"Zero-Shot" Super Resolution Report

Benichou Yaniv, Bonnefoy Nicolas, Dahy Simon, Guckert Mathis

March 2024

Abstract

"Zero-Shot" Super-Resolution (ZSSR) [3] introduces a new unsupervised CNN-based method for single-image super-resolution. An image-specific CNN is trained on the low-resolution input image, and learns to recover the original resolution from downscaled versions of the image generated by data augmentation, by taking advantage of natural regularities. This process works well in scenarios with non-ideal downscaling kernels and poor-quality low-resolution images (e.g. historical images, smartphone shots...), and outperforms state of the art methods on this kind of images.

1 Context and interest of the approach

1.1 Context

Published in 2018, this paper follows suit to a range of results attained in the field of super-resolution (SR) using deep learning neural networks [2]. Those methods were able to overperform previous methods by a large margin. Nevertheless, what makes their strength is also their main weakness: they are trained on large datasets, requiring a very important training time. It means that albeit particularly efficient on images similar to those in the training set, those models face difficulties in handling images whose degradation pattern follows another pattern than those featured in their training. Indeed, such training sets are often constructed with a predefined specific downgrading kernel without distracting artifacts and with a fixed SR scale factor. They therefore greatly lack in adaptability, even more so when we consider that naturally obtained low resolution (LR) images don't follow such clear artificial downscaling patterns. It means that apart from artificially downgraded images, the usefulness of those models is limited.

1.2 Interest of the Approach

Taking the opposite approach, this paper aims to conceive an unsupervised CNN-based Super Resolution method, which would be a first according to the

authors. Such an approach tackles a lot of the issues aforementioned. First of all, given that the training is specific for each image, there is no dependance to previously chosen downgrading patterns, which makes this method better suited to manage "naturally" obtained LR images. As such, it is able to generalize the performance to cases where the downgrading pattern is unknown and often non ideal, precisely where state of art methods still face some difficulties. Although it needs to be trained each time we send it a new entry image, this training time stays reasonable (a few minutes without GPU), while avoiding a heavy pre-training required for the other methods. Furthermore this Zero-Shot method takes advantage of the internal recurrences that characterizes natural images [1, 4]. Thanks to cross-scale internal recurrence, it is able to mobilize internal image statistics, learning from similar patterns in the rest of the image.

2 Algorithm description

The algorithm described in the paper, by definition, only consists in the training of an image-specific convolutional neural network. We therefore detail in this part the construction of the image-specific dataset, the definition of the CNN and its training algorithm.

2.1 Building the image specific Dataset

To train the convolutional neural network to generate a higher resolution of itself on more than the test image, we use data-augmentation. The training set is built by first downscaling the LR image I into smaller versions and then downscaling these new images again by the SR factor set in parameter. The goal is to learn to reconstruct all these pairs (perform SR with the same scale factor we will apply to the test image). To expand even more the dataset, we apply four rotations and two mirrors to each pair, creating 8 times more images.

2.2 Characteristics of the convolutional neural network (CNN)

The image-specific neural network is defined as follow:

- input interpolated to have the same size as the output size,
- composed of 8 hidden layers, 64 channels with ReLU activations,
- minimize the L_1 loss, using an Adam optimizer,
- initial learning rate of 10^{-3}, divided by 10 if when making a linear fit of the reconstruction error, if the standard deviation $\sigma \geq \alpha * \text{slope}$ (until a minimal value of 10^{-6}) for which the algorithm stops.

2.3 Training algorithm

To be more precise on the training of the CNN, it learns the residual between the interpolated LR and high resolution (HR) parent (the difference between the two). To do this, at each iteration, we choose a single father-son pair with a probability depending on the size of the HR father image (the higher the size-ratio between HR father image and the test image, the higher the probability to be chosen). To speed the training process, instead of working on the whole pairs, a random crop of fixed size (128x128 by default) is made, and the CNN learns on these small patches. Finally, when applying the CNN to the test image, we in fact apply it on 8 different versions of it, generated with rotations and mirrors, and taking the median in the end.

3 Numerical illustration

We started by updating and "cleaning" the code written by the authors (repository link: `https://github.com/assafshocher/ZSSR`), that is refactoring the code for Python 3 and solving library deprecation issues. We tested the model described in the paper on several images, using the default settings. Here are some of the most important simulation parameters:

- scale factor: 2,
- maximal number of iterations: 3000,
- start learning rate: 10^{-3},
- downscaling kernel: bicubic.

To assess the similarity between the images, we used two metrics:

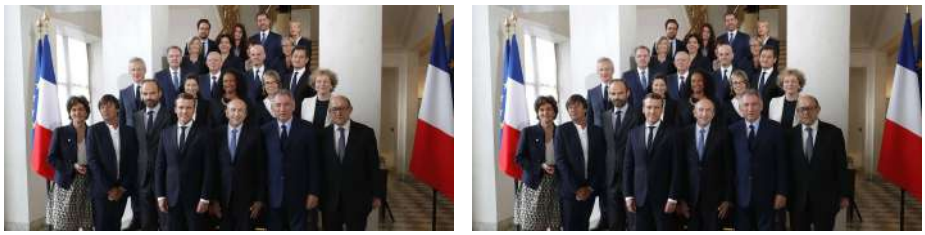
- the Peak Signal-to-Noise Ratio (PSNR):

$$PSNR = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{1}$$

  where MAX is the maximum possible pixel value of the original image, and MSE is the mean squared error between the original and the reconstructed image. The higher the PSNR (in dB), the better the reconstruction.

