Image manipulation with generative adversarial networks (GANs)

Mohamed EL MENNAOUI - Mohamed KERROUMI
mohamed.el-mennaoui@student.ecp.fr kerroumimohamed1997@gmail.com

Abstract

In this paper, we will present a short overview of SinGAN , published on 2019, a new unconditional generative model proposed by the authors of [5] that is trained on single natural image. This model could be used for a variety of image manipulation tasks. We will try to use it to do image inpainting, one of the challenging image restoration tasks.

1. Introduction

SinGAN is an unconditional generative model that captures the patches distribution of a single training image x. To deal with all kind of general natural images, the model needs to capture the statistics of complex image structures at many different scales and the global properties of the image such as the arrangement and shape of objects in the image.

2. SinGAN Model & Applications

2.1. Model Description

The model can capture fine details and texture information thanks to its multi-scale pipeline (the figure 1 shows the pipeline of the model). The model consists of a pyramid of GANs, where each is responsible for capturing the patch distribution at a different scale of the training image.

Both training and inference are done in a coarse-to-fine scale. At each scale the generator G_n learns to generate image that cannot be distinguished by the discriminator D_n, the generator G_n takes as inputs a random noise image z_n and the generated image from the previous scale $\tilde{x}_n$ upsampled with the scale factor r.

$$\tilde{x}_n = G_n(z_n, \tilde{x}_{n+1\uparrow^r})$$

Except for the coarsest scale which is purely generative.

$$\tilde{x}_N = G_n(z_N)$$

Then the discriminator tries to distinguish the output of the generator from the down-sampled version of the training

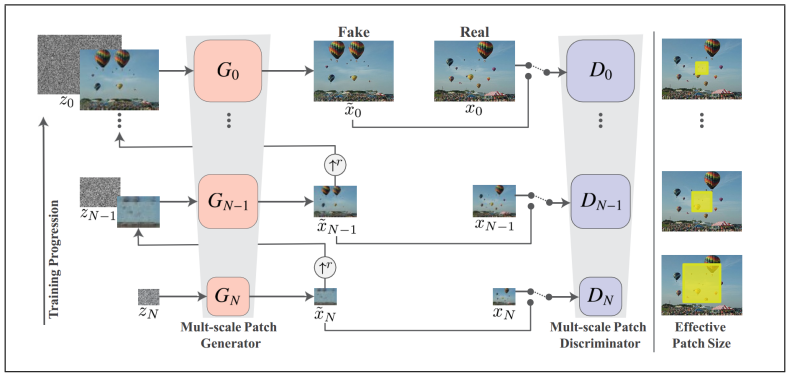


Figure 1: SinGAN's multi-scale pipeline;Image extracted from the original paper [5]

image x_n. Through adversarial training G_n learns to produce realistic patches. The coarser levels are responsible for the generation of the object global structure, and finer levels adds details that were not generated by the previous scales.

2.2. Training

The training of the model is done sequentially, we start by training the coarsest scale to the finest one, when a given scale is trained we keep the parameters of the corresponding GAN fixed. The training loss of each scale contains an adversarial term and a reconstruction term.

$$\min_{G_n} \max_{D_n} \quad L_{adv}(G_n, D_n) \quad + \alpha L_{rec}(G_n)$$

- Adversarial Loss The adversarial loss is used to penalize the distance between the distribution of patches in x_n and the distribution of patches in generated samples $\tilde{x}_n$. The authors of the paper used the WGAN-GP loss.

$$\min_{G_n} \max_{D_n} \quad E_{x\sim P_r}[logD_n(x)] \quad + E_{\tilde{x}\sim P_g}[log(1-D_n(\tilde{x})]$$

This increases training stability. The generator and discriminator have the same architecture (5 Conv-layers)

- The reconstruction loss is used to ensure the existence of a set of noise map that can reproduce the original image. Hence during training we fix a noise map z^* and we input the following random noise image at each scale : $\{z_N^{rec}, z_{N-1}^{rec}, ..., z_0^{rec}\} = \{z^*, 0, ..., 0\}$ and the reconstruction loss has the following expression:

$$L_{rec} = \|G_n(0, (x_{n+1}^{rec}) \uparrow^r) - x_n\|^2$$

Where x_{n+1}^{rec} is the generated image at the n+1th scale when using the previous noise map.

2.3. Applications

We tried to reproduce some of the results presented in the paper to make sure that we understood the paper and that everything is working correctly. So we trained our model on some natural images and generated some random samples and ensure the the output images preserve the visual content of the training image. As we can see on figure 3 , the generated images depict new realistic configuration of objects while preserving the global structure of objects.

Training image

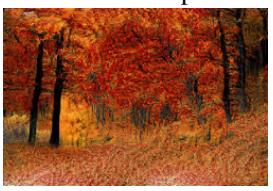
Random Sample 1

Random Sample 2

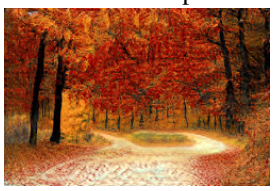
We tested SinGAN also on texture images, and the model handles well this kind of images, As can be seen on the figure below, the generated images contain new combinations of patches that do not exist in the training image. The generated images have a complex texture and a non-repetitive global structure.

