Generative AI Assignment 2

Advik Sinha, Rohan Mathew Sabu

Abstract

Generative Adversarial Networks (GANs) aim to generate realistic synthetic data using two neural networks: a generator which generates fake samples and a discriminator that tries to distinguish them from real data. Through this process, the generator improves over time, generating increasingly realistic outputs. The paper ***SinGAN: Learning a Generative Model from a Single Natural Image*** *[1] introduces a novel unconditional generative model that can be trained on a single natural image.*

1 Introduction

Generative Adversarial Networks (GANs) are widely used for tasks such as image generation, style transfer and data augmentation. Most GAN-based approaches are trained on class-specific datasets and often condition the generation of another input signal. Prior works on single image GAN approaches are largely conditioned on an input image.

The paper proposes a novel approach of unconditional generation using a single natural image. It shows how the internal statistics of patches within a single image carry enough information for learning a powerful generative model. The model differs from a conventional GAN in that it uses patches of a single image as the training samples rather than a set of images. Learning the underlying distribution of the patches within a single image has important use in several computer vision tasks such as denoising, deblurring, super resolution, dehazing, and image editing.

2 Architecture

The model architecture consists of a pyramid of generators, G: $\{G_0, \ldots, G_n\}$ that learns the distributions of the input images at different scales. Each generator is trained using the corresponding image from an image pyramid of the input image x: $\{x_0, \ldots, x_n\}$. The model also consists of a pyramid of discriminators D: $\{D_0, \ldots, D_n\}$, and each generator attempts to fool the associated discriminator by producing realistic image samples w.r.t. the patch distribution of its corresponding image.

The model functions sequentially, with the first image generation starting at the lowest resolution. The first image generation is entirely generative, with G_n generating an im-

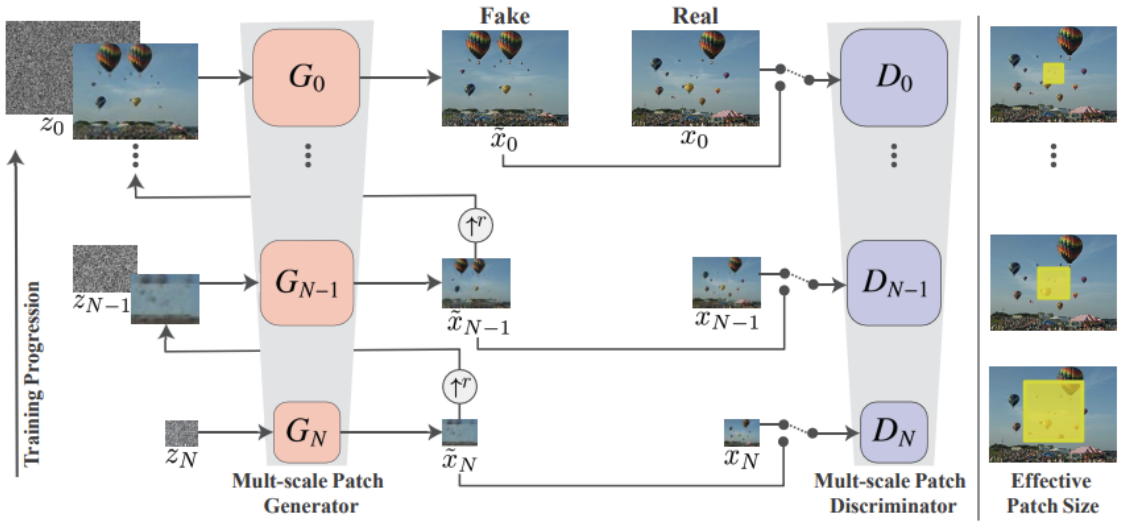

Figure 1: SinGAN Pipeline

age from white Gaussian noise. Every subsequent layer takes an upsampled version of the image generated by the previous layer along with the Gaussian noise. As described in the paper [1], the nth generator performs the following function:

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

The full model pipeline is shown in Figure 1.

The generator model is made up of a fully convolutional layer with 5 conv-blocks of the form Conv(3 x 3)-BatchNorm-LeakyReLU. The convolutional layer initially has 32 kernels per block and increases by a factor of 2 every 4 scales. Each generator performs the operation

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

where ψ_n is the fully convolutional network described above. The architecture of D_n is the same as the net ψ_n within G_n.

3 Training Details

The model is trained sequentially, from the lowest resolution to the highest. The training loss for the nth GAN is as follows:

$$\min_{G_n} \max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha \mathcal{L}_{\text{rec}}(G_n)$$

where $\mathcal{L}_{\text{adv}}$ aims to ensure that the generated samples are similar to the actual samples and $\mathcal{L}_{\text{rec}}$ aims to ensure that a set of noise maps exist which can reconstruct the input x.

The adversarial loss is calculated using the WGAN-GP loss [2], which adds an extra gradient penalty to the existing critic loss to stabilize training. The final discrimination score is calculated as the average over the patch discrimination map. The loss is calculated over the entire input image.

