ON HOW TO COMBINE SINGLE IMAGE SUPER-RESOLUTION ALGORITHMS

Robert STANCA[30]
Eduard-Marius COJOCEA[31]
Cristian AVATAVULUI[32]
Costin-Anton BOIANGIU[33]

Abstract: *In this paper, we present several models for super-resolution and how the performance can be increased on the DIV2K dataset using voting techniques. Combining various models and voting techniques we show that some state-of-the-art algorithms can be furtherly improved. At the same time, using voting or multiple experts' decision, we achieve more robust systems, which have stable performance, high subjective evaluation and also encouraging users' confidence. The results proved that the proposed method delivers accurately-enough results while ensuring strong reliability.*

Keywords: *super-resolution, voting systems, deep learning, convolutional neural networks, generative adversarial networks*

1. Introduction

In the last decade, the field of Computer Vision has grown exponentially both in popularity and in high-performance models, offering a new way of approaching problems involving image processing. Such a problem is super-resolution, which requires the transformation of the images from a lower resolution to a higher resolution. Super-resolution has significant applications in game development, multimedia content creation, advertising, medical imaging, security image analysis, autonomous vehicles and many others.

The super-resolution process involves transforming an image from a lower resolution to a higher resolution by supplementing it with information that must be approximated based on the lower quality image. Thus, Machine Learning algorithms are well fitted for this problem, making it possible to learn better approximations than the heuristic approaches. Deep learning techniques are a good choice since they can extract more complex and finer information in the learning

[30] Engineer, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, robert.stanca@stud.acs.upb.ro
[31] PhD Student, Eng., Research and Development Department, OpenGov Ltd., Bucharest 011054, Romania; marius.cojocea@opengov.ro
[32] PhD Student, Eng., University Politehnica of Bucharest, cristianavatavului@gmail.com
[33] Professor, PhD Eng., Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, costin.boiangiu@cs.pub.ro

process when compared to shallow networks, which usually are able to extract only coarse or superficial information from images. Convolutional Neural Networks (CNN) and Generative Adversarial Network (GAN) are the two main model classes that are used to obtain excellent details for a given image.

The main advantage of using Machine Learning for this task is the fact that such models do not require manually extracted features, CNNs and GANs being able to act as automatic feature extractors before they generate the results. This makes it possible for identifying filters and features that usually perform better than those extracted manually. Also, describing mathematically an object, an animal, a person, or how clear an image is to the human eye is very hard to do, which means that heuristic methods can only go so far, being dependent on the features extracted using human insights. But a CNN can extract such a description, which, if visualized by humans, might seem not related to the task. Despite this, CNNs can learn such descriptions during training. Sometimes, they even overcome human performance in some tasks.

Some of the most popular methods approaching the super-resolution problem are:

- ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [1]
- Residual Dense Network for Image Super-Resolution [2]
- Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [3]

In this article, we propose a voting system based on these three methods in order to obtain an improved and robust super-resolution technique.

2. Related work

Initially, a heuristic solution for super-resolution was represented by bilinear interpolation algorithms [3], which allow the estimation of pixel values based on nearby pixels. Interpolation works using known data to estimate values at unknown points. In image interpolation, each pixel's intensity is estimated based on the values at the surrounding pixels. This operation is computationally intensive and the results are not the greatest. Also, bilinear interpolation introduces a blur effect that comes from the linear equations used for interpolation.

Dong. [4] presents one of the earliest solutions based on Deep Learning to solve super-resolution. The proposed architecture is called SRCNN and learns to map images from a lower resolution to a higher resolution using an end-to-end learning system where all essential features are automatically discovered, between stages data entry and output stage.

Later, a new series of approaches were presented using different architectures. The most significant are residual learning architectures [5], Laplacian pyramid structure [6], residual blocks [7], dense networks connected [8], and densely connected residual networks [9].

Generative adversarial network (GAN) architectures are also used in super-resolution. Unlike most networks, GANs use a pair of two networks to learn. This mechanism involves a generative network and a discriminative network, each having a crucial role in this architecture. The role of the generative network is to generate the best possible candidates to mislead the discriminative network into considering the candidate is real and not generated. The generative network is receiving feedback from the discriminative network. On the other hand, the discriminative network has the role of detecting candidates that were poorly generated, which aren't indistinguishable from real candidates. The two networks use learning similar to agent learning, where each component reacts to the feedback received from the other component.

