SRGAN for Super-Resolving Low-Resolution Food Images

Yudai Nagano
The University of Tokyo
nagano@nlab.ci.i.u-tokyo.ac.jp

Yohei Kikuta
Cookpad Inc.
yohei-kikuta@cookpad.com

ABSTRACT

Single image super-resolution, especially SRGAN, can generate photorealistic images from down-sampled images. However, it is difficult to super-resolve originally low resolution images that contain some artifacts and were taken many years ago. In this paper, we focus on the food domain because it is useful for our recipe-based web service if we can create better looking super-resolved images without losing content information. Based on the observation that SRGAN learns how to restore realistic high-resolution images from down-sampled ones, we propose two approaches. The first one is a down-sampling method using noise injection to create desirable low-resolution images from high-resolution ones for model training. The second one is to train models for each target domain: we use the beef, bread, chicken and pound cake categories in our experiments. We also propose a novel evaluation method, Xception score. Compared with existing methods using qualitative and quantitative experiments, we find the proposed methods can generate more realistic super-resolved images.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; *Reconstruction*; Neural networks;

KEYWORDS

Super resolution, Neural networks

ACM Reference Format:
Yudai Nagano and Yohei Kikuta. 2018. SRGAN for Super-Resolving Low-Resolution Food Images. In *CEA/MADiMa'18: Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management in conjunction with the 27th International Joint Conference on Artificial Intelligence IJCAI, July 15, 2018, Mässvägen, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3230519.3230587

1 INTRODUCTION

1.1 Super-resolution with deep learning

Super resolution is one of the research fields that has achieved remarkable developments thanks to deep learning. SRGAN [3] has shown outstanding performances over other architectures in MOS testing, which is based on human senses.

In this paper, we aim to restore the original attractive texture of food by applying super-resolution processing to low-resolution images that were published in the past. In particular, we examine and experiment with how to bridge the apparent difference between a low-resolution image posted in the past and recent high-resolution images and resolution, and clarify the possibility of application to actually published images.

SRGAN, which is an existing method, is suitable for our objective because it is capable of performing 4 × super-resolution and is better at restoring the appearance according to human perception than pixel units based measures. However, because SRGAN is trained with data consisting of a reduced image constructed from a very sharp high-resolution images, there are cases wherein it is not possible to restore a desirable texture, even if super-resolution is performed on an image with a low resolution, as illustrated in Figure 1.

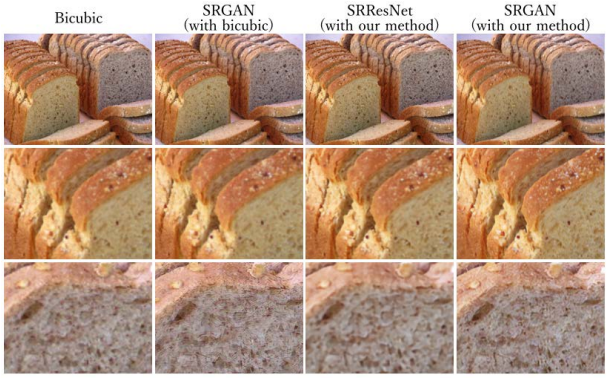

Figure 1: Comparing results of a super-resolved bread image (4x upscaling). The image is a Creative Commons image. (http://flic.kr/p/4pizmd).

2 BACKGROUND

2.1 Recipe images and super-resolution

Photos of cooking are one important factor for attractive cooking recipes on a website. However, in our web service Cookpad, images are a mix of low-resolution images that were taken about 20 years ago and recent high-resolution images. The images of recipes posted in the early years of the service have a resolution that is a fraction of the resolution available in recent years, or they include noticeable camera noise. To solve this problem, our aim is to super-resolve low-resolution cooking images.

2.2 SRResNet and SRGAN

To compare the outputs of non-adversarial and adversarial models, we experimented with two models called SRResNet and SRGAN.