The reconstruction loss is used to ensure that there exists a set of noise maps that generates the input image x. As mentioned

Figure 2: SinGAN Generated Synthetic Image

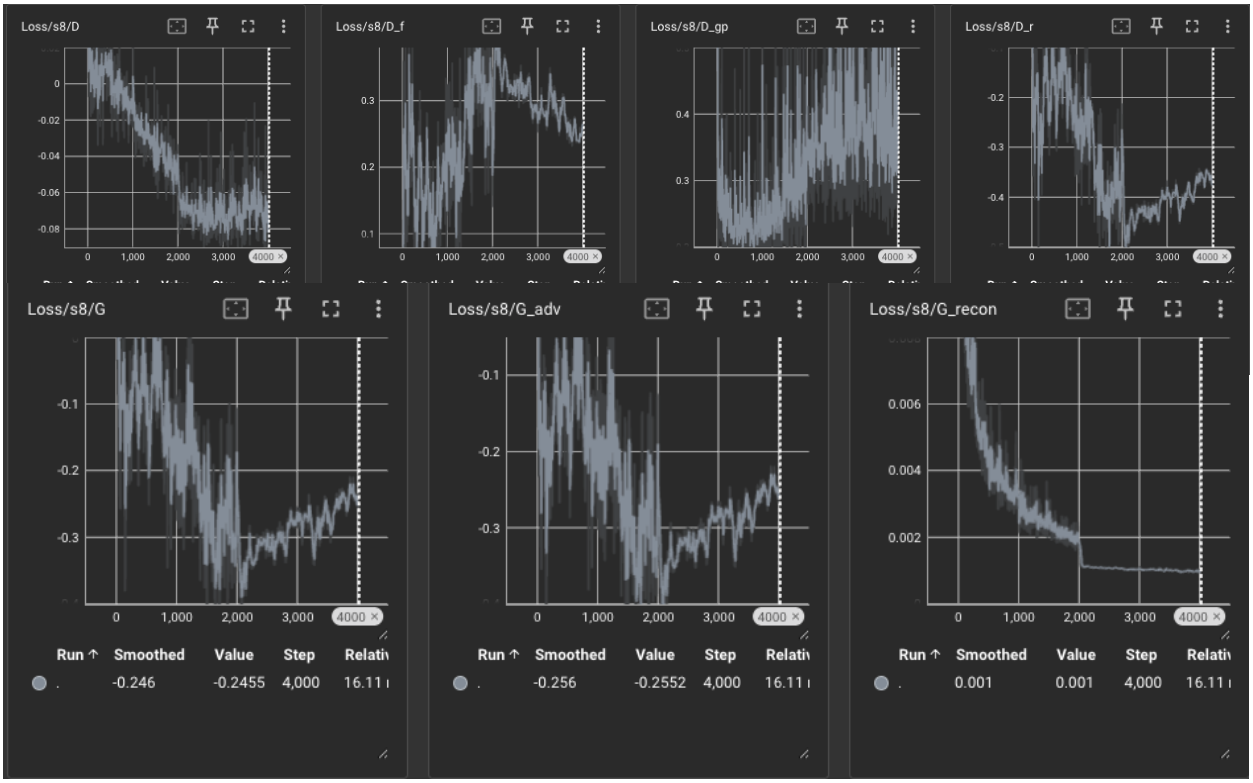

Figure 3: SinGAN Loss Plots

in the paper, we choose some

$$\{z_N^{\text{rec}}, z_{N-1}^{\text{rec}}, \ldots, z_0^{\text{rec}}\} = \{z^*, 0, \ldots, 0\}$$

, where z^* is a noise map that is kept fixed during training. If $\tilde{x}_n^{\text{rec}}$ denotes the image generated by the nth generator when using those maps, the for $n < N$

$$\mathcal{L}_{\text{rec}} = \left\| G_n\left(0, \left(\tilde{x}_{n+1}^{\text{rec}}\right) \uparrow^r\right) - x_n \right\|^2,$$

and for $n = N$

$$\mathcal{L}_{\text{rec}} = \|G_N(z^*) - x_N\|^2$$

The model was trained for 3000 iterations per scale. The number of scales depended on the input image size. Training was performed with a learning rate of 5×10^{-4}, using the Adam optimizer for both the generator and discriminator, with $\beta_1 = 0.5$ and $\beta_2 = 0.9$. We used a reconstruction loss weight of 10, an adversarial loss weight of 1, and a noise weight of 0.1.

The generator employs 3×3 convolutional kernels. At coarser scales, these kernels cover a larger receptive field due to downsampling, thus capturing broader structural information. At finer scales, the same kernels affect smaller regions of the image, allowing the network to model finer details.

4 Results

The authors of the paper evaluated the model using various qualitative and quantitative methods on a large number of images. As part of the qualitative evaluation, they presented several images generated by the model at various scales. Quantitative analysis was done using the AMT perceptual test, where workers were presented with several images and asked to classify if the image was real or fake.

The paper also presents the use of SinGAN for various applications, such as Super-Resolution (*Increase the resolution of an input image by a factor s*), Paint-to-Image (*Transfer a clipart into a photo-realistic image*), Harmonization (*Realistically blending a pasted object with a background image*), Editing (*Produce a seamless composite in which image regions have been copied and pasted in other locations*) and Single Image Animation (*Create a short video clip with realistic object motion, from a single input image*).

We provide the loss plots obtained on training the model in 3.

Some synthetic images of the image the model was trained on are provided in Figure 2. These images were generated based on the noise provided. The core idea of the paper is that the model learns the intricacies of the input image such that it makes realistic changes to the image.

References

[1] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," 2019.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.