- the Structural Similarity Index (SSIM): the structural similarity is a classical metric that allows to capture more general differences than simply pixel by pixel differences. It is computed on several windows of fixed size, using mostly the mean and covariance of the pixels in these windows. The SSIM is a value between -1 and 1, where 1 means that the two images are identical.

In the examples in the paper, the authors manage to achieve PSNRs between 25 and 27, and SSIM ratios around 0.8. We tested the model on images by first downscaling the ground truth by a factor 2 (using by default a bi-cubic interpolation), and then applying ZSSR, allowing us to compare the ground truth and the augmented image:



(a) Ground Truth (b) ZSSR

Figure 1: Original image and reconstructed image using ZSSR

For the first image (figure 1), we see that the reconstruction is quite neat, and the difference is not obvious at this scale. The obtained PSNR is 32.7 and the SSIM is 0.85.

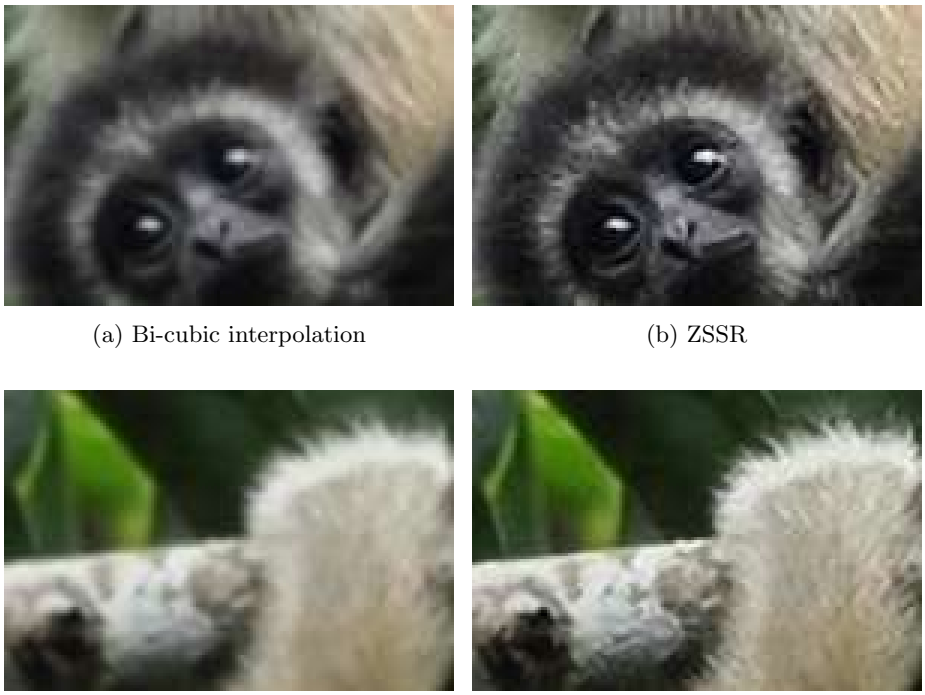
To better illustrate the impact of ZSSR, we tested the model on an image with more details (a picture of gibbons).

(a) Ground Truth (b) ZSSR

Figure 2: Original image and reconstructed image using ZSSR

For the image tested in figure 2, we obtained a PSNR of 38.6 and a SSIM of 0.96, which are very good results.

To emphasize the impact of the algorithm, we also show, on figure 3, the difference between the image upscaled using bi-cubic interpolation and the image upscaled using ZSSR.



(a) Bi-cubic interpolation (b) ZSSR

(c) Bi-cubic interpolation (d) ZSSR

Figure 3: Comparison of patches between bi-cubic interpolation and ZSSR

Our simulations remain limited compared to the original paper. Indeed, when applying the mode on the images used in the paper (mostly the historical ones), we obtained poor results that were not very interesting to develop here. We can particularly mention the fact that the authors emphasize how combining the ZSSR algorithm with an estimation of the downscaling kernel can improve the results, but no such estimation was implemented in the code we used.

4 Conclusion

This zero-shot method explores a new approach to super-resolution problems. It is able to outperform standard deep learning methods for naturally downgraded images, by taking advantage of the regularities present in natural images. It does not require a heavy pre-training, and is able to adapt to the specificities of each image. However, unlike other pre-trained models, it requires a training time for each new image, which can be a limitation in some cases.

References

[1] S. B. D. Glasner and M. Irani. Super-resolution from a single image. 2009.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. July 2017.

[3] A. Shocker, N. Cohen, and M. Irani. "zero-shot" super-resolution using deep internal learning. June 2018.

[4] M. Zontak and M. Irani. Internal statistics of a single natural image. June 2011.