SRResNet is superior to existing super-resolution methods such as the bicubic method with respect to image quality evaluation indices such as PSNR, and SSIM. It is a model that is trained without

CEA/MADiMa'18, July 15, 2018, Mässvägen, Stockholm, Sweden

ACM ISBN 978-1-4503-6537-6/18/07.
https://doi.org/10.1145/3230519.3230587

the GAN structure in SRGAN. In this experiment, we trained SRResNet using only the MSE loss without perceptual loss.

Despite the low score of PSNR and SSIM, SRGAN can generate photorealistic images from a low-resolution single image. SRGAN's essential advantage is that it considers perceptual similarity loss and adversarial loss. First, perceptual similarity is considered by comparing the pixel wise high-level features extracted from the VGG's middle layer outputs. Second, the discriminator distinguishes between true high-resolution images and super-resolved images generated from generators, and these competitive learning procedures makes it possible to generate aesthetically pleasing images for humans.

We hypothesized that SRGAN could generate perceptually attractive images, especially food cooking images.

2.3 Datasets

DIV2K dataset [1] This dataset contains pairs of high- and low-resolution images from various categories, such as buildings, insects, mountains, etc. The number of image pairs used for training was 800, and 100 image pairs were reserved for testing.

Our dataset We gathered high-resolution recipe images from Cookpad so that the number of images was the same as that of the DIV2K dataset. For our dataset, we manually selected appropriate images and excluded duplications with insufficient resolution.

Images were collected from four categories: chicken, beef, bread, and pound cake. In addition, we defined two types of chicken categories [chicken(A) and chicken(B)] for checking the same categories return similar results. These were selected from the Cookpad recipe categories considering texture quality and differences in the categories.

3 METHOD

3.1 Making low-resolution images

Typically, a general super-resolution approach does not also reduce noise, but if we introduce super-resolution into food images, we should remove noise within our super-resolution framework. This can be implemented using a down-sampling method.

In this paper, we propose the method to reduce the resolution by adding artificial jpg-type noise and blur, especially concentrating on images that are out-of-focus because of the compression process used for handling images within the service and old camera performance. We show the procedure for creating low-resolution images in Figure. 2.

Inducing jpg noise Jpg noise artifacts causes serious problems when super-resolving images. To reproduce the noise generated when an image, jpg noise was artificially added using Python Pillow library. Depending on the quality of the jpg compression (a uniformly distributed random number from 20 to 100), the image quality was lowered.

Inducing blur The performance of old cameras causes various types of bad image representations. As for the jpg problem, we reproduce bad representations using image processing (Figure 2).

Filtering was applied to reproduce out-of-focus images or blur due to camera performance using Python scikit-image library. There are also various types of blurring, and in this study we used a median filter and a Gaussian filter, chosen while checking the quality of the image. Of the entire dataset, 50% were processed using a median filter, 25% were processed using a Gaussian filter, and the rest were not filtered. The median filter was used with the default parameters, and the Gaussian filter had a sigma of 1 for half of the images it processed and 2 for the other half.

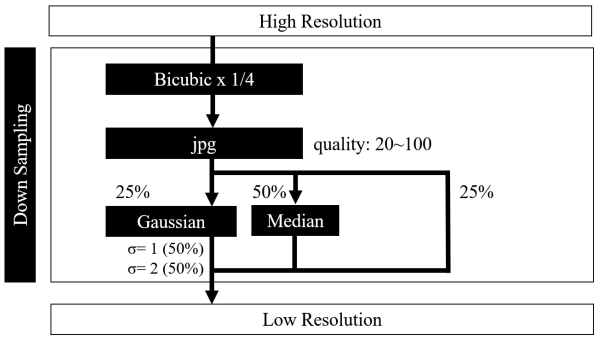

Figure 2: Procedure for creating low-resolution images

3.2 Xception score

It is difficult to evaluate images generated from GANs. Hence, in recent work, scoring using pre-trained networks has been used (e.g., *Inception score*). In this study, we imitate this approach and propose a similar evaluation method.

That is, we propose using the average value of the output softmax value p_c (c means categorical) of an ImageNet pre-trained 1000 class classifier:

$$s_c = \frac{1}{N}\sum_{i=1}^{N} p_c(x_i). \tag{1}$$

Here, N is the number of test data and x_i is an input image. We assume if the outputs can reconstruct appropriate texture, the score of (1) will be higher than if the texture is inappropriate.

