Are GANs Created Equal? A Large-Scale Study

Seungmin Lee (profile2697@gmail.com; 2013-11420), Dept. of Computer Science and Engineering, Seoul National University

1. Motivation

Even though an enormous number of Generative Adversarial Networks (*GAN*) algorithms have appeared, there is no consensus on what is the proper way to compare the algorithms. This paper tries to compare the performance of the algorithms with carefully designed experiments. This paper mainly focuses on an evaluation metric called Fréchet Inception Distance (*FID*), but it also proposes a dataset with precision, recall, and F1 score as complementary metrics. Based on the results, this paper claims that there is no evidence that other methods work better than the firstly proposed GAN. Furthermore, this paper also argues researchers should present the distributions of performances according to the budgets.

2. Settings for the Fair Comparison

2.1. Algorithms and Architectures

In this paper, the authors only care about unconditional GANs that only utilize unlabelled images. Seven algorithms, including the first GAN, are selected. In the experiments of this paper, all algorithms except for BEGAN use InfoGAN architecture, and BEGAN uses an architecture that has a similar number of parameters.

2.2. Budgets and Randomness

Although architectures and datasets are the same, the GAN algorithms can be affected by hyperparameters and random seeds. The authors report the performance changes according to the hyperparameters and random seeds. For hyperparameters, the authors also concern hyperparameters' search budgets. Algorithms that have fewer hyperparameters are more advantageous in these settings.

2.3. Metrics

Fréchet Inception Distance (*FID*) To measure the FID, assume that there is a pre-trained classifier. Then using the classifier, embed samples from both real images and generated images, respectively. After that, the FID can be measured by modeling each group of embedded samples as a gaussian distribution:

$$FID(d,g) = \|\mu_d - \mu_g\|^2 + Tr(\Sigma_d + \Sigma_g - 2(\Sigma_d \Sigma_g)^{\frac{1}{2}}) \quad (1)$$

where (μ_x, Σ_x) indicates the mean and covariance of embedded features.

The authors empirically show that the FID is a useful metric for measuring mode collapsing compared to the Inception Score. However, they also point out the FID fails to detect overfitting. If a model that memorizes all the training examples like $memoryGAN$, the model gets a perfect score at FID. Nevertheless, the authors adopt the FID as a main metric.

Precision, Recall, and F1 Score For alleviating the problem of the FID, the authors suggest precision, recall, and F1 score as complementary metrics with a newly proposed dataset. The proposed dataset is a set of convex polygons, which is designed to be harder than MNIST and easier than ImageNet so that it can efficiently calculate the proposed metrics. Distances between samples and data manifolds measure the proposed metrics. For example, precision becomes higher when generated samples settle close to a manifold. Similarly, recall becomes higher when generated samples can cover manifolds as many as possible. This paper also suggests how to approximate the distance. The distance can be measured using the following equation with a generator G and a given image $x \in \mathbb{R}^{d^2}$:

$$\mathcal{L}(x, x^*) = \|x - x^*\|^2 \quad (2)$$

$$\text{where } x^* = G(z^*) \quad (3)$$

where $z^* = \arg\min_z \|x - G(z)\|^2$, which indicates a latent vector used for generating the given image's closest sample from the manifold.

3. Results

Using the settings mentioned in the previous section, the authors conducted many experiments. For hyperparameters, the GAN algorithms are highly sensitive, so no algorithm stably achieves the best FID score or F1 score. Similarly, no algorithm dominates others when using mean FID. With fixed budgets, the authors argue that they could not find a statistically superior algorithm over other algorithms. Therefore, they claim that other papers should report the distribution of the performances over varying budgets. For randomness, some algorithms are unstable, and all algorithms become volatile as the complexity of the dataset increase.

4. Personal Memo

Despite the carefully designed experiments, I think the conclusion is quite impractical and straightforward. Moreover, the selected algorithms seem not state-of-the-art. The latest algorithm was published in 2017. Additionally, I believe the used hyperparameter search policy, which is similar to grid search in a wide range, is impractical. I even think it could be unfair to some algorithms if their hyperparameters can be easily inferred by a human but in a narrow range.