GAN method can be successfully applied in the generation of photo-realistic images [10]. Also, in this sub-domain were introduced various mechanisms to improve the performance of GAN networks such as: using Wasserstein distances, regularizing the discriminative network through a technique called gradient clipping that limits the maximum value of the gradient.

2.1 Problem motivation

The interest in super-resolution research has grown with the expansion of online services and the internet. Even though camera resolution has increased significantly in recent years, super-resolution is a technique that can benefit from reducing the size of transferred data by sending a downscaled image and reconstructing it with a certain degree of accuracy to the recipient.

Also, developing multimedia content for various resolutions is an expensive process and a possible solution is to develop at a minimum resolution and turn it into a higher resolution using deep learning techniques only when needed.

Other applications include interpreting data from surveillance video cameras, which provide lower resolution to make it easier to store this data for later use. The size of files saved by a surveillance camera is directly proportional to the quality at which they are recorded. Because the storage solutions are generally limited, they record video content in 720p or smaller format, which does not focus on the details in the frame.

Thus, the development of these types of solutions would bring a significant addition to the quality of the images generated at a lower resolution by using machine learning techniques.

3. Proposed methods

In this paper, we will focus on combining the results obtained using well-known algorithms from the super-resolution domain to achieve better results compared to using these algorithms individually.

To combine the results obtained from the algorithms, we will use different voting techniques to determine the best result. The results of the voting algorithms will be

presented in the following sections, along with some insights and observations regarding the voting process.

In the following subsections, we will present the three Deep Neural Networks used in our experiments.

Residual Dense Network - RDN

This architecture brings significant improvements compared to the previous versions, by introducing new components named Dense Residual Blocks.

The architecture is presented in figure 1. It is composed of the following:

- shallow feature extraction net (SFENet)
- residual dense blocks (RDBs)
- dense feature fusion (DFF)
- up-sampling net (UPNet)

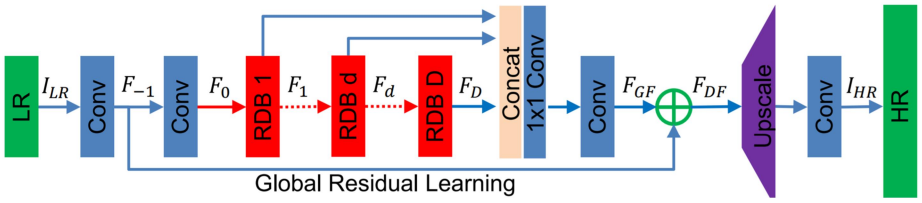

Figure 1. The proposed architecture of Residual Dense Network in [2]. Image taken from [2]

The first part (SFENet) is using convolution operations to extract shallow layer features (F-1, F0). These are afterward introduced in the residual dense blocks to generate another set of features.

These blocks are presented in figure 2. They are made of convolution operations and the ReLU (rectified linear units) activation function. The general convolution blocks differ from the ones used in this paper due to the usage of dense links. A final feature is generated by using dense links with all the features generated by the convolution operations. Moreover, this final feature is used along with the input feature to generate a global feature at the block level.

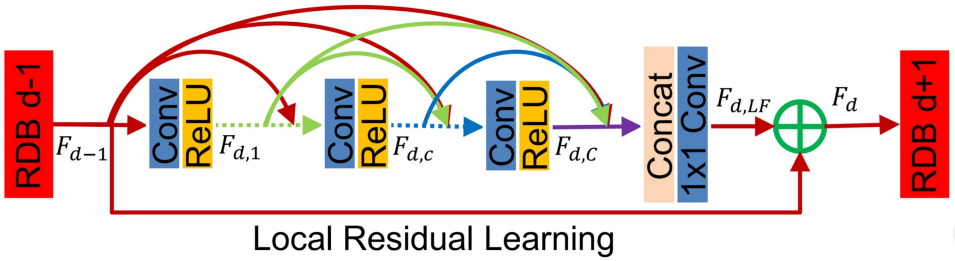

Figure 2. Residual Dense Block architecture. Image taken from [2]

Being composed of multiple residual dense blocks, each block generates a global feature, which is then used to generate a feature map.