4 EXPERIMENTS

To clarify the effectiveness of the down-sampling method, we trained various SRGAN models. First, to confirm the effectiveness of our proposed down-sampling method, we trained two models, one that included and one that did not include the down-sampling method. Next, to investigate the difference in appearance due to changes in the image category, such as cooking type, we trained SRGAN six times using different datasets. Specifically, we used the following categories: DIV2K jpg, chicken (A), chicken (B), beef, bread and pound cake. Finally, we trained SRResNet, which is good at according to extisting evaluation methods like PSNR to determine if SRGAN is really the best tool for super-resolution.

4.1 Training details

The training of SRGAN is divided into two stages: the first stage optimizes the MSE of the output of SRResNet and the second stage

adds the loss function of the middle layer MSE and GAN to the loss function and trains the SRGAN. Although the intermediate layer output has arbitrariness, the VGG54 used in the original paper is also used here. The model was implemented with TensorFlow, the batch size was 16, and the SRGAN were optimized using Adam.

For the first stage of training the SRGAN, the learning rate of Adam is 10^{-4} and the total epochs were set to 100. In the second stage, the learning rate of Adam was the same, but the total number of epochs was $2,000$ and the learning rate was divided by 10 after epoch $1,000$.

The training of SRResNet is simpler than that of SRGAN. The architecture of SRResNet is SRGAN's own generator without the discriminator. Similar to SRGAN, initial learning rate of Adam was 10^{-4}, and total epochs were $2,000$. Moreover, the learning rate was divided by 10 after epoch $1,000$.

It takes about one day in the case of AWS P3 to train a single SRGAN model. In addition, SRResNet takes almost 8 hours.

5 EVALUATION

5.1 Qualitative evaluation

Tables comparing the resulting images showed in Figures 3 and Figures 4. Here in order to compare the detailed texture, a small section of the results are enlarged. Because of space limitations, we could not show all results. Hence, we show only the bread and beef results.

The result obtained using DIV2K (Bicubic) show that jpg noise is emphasized and image quality is not substantially improved. Although the method trained on DIV2K (jpg) has been able to cleanly denoise the images in the upper and middle rows, some undesirable textures are generated in the lower-row images. This is thought to be because the model trained with the various images in DIV2K has not sufficiently learned the texture that frequently occurs in this category, especially in images in the food domain.

In addition, the output images for the beef, chicken (A), and chicken (B) categories tend to have an unnatural glossy texture. This is considered to be caused because the method strongly reproduces a texture from the meat domain, in which there are many glossy images because of oil. Although chicken (A) and chicken (B) images have different output, they are both glossy, and the tendencies of these generated textures appear to be similar. From this result, it can be inferred that SRGAN captures the feature of each category.

On the contrary, the network trained with bread and pound cake induces a texture that is close to the ground truth. This appears to indicate that the appropriate texture can be learned by the appropriate dataset.

As in the case of bread, DIV2K (Bicubic) could not improve results but emphasizes the noise. Although the output image of DIV2K (jpg) is clear, compared with the result of DIV2K (Bicubic), the gloss and texture of the meat appears to be unreproduced, and it does not restore well the original aesthetics of the cooking image.

The beef, chicken (A), and chicken (B) categories create images that are rich in gloss and reproduce meat fat well. On the contrary, it seems that this is too much to reproduce, and SRGAN can learn the features of the texture, remarkably well.

In the bread and pound cake categories, the results are not as bad as for DIV2K (jpg), but the images look less shiny. This is also considered to be the result of appropriately reflecting the characteristics of the categories; that is both bread and pound cake have matte textures.

5.2 Quantitative evaluation

We evaluated the output images using the Equation (1). Although, ImageNet (1,000 classes) does not include the same categories as bread and chicken used in this article, categories close to these domains such as "meat loaf" and "French loaf" are included.

We conducted this evaluation with a pre-trained VGG16 and Xception [2]. We show only the Xception result because both results are similar. Note that despite the super-resolution, the super-resolved input image is resized to 299×299 to match the receptive field of Xception when predicting the output.

