary piece of software in which the algorithms used are unknown. This prototype application differs from both of these in that it is a free application that makes practical use

of more up-to-date, publicly available image scaling networks in a user-friendly manner through a GUI; by compiling the current and more recent state-of-the-art models together in the application, instead of a single model used in either program mentioned.

A qualitative survey was created to test the results of the networks used in the application on human perception. The same 5 images from the previous test were used, each run through the application with the downscale option selected. The image scaling options for each image were; Bicubic, ESPCN, SRResNet, LapSRN, and ProGanSR. This resulted in a total of 25 images that were given to participants. 20 Participants responded to the survey. Participants were asked to rank the images in order of visual quality and realism, where a rank of 1 is the highest quality and most visually pleasing image, and a rank of 5 is the lowest quality and least visually pleasing image. Participants are not given the Ground Truth image as reference, and the names of each model are not divulged.

3 Results and Evaluation

3.1 PSNR and SSIM

On the 'statuette' image set, bicubic scaling appears to give the highest value results for both SSIM and PSNR. It is unclear why this happens, but it is only the case on this image. This example is some justification as to why PSNR and SSIM alone are not a concrete metrics for judging image quality. SRResNet has the most occurrences of the highest values of PSNR and SSIM on the 5 sets of test images, in both test runs. SRResNet also outperforms ProSR when tested against these metrics, which is to be expected. ESPCN falls behind bicubic scaling in many of these test cases, in both SSIM and PSNR. The majority of results gathered in this test show that error-focused architectures do outperform both traditional scaling methods and perceptual-focused architectures. It is clear when looking at the images PSNR and SSIM alone do not provide the optimal method for judging the visual quality of a super-resolved image.

Table 1. The results of the PSNR and SSIM Test on a custom set of 5 images.

Image Set	Test	ProSR	SRResNet	LapSRN	ESPCN	Bicubic
Sign	PSNR (dB)	27.674	27.792	28.006	24.283	24.844
-	SSIM	0.888	0.883	0.885	0.696	0.787
Dog	PSNR (dB)	25.325	27.146	25.478	25.738	26.438
-	SSIM	0.753	0.800	0.791	0.741	0.776
Statuette	PSNR (dB)	24.300	26.054	25.007	24.004	27.766
-	SSIM	0.836	0.826	0.821	0.813	0.856
Bluebell	PSNR (dB)	22.604	23.815	22.767	21.235	22.061
-	SSIM	0.746	0.788	0.765	0.689	0.693
View	PSNR (dB)	21.950	23.504	22.735	22.367	21.083
-	SSIM	0.644	0.704	0.697	0.663	0.574

3.2 Perceptual Study

ESPCN was ranked lowest of the tested group, only barely contesting bicubic scaling in most cases. Looking at the images, there is only a minute difference between ESPCN and Bicubic, with ESPCN looking slightly sharper than the Bicubic images. As expected, Bicubic scaling provides the worst quality results and this is reflected in the participants' response. In 3 out of 5 test cases, Bicubic scaling is ranked higher than or equal to ESPCN. Therefore, it can be determined from this that ESPCN provides an alternative to Bicubic scaling, not a true replacement as was expected with the other models.

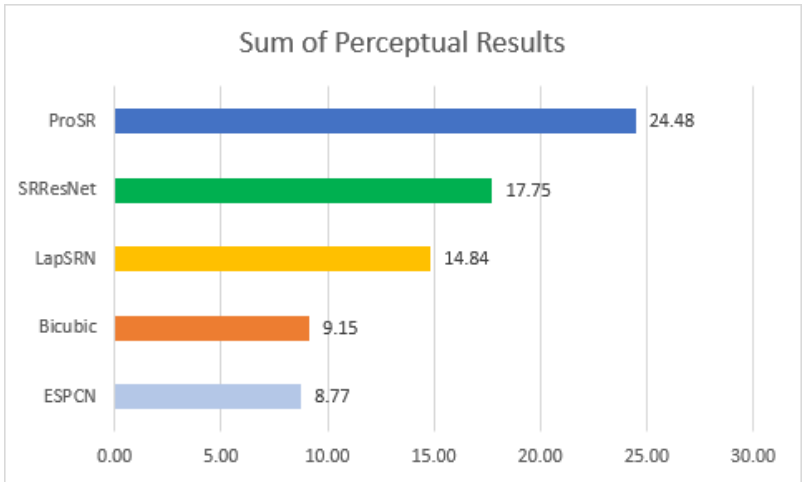

Fig. 3. A bar graph showing the total aggregate results of perceptual testing on the same 5 images.

4 Conclusion

This paper has presented a study and prototype implementation of state-of-the-art techniques for super-resolution within a x64 and Unix compatible application, allowing for any user to upscale a desired image using these techniques without the need for knowledge of programming or deep learning. Through testing, we find that the error-focused architectures (based around PSNR, SSIM, and Mean-Square Error testing) provide some excellent techniques for objective super-resolution, but result in often murky and smudged images. The perceptually-focused architectures, a more recent development making use of adversarial networks, give promising results that better represent true, high-resolution images able to fool the human perception. In the context of applications of these models, perceptual approaches that hallucinate finer detail might be less suited for medical applications or surveillance because the data they produce is technically not present within the original image, giving an advantage to error-focused

approaches. Perceptual approaches may therefore be more useful for applications that do not specifically require the content of the images to be accurate (such as personal photos). This gives merit to the suggestion that one path for super-resolution is not necessarily better than another.

5 Further Work

The application can be extended to work on other forms of media with further training, such as drawings or animations. Re-training each network with more data is another viable further step, in order to provide more optimal results on photographic images. The tool could also be extended to process larger images in a memory-saving manner, as larger images currently require high amounts of VRAM to process. Further optimization techniques can be utilised to streamline the process and make it more practical for real time and networking applications. A further study could be conducted to compare the relative quality of each architecture with and without GAN; current testing only shows that human participants prefer GAN-processed images, but not which GAN architecture specifically.

References

1. Agustsson, E. and Timofte, R. (2017). NTIRE 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 126-135.
2. Arbelaez, P., Maire, M., Fowlkes, C, and Malik, J. (2011) Contour Detection and Hierarchical Image Segmentation. IEEE TPAMI May. 33(5), pp. 898-916
3. Dong, C., Change Loy, C., He., K. and Tang, X. (2015) Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2), June, pp. 295 – 307.
4. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A., Bishop, R., Rueckert, D. and Wang, Z. (2016) Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. IEEE.
5. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H. and Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. Signal Processing, 128, pp.389-408.
6. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W. (2017) Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-16, 2017, Honolulu, HI, USA. IEEE.
7. Lai, W., Huang, J., Ahuja, N. and Yang, M. (2018) Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
8. Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O. and Schroers, C. (2018) A Fully Progressive Approach to Single-Image Super-Resolution. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). June 18-22, 2018, Salt Lake City, UT, USA. IEEE.