This feature map is generated by concatenating all the global features generated until that moment. Also, the upscale network uses as input both the previously generated feature map and the shallow layer features generated at the beginning of the process.

In this paper, convolutions of size 3x3 are used to generate global and local features. Convolutions of size 1x1 are used to aggregate all the generated features.

Super-Resolution Generative Adversarial Network - SRGAN

The paper proposes a loss function called perceptual loss. It is made of two-loss functions: adversarial loss and content loss. Adversarial loss is used to differentiate between photo-realistic images and images created by the generative network.

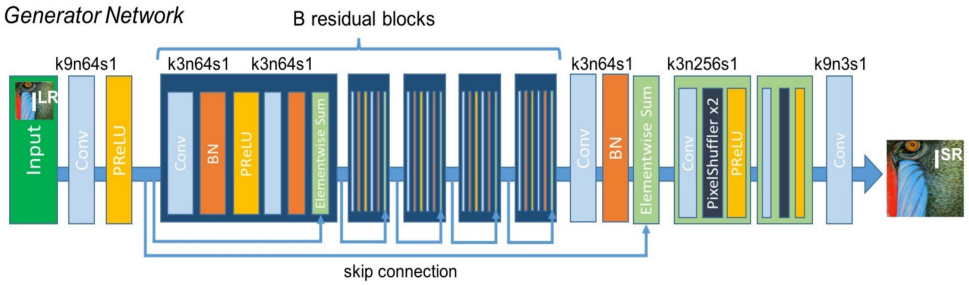

Figure 3. Generator Network. Image taken from [3]

Also, the authors note that perceptual loss is an improvement over pixel similarity, thus allowing the network to generate textures similar to photo-realistic ones.

The generator has a similar structure with the architecture presented in [2], both using residual blocks. The residual blocks described in this paper are more complex than the residual blocks from [2], and they are made of the following:

- Convolution operation
- Batch normalization
- Parametric ReLU
- Batch normalization
- ElementWiseSum

In contrast with the previous paper, the authors use two batch normalization operations, two convolutions, and an elementwise sum instead of concatenation to aggregate the features for each block.

The generative network receives as image input a lower resolution image. It tries, using the feedback received from the discriminator, to improve the upscale process.

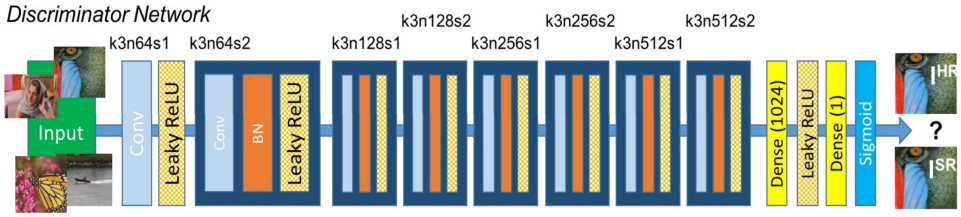

Figure 4. Discriminator Network. Image taken from [3]

However, the discriminator structure is a classic one composed of convolution operations, batch normalization, and the parametric ReLU activation function. This network tries to determine whether the input image is received as a high-resolution image or generated using super-resolution.

$$\boldsymbol{l^{SR} = l_X^{SR} + 10^{-3} l_{Gen}^{SR}} \tag{1}$$

Where: l^{SR} is the perceptual loss, l_X^{SR} the content loss, $10^{-3} l_{Gen}^{SR}$ the adversarial loss.

The main contributions of this paper in the super-resolution domain are the introduction of a loss that is able to generate textures that are similar to photo-realistic textures.

$$l_{MSE}^{SR} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left(I_{x,y}^{H\,R} - G_{\theta G}(I^{LR})_{x,y} \right)^2 \tag{2}$$

Enhanced Super-Resolution Generative Adversarial Network - ESRGAN

The system presented in the [1] brings some significant improvements over SRGAN by using the concepts previously mentioned in the GAN framework.