Training image

Random Sample 1



Random Sample 2



3. Model extension to Inpainting

Inpainting is a process of restorative conservation, where damaged, deteriorating, or missing parts of an artwork are reconstructed, ultimately with the goal of presenting the work as it was originally created (wikipedia).
Many efforts have been put by researchers in the field to use GANs for Inpainting task, such as [3] and [2] where researchers proposed a new and innovative was of using generative models combined with a special type of convolutions named gated convolution.
In this article, we will try to use SinGAN model to perform the same task.

3.1. Approach

In our approach we assume that we already have a reference image which is quite similar to the image to be inpainted. That is usually the case if one wants to use this extension for subtitles removal in a video. The user could easily extract a reference image from an adjacent frame without subtitles.
Once we get the reference image, we extract the mask of the subtitle or the occluding object using binarization. One could also use object detection state-of-the-art algorithms such as mask R-CNN to automatically extract masks.
Finally, we train SinGAN on the image with subtitles or occluding object. Then inject a downsampled version of the reference image into one of the coarse scales. In practice, we take the last one. We then combine SinGAN's output at the edited regions, indicated by the mask, with the original image. This results in unnoticeable inpainting as we'll see in the results.

3.2. Metrics

To evaluate the Inpainting process, one could show the resulted image to people and do a survey on the quality of inpainting. However, this is a subjective evaluations and very difficult to scale. Instead, in order to quantify inpainting processes objectively, we will focus on the two following metrics [4]:

- **Structural Similarity Index (SSIM):** a metric used to predict the perceived quality of a digital picture.

$$ssim(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_1)}$$

- **Average Squared Visual Salience (ASVS):** is a non reference (NR) metric. It relates to the noticeability of the inpainted pixels compared to the overall scene.

$$ASVS = \frac{1}{||\Omega||} \sum_{p \in \Omega} |S^{'}(p)|^2$$

4. Results

We tested our method for both subtitle removal and occluding objects removal. The second setup is an artificial experiment. Since, we assume having a good reference image. We first trained SinGAN on clean images then added the occluding object and used the random samples for SinGAN's output as a reference image.
We compared our method with the NVDIA inpainting tool, that could be found here [1], and an openCV built-in inpainting algorithms based on Fast Marching Method and developed by Alexandru Telea in 2004.
Keep in mind that for the quantitative evaluation, we want the **SSIM** to be as close to 1 as possible, and the **ASVS** be as small as possible, meaning that the inpainting is unnoticeable.

4.1. Qualitative evaluation

Please refer to the annex for the qualitative evaluation. The space here was not sufficient to provide quality figures.

4.2. Quantitative evaluation

Image	OpenCV	NVIDIA	SinGAN
Movie	0.92	0.32	0.88
Birds	0.94	0.96	0.95
Mountains	0.21	0.16	0.56
forest	0.24	0.18	0.92
car	0.95	0.95	0.79

Table 1: SSIM metric between inpainted image and a reference image

Image	OpenCV	NVIDIA	SinGAN
Movie	0.005	0.045	0.006
Birds	0.068	0.051	0.066
Mountains	0.059	0.053	0.049
forest	0.034	0.037	0.034
car	0.15	0.135	0.181

Table 2: ASVS metric calculated over the inpainted region

Avg. metric	OpenCV	NVIDIA	SinGAN
SSIM	0.65	0.51	0.82
ASVS	0.063	0.064	0.067

Table 3: Average metrics over samples for comparison

5. Conclusion

We examined through this article the abilities of the new unconditional generative model, SinGAN, to perform various Image manipulation tasks. We also tried to extend it to perform Image Inpainting, and we compared it with the OpenCV built-in method based on Fast Marching Method and the NVIDIA tool. Over the samples that we were able to get given the time and cloud credit, we conclude that SinGAN achieves promising results if given a good reference image. In practice this might be an issue.
Hence, a possible extension of the model could be to add a contextual layer to try Inpainting during the training part.

References

[1] https://www.nvidia.com/research/inpainting/. 3

[2] Jimei Yang Xiaohui Shen Xin Lu Jiahui Yu, Zhe Lin and Thomas Huang. Generative image inpainting with contextual attention, 2018. 2

[3] Jimei Yang Xiaohui Shen Xin Lu Jiahui Yu, Zhe Lin and Thomas Huang. Free-form image inpainting with gated convolution, 2019. 2

[4] Azeddine Beghdadi Muhammad Ali Qureshi, Mohamed Deriche and Asjad Amin. A critical survey of state-of-the-art image inpainting quality assessment metrics. *Journal of Visual Communication and Image Representation*, 49:177–191, 2017. 2

[5] Tali Dekel Tamar Rott Shaham and Tomer Michaeli. Singan: Learning a generative model from a single natural image, 2019. 1

ANNEX : Qualtitative Evaluation