The top three scores for the image of the pound cake are shown in Table 2. The scores of "French loaf" of bread and pound cake are higher than those of meat-based domain, and the score of "meat loaf" of beef, chicken (A), chicken (B) is higher than that of bread.

From this result, it can be inferred that SRGAN correctly learns the texture of each domain, and the resulting restored image is more easily recognized as a domain category that is closer to the model of ImageNet. However, the score of "French loaf" of DIV2K (jpg) is also high, and although this index captures a part of the features of SRGAN, there are also inadequacies in the output.

We used existing methods PSNR and SSIM. As mentioned above, this is not an appropriate indicator, but it is the main application to investigate whether there is a deviation from the previous research. We calculated the PSNR and SSIM scores, which are generally used for image quality evaluation. The results agree well with previous research [3]. In addition, the results reflect the presence or absence of the low-resolution method and differences caused by the domain were not obtained.

6 DISCUSSION AND CONCLUSION

In this work, we demonstrate that devising a down-sampling procedure and dividing the categories of the datasets is effective for training SRGAN, especially in the food domain. In addition, using our method, the performance of SRGAN was improved qualitatively and quantitatively.

However, the quantitative evaluation performed in this study is not sufficiently thorough, so other evaluations other than PSNR and SSIM were also performed. First, we compared the loss in the frequency domain using a two-dimensional discrete Fourier transform. Second, we calculated the *Inception score* of the output image. However, insufficient differences were found using both metrics. Since there are various evaluation metrics for image aesthetic quality, we'll apply those metrics to the proposed method.

Because the proposed method in this paper does not work well for some of the images, it would be difficult to fully introduce it into our recipe-based service. Therefore, we are considering introducing it by appropriately targeting its application and combining with manual checking.

REFERENCES

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

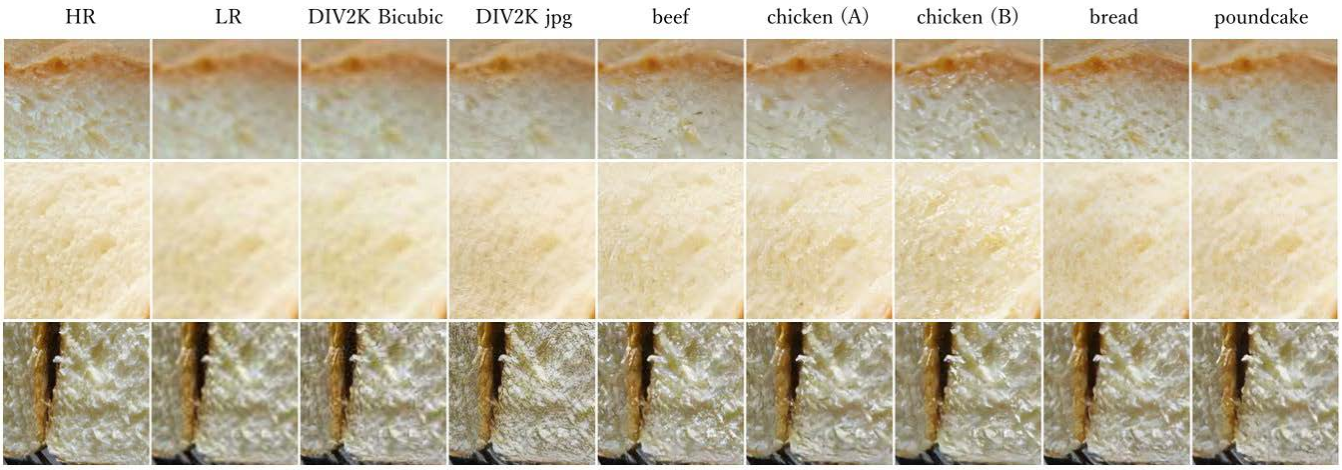

Figure 3: Cropped super-resolution images of bread obtained using each trained model. Each image is a recipe images of (Recipe name, Recipe author and URL) in order from the top as follows: (HB で☆リッチな卵ミルク食パン, ジョンリークッカー, https://cookpad.com/recipe/690987), (HB*超リッチ♪生クリームと練乳の絹パン, ねっちゃんっ, https://cookpad.com/recipe/983984) and (ハチミツレモンブレッド, はなむん, https://cookpad.com/recipe/397851). These are acquired on March 8, 2018.