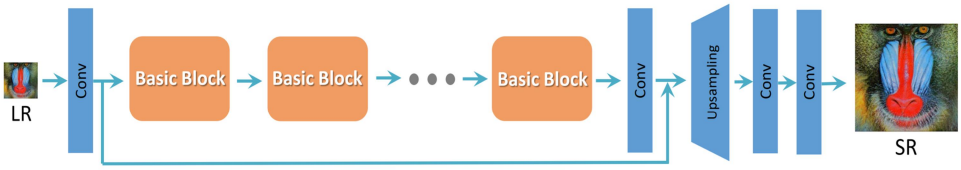

Figure 5. Basic architecture SRResNET. Image taken from [1]

The authors keep the general structure of the generator but replace the residual blocks with a new type of residual blocks called RRDB. They keep the convolution operations and replace parametric ReLU with Leaky ReLU but add block-level residues using a Beta factor to scale these residues.

Significant changes have also been in the structure of the discriminative network. So far, a discriminative determined the probability that an image was real and natural, and the authors introduce a new type of relativistic discriminative that brings improvements for specific tasks.

The relativistic discriminative is based on the following idea: instead of determining whether an image is natural or not, they decide whether an image is more natural than a fake one. This fake image is generated using the average of the values of the predicted fake images, thus introducing a comparative degree between real and fake images.

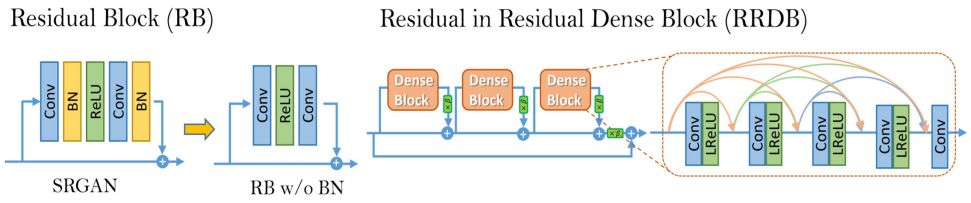

Figure 6. Residual Block and Residual in Residual Dense Block. Image taken from [1]

Another remark in the previous article is the order of use of the characteristics and activation. They argue that the brightness suffers when first using the activation followed by characteristic and propose the parameterization of the loss.

$$L_G = L_{percep} + \lambda L_G^{Ra} + \eta L_1 \tag{3}$$

Voting algorithms

The previous algorithms are used to generate the most realistic images that we further use for the voting algorithm to determine the best representation.

The algorithms chose for voting are as follows:

- Average Voting (pixel-wise)
- Furthest away – the distances between the pixel values for each algorithm are calculated, and the one with the maximum distance is chosen
- Noise estimation – the image is transformed into BW, and it is estimated how much noise is in the image, the image with the least noise is chosen (patch-wise)

4. Experiments

Dataset

DIV2K [11] is the most widely used data set to evaluate the performances of super-resolution algorithms. It contains 800 high-resolution images for the training stage, 100 high-resolution images for validation, and 100 high-resolution images for the test stage.

For each high-resolution image, it also offers a variant whose quality has been reduced using one of the following procedures:

- Bicubic interpolation
- Unknown operator – the procedures that generated the low-quality image are kept hidden in order to avoid finding a method that has results only for bicubic interpolation

In this dataset, there are also three scaling versions (x2, x3, x4) that determine the final size of the picture with reduced image quality.

For the experiments in the article, we used the x4 scaling and the unknown operations to measure the performance of the voting algorithm.

Evaluation methods

To evaluate the quality of images, we use the most common metric in specialized publications, and we compare it with the results obtained by the individual running of the Deep Learning algorithms.

We can measure the success of a network by finding out how well it reduces the mean squared error (MSE) between the output pixels and the original version.

The best result is an MSE of 0, which means that the original high-resolution image and the high-resolution version generated by the network are identical.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(I_{(i)} - \widehat{I_{(i)}}\right)^2 \tag{4}$$

$$PSNR = 10 * log_{10}\left(\frac{L^2}{MSE}\right) \tag{5}$$

The purpose of the PSNR equation is to calculate the compensation between the MSE and the maximum value of the pixels. A higher PSNR represents high quality generated images. It is worth noting that PSNR is not a completely objective metric since PSNR-oriented approaches tend to output images overly smoothed, lacking

enough high-frequency details. Despite this, it is still a useful metric, which can measure how well an image is being upscaled.

