Review on Paper
SinGAN: Learning a Generative Model from a Single Natural Image

Ilia Kobelev
MIPT

This paper was originally submitted to the ICCV 2019 conference by Tamar Rott Shaham, Tali Dekel and Tomer Michaeli and received the Best Paper Award.

1 Introduction

Following the recent success of Generative Adversarial Networks in producing high-quality visual data the authors of the article present SinGAN – a generative model that needs only one natural image for the whole training procedure. Furthermore, the method presented in this article allows to train GAN which does not require both conditioning on input image or other prior signal and pre-training on the class-specific task. Thus, SinGan can generate samples with complex natural objects purely from random Gaussian noise in contrast to previous works which used similar training methods but in the realm of texture generation. Still, in my opinion, examples of the model's output included in the Supplementary Materials demonstrate that it is capable of generating realistic image samples. It is also confirmed by the fake/non-fake user studies conducted by the authors as well as some other quantitative metrics.

Apart from image manipulation tasks such as editing, harmonization, super-resolution and animation, described by the authors, I think SinGAN can be widely used for the data augmentation as it is able to modify both coarse and fine structures of the input image while generating new samples. However, I should admit that this approach is most likely intractable at present time, since the model needs 30 minutes of training time for the image of size 256×256 pixels, therefore augmenting the data can be more costly than its manual labelling.

2 Method

The SinGAN model makes extensive use of a multi-scale pipeline. Basically, the model's architecture is a pyramid of fully-convolutional GANs each responsible for generating patches at different scales – from the coarsest at the Nth layer to the finest (i.e. scale of the original image) at the layer 0. Both during training and inference time layers are processed in the reverse order to their numbers, i.e, in coarse-to-fine fashion. Although authors repeatedly emphasized that

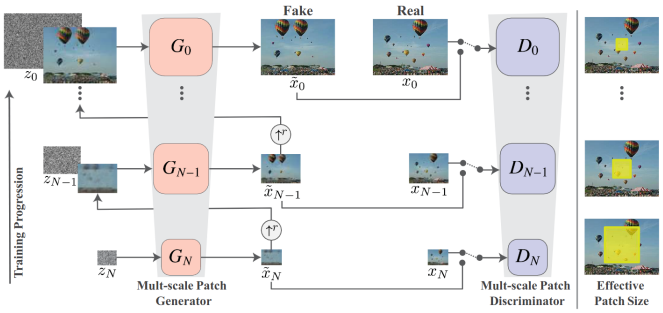

Figure 1: SinGAN multi-scale pipeline

the model concerned is purely unconditional, actually it is true only for the Nth layer that is supposed to generate downsampled images from the tensor of Gaussian noise of the same size. Each next layer i takes as input both the sample of noise and the upsampled output of GAN at the layer $i+1$ and is supposed to fix the latter (i.e. to generate additive) to match the image at the current scale as well as possible. I can't help wondering if more complicated algorithms for upsampling between SinGAN's layers were tested apart from simple linear interpolation. Specifically, whether using transpose convolution with learnable parameters can benefit the model's performance.

Each layer consists of generator and discriminator that both have the same architecture – 5 stacked convolution layers with non-linearity and Batch Normalization between them. However, it is not mentioned in the paper whether any other architecture was tested. For instance, one could propose to experiment with Residual or Squeeze-and-Excitation blocks or add Highway connections between layers. The authors, though, explained that the architecture was intentionally chosen to be as simple as possible to learn patch statistics just at one aspect ratio and not overfit the whole image.

The discriminator in each GAN effectively models an image as Markov random field assuming that pixels depend only on the local neighbourhood. In simple words, discriminator is applied to the image in strided fashion and for each patch outputs the score whether it is real or fake. Scores for all spatial locations are then averaged to produce the loss value for the whole image. Authors argue that this technique helps to achieve significant variability, yet maintain patch distribution close to the original image.

SinGAN is trained layer by layer, i.e. once the ith GAN has converged it is fixed for the rest of the optimization process and training proceeds with the layer $i-1$. It is not mentioned, though, if an end-to-end optimization of the whole pipeline was tested. I could assume that this scenario would almost surely

lead to vanishing of the gradients, and this problem can not be easily solved with residual connections, since GANs at different layers operate with the tensors of different sizes which can not be added pixelwise.

One more idea from the paper, which I deem to be quite sensible, is worth mentioning. At each scale we are to decide on the parameters of spherical Gaussian distribution to be used for drawing noise samples. Naturally, the mean is set to 0, since we do not need any additional bias while generating images. But the standard deviation is chosen to be proportional to the RMSE between the upsampled output of the coarser layer and the target image. This approach seems to be reasonably rational if we recall that all variability added at the current scale is generated from the random noise, which therefore would better be of the same order as the additive we are aimed to predict.

3 Results

Briefly, SinGAN was tested on images depicting complex objects such as buildings, mountains, baloons and successfully proved to be capable of producing new structures in global arrangement as well as preserving fine structures like snow or sand grains.

I would rather focus on the features of the pipeline that allow to alternate between generating more or less coarse structures in the resulting images. Experiments conducted by the authors clearly show that the diversity of the output is largely affected by the first SinGAN layer where the noise is actually sampled during the inference. More precisely, the sequence of the noise samples $\{z_N, z_{N-1}, \dots z_0\}$ is fixed during training and the model is optimized to reproduce exactly the original image from this sequence. Then, while evaluating the model, this sequence is unfreezed at some point with coarser noise samples remaining fixed and finer ones becoming variable. Starting noise sampling at more deep SinGAN layers leads to increased variability in global structure, therefore producing sometimes unrealistic images. On the other hand, if one unfreezes a few top layers, then the model will preserve global arrangement and modify only fine textures of objects in the input image.

The realism of the generated images was quantified by the user study that followed two protocols: paired (real vs. fake) and unpaired (either real or fake). Both approaches demonstrated high confusion rate with the unpaired one achieving 47% rate for the images generated from the scale $N - 1$ (perfect confusion rate equals 50%, and more is better). Another metric for quantification of the SinGAN introduced by the authors is Single Image Frechet Inception Distance. It makes use of activations of the convolution layer just before the global average pooling of the Inception network and treats each vector as corresponding to the patch in the input image. FID scores computed between real

and fake activations for each patch are averaged and give the final SIFID value which is below 0.1 for the pool of 50 test images.

4 Conclusion

To sum up, the paper presents a generative model that is trained in adversarial fashion and is aimed at learning patch statistics of the input image at different scales. The model can be directly applied to the wide range of image manipulation tasks with the same architecture and does not require any additional information or pre-training beyond the single image. The overall architecture of the SinGAN pipeline is rather simple and easy to implement. However, in my opinion, the optimization process is highly unstable and requires a number of heuristics and advanced solutions to converge (for example, Wasserstein GAN loss and restricting norm of the gradients).

The authors conducted thorough examination of the success and failure cases and evaluated generated images both quantitatively (SIFID metric) and qualitatively (user study).

To my mind, one of the key features of the introduced model is the flexibility of the GAN pyramid which makes it possible to inject randomness at different scales and in such a way to control coarse and fine structures in the output. SinGAN completely removes the need for big data in such applications as super-resolution, harmonization, single image animation and others.