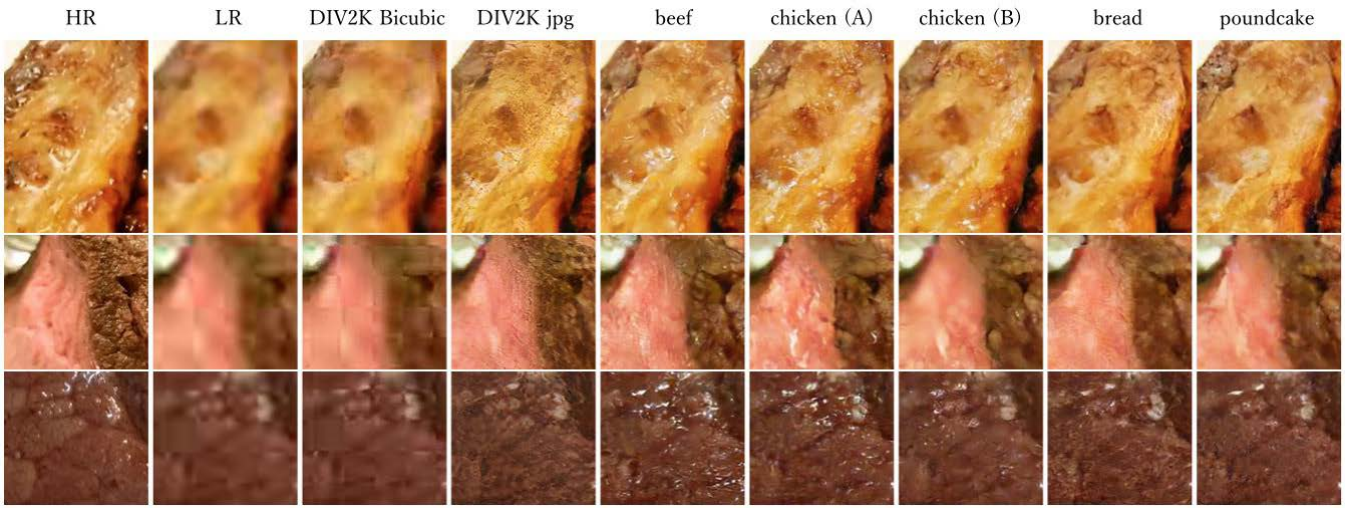

Figure 4: Cropped super-resolution images of beef obtained using each trained model. Each image is a recipe images of (Recipe name, Recipe author and URL) in order from the top as follows: (どんな材料でもお手軽☆味噌漬け焼き, AyakoOOOOO, https://cookpad.com/recipe/1008812), (ローストビーフをフライパンで作ろう！, はーたんのおっかさん, https://cookpad.com/recipe/1900893) and (安いお肉を美味しいステーキにする方法♪, ElisabethF, https://cookpad.com/recipe/2767568). These are acquired on March 8, 2018.

[2] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. arXiv preprint (2017), 1610–02357.
[3] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. CVPR (2017).

Table 1: Result of the proposed evaluation method for test image (pound cake) using Xception.

rank	LR	HR	DIV2K (jpg)	beef	chicken (A)	chicken (B)	bread	pound cake
1	French_loaf (**0.34**)	French_loaf (**0.31**)	French_loaf (**0.34**)	French_loaf (0.20)	French_loaf (0.23)	French_loaf (0.23)	French_loaf (**0.34**)	French_loaf (**0.32**)
2	plate (0.15)	plate (0.16)	plate (0.12)	meat_loaf (**0.18**)	meat_loaf (**0.17**)	plate (0.16)	plate (0.12)	plate (0.13)
3	potpie (0.11)	meat_loaf (0.11)	meat_loaf (0.08)	plate (0.18)	plate (0.16)	meat_loaf (**0.14**)	meat_loaf (0.07)	meat_loaf (0.11)