5. Results

In Table 1 we present the PSNR values obtained for the three individual methods and then the values obtained by using the proposed voting strategies. We can observe that the average voting system performs worse than the individual methods. The same is true for the furthest away strategy. But the noise estimation strategy is performing better than RDN and SRGAN, although still is lacking in comparison to ESRGAN. Despite this, this strategy offered lower fluctuation in performance, making the system to be more robust. Thus, it is worthwhile taking into consideration a voting system when approaching super-resolution, which is harder to objectively describe than other Computer Vision problems.

Algorithm	**PSNR**
RDN	19.17
SRGAN	19.36
ESRGAN	19.85
Average Voting	18.98
Furthest away	19.09
Noise estimation	19.37

Table 1. The PSNR values obtained by using the three individual methods and the three proposed voting systems.

A snapshot of some visual results is presented in Figure 7.

High Resolution RDN SRGAN ESRGAN

Average Furthest Away Noise estimation

Figure 7. The original high-resolution image and the upscaled images obtained by using the three individual methods and the three proposed voting systems.

6. Conclusion

In this article, we tested various deep learning models that solve the problem of super-resolution and we used voting algorithms to improve their performance. Since many models used for super-resolution achieve high PSNR, but they generated images still lacking for the human eye, we can conclude that this problem has a visual perception component, represented by the subjective evaluation of human observers. Since a perfect recreation of the downscaled image is unlikely, it is necessary to take into consideration the human perspective of the results, since for practical purposes, it is useless for an image to have high PSNR and low subjective quality (as determined by human observers). Thus, results should always be correlated with human observations.

In the near future, the current super-resolution approach will be integrated alongside other voting-based approaches [12-14] in a fully-unsupervised processing pipeline destined for analysis and processing of image documents.

Acknowledgement

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689 / „Lib2Life- Revitalizarea bibliotecilor și a patrimoniului cultural prin tehnologii avansate” / "Revitalizing Libraries and Cultural Heritage through Advanced Technologies", within PNCDI III

7. References

[1] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018. *ESRGAN: Enhanced super-resolution generative adversarial networks*, Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[2] Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y., *Residual dense network for image super-resolution*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472-2481, 2018.

[3] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., *Photo-realistic single image super-resolution using a generative adversarial network*, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681-4690, 2017.

[4] Dong, C., Loy, C.C., He, K. and Tang, X., *Image super-resolution using deep convolutional networks*, IEEE transactions on pattern analysis and machine intelligence, 38(2), pp.295-307, 2015.

[5] Kim, J., Kwon Lee, J., Mu Lee, K., *Accurate image super-resolution using very deep convolutional networks*, CVPR, 2016.

[6] Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H, *Deep Laplacian pyramid networks for fast and accurate super-resolution*, CVPR, 2017.

[7] He, K., Zhang, X., Ren, S. and Sun, J., *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016

[8] Zhang, Y., Tian, Y., Kong, Y., Zhong, B. and Fu, Y., *Residual dense network for image super-resolution*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472-2481, 2018.

[9] Tai, Y., Yang, J., Liu, X. and Xu, C., *Memnet: A persistent memory network for image restoration*, Proceedings of the IEEE international conference on computer vision, pp. 4539-4547, 2017.

[10] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., *Photo-realistic single image super-resolution using a generative adversarial network*, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681-4690, 2017.

[11] Agustsson, Eirikur and Timofte, Radu*, NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study*, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, July 2017.

[12] Costin-Anton Boiangiu, Radu Ioanitescu, Razvan-Costin Dragomir, *Voting-Based OCR System*, The Journal of Information Systems & Operations Management, Vol. 10, No. 2, 2016, pp. 470-486.

[13] Costin-Anton Boiangiu, Mihai Simion, Vlad Lionte, Zaharescu Mihai – *Voting Based Image Binarization*, The Journal of Information Systems & Operations Management, Vol. 8, No. 2, 2014, pp. 343-351.

[14] Costin-Anton Boiangiu, Paul Boglis, Georgiana Simion, Radu Ioanitescu, *Voting-Based Layout Analysis*, The Journal of Information Systems & Operations Management, Vol. 8, No. 1, 2014, pp. 39-47.