How latent space affect the images generated by CGAN model

Kaiyu He
Columbia University, New York
kaiyuhe998@gmail.com

Abstract

In this study, I examine the relationship between the dimensionality of the random latent vector used to generate images with a Conditional Generative Adversarial Network (CGAN) and the quality of the generated images, as measured by the Fréchet Inception Distance (FID) score. Our results show that increasing the dimensionality of the latent vector leads to the generation of higher-quality images, but that there may be an optimal point beyond which further increases have little effect. Our findings suggest that the FID score can be a useful measure of the quality of images generated by a CGAN and that the dimensionality of the latent vector plays a role in determining this quality. And also I find a way to evaluate information contained in a dataset.

1 Introduction

Generative adversarial networks (GANs) have gained significant attention in recent years for their ability to generate high-quality images that are difficult to distinguish from real ones. GANs consist of two neural networks, a generator and a discriminator, that are trained in a two-player minimax game. The generator tries to generate realistic images that the discriminator cannot distinguish from real images, while the discriminator tries to correctly classify the generated images as fake. This adversarial training process results in the generator learning to produce increasingly realistic images over time.

In this study, I investigate the use of the FID score to evaluate the quality of images generated by a Conditional Generative Adversarial Network (CGAN). I focus specifically on how the dimensionality of the random latent vector used to generate the images affects the FID score. Latent vectors are often used to provide noise or randomness to the generation process and are typically high-dimensional. However, little is known about the optimal dimensionality for these vectors and how it might affect the quality of the generated images. By studying the relationship between latent vector dimensionality and image quality, I aim to provide insights that may be useful for improving the performance of CGANs in a variety of applications.

2 Data

Here I tried CGAN model on two datasets. One is Clothing Model dataset and one is MNIST handwritten digits dataset.

2.1 Clothing Model

The Clothing Models dataset on Kaggle is a collection of images of clothing and models wearing the clothing. The dataset consists of 16.2k images and most of them are cloth with models. To fit the data for binary CGAN, I manually labeled data into two groups, only cloth, and cloth with models. Since the data set is highly imbalanced, I end up with 1870 images in each group.

2.2 MNIST

The MNIST dataset is a collection of 70,000 28x28 pixel grayscale images of handwritten digits from 0 to 9, along with their corresponding labels. It is commonly used as a benchmark for evaluating machine learning models, particularly in the field of image classification. The MNIST dataset has also been used for tasks such as object recognition, image generation, and anomaly detection.

3 Method

3.1 Model structure

The architecture of the Generator used for generating kaggle images is a transpose convolution neural network with 3 convolution blocks with filter size (4*4) and stride (2*2) followed by the Relu activation layer.

The generator for MNIST images is simpler, I use the same transpose convolution neural network architecture with only 2 layers of convolution block with filter size (3*3) and stride (2*2) with LeakyRelu activation layer.

To keep the complexity of the generator and discriminator to be the same, I keep the architecture of the discriminator the same as the generator but change the transpose convolution neural into a traditional convolutional layer with the same number of filters and stride and filter size.

3.2 Objective function

$$min_G max_D V(D,G) = E_{x\tilde{}p_{data}(x)}[logD(x)] + E_{z\tilde{}p_z(z)}[log(1 - D(G(z)))]$$

Above is the objective function used for the generator and discriminator. In my experiment, before inputting the latent vector into the generator, label information will be concatenated into it in this form. $(L_1, L_2...L_n, label_1, label_2, ...label_m)$

For the discriminator, since the input is a 3D matrix, the label information is represented by a concatenated channel.

3.3 Training

Cloth and model: Adam optimizer is been used for training both generator and discriminator. And to get a better convergence, I also use learning rate scheduler which start with base learning rate $lr\_discriminator = 0.0003$ and $lr\_generator = 0.0001$, the learning rate will decreased by half every 200 epochs.

MNIST: Also uses Adam optimizer, but with fixed learning rate for both generator and discriminator with $lr\_discriminator = 0.0003$ and $lr\_generator = 0.0003$ and only trained for 20 epochs.

3.4 Metric

Accurately evaluating the quality of these generated images, however, remains an active area of research. Traditional metrics such as mean squared error or structural similarity index do not always accurately reflect the perceived quality of the images, as they are sensitive to low-level image features such as pixel values and may not capture higher-level features such as global structure or semantics. In particular, these metrics may not provide an accurate measure of the quality of the images from a human perspective. To address this issue, the Fréchet Inception Distance (FID) score was proposed as a measure of the quality of generated images. The FID score is based on the idea of using a pre-trained convolutional neural network (CNN) to extract features from images and compare the distribution of these features between the generated images and a reference dataset. The FID score is calculated as the Fréchet distance between the feature distributions, which is a measure of their similarity. If the generated images have low FID score, it means the generated images are considered "close" to the true images.

And in this study, InceptionV3 model which is pre-trained on ImageNet for image classification is loaded for calculating FID score for both dataset.

4 Experiment setup

4.1 Cloth and model

Since the limited computation power and for visualization perpuse, I only trained three GAN model with input dim of 3, first GAN is unconditional GAN model and 3 dimensional input is pure random noise sampled form normal distribution. Second GAN model is a conditional GAN model, the first and second dimension of input is the same random noise sampled form normal distribution and the third dimension is binary label. The third GAN is also an conditional GAN but only the first dimension of input is random noise, the second and the third dimension is softmax label.

After trained these three GAN model, FID score between true images and generated images are then calculated for further analysis.

4.2 MNIST

Since MNIST provides a balanced dataset with 70000 samples and is much easier to train compared with Cloth and model data. I trained 64 CGAN models whose inputs are fixed 10-dimensional softmax label information concatenated with random noise vectors sampled from standard normal distribution whose dimensions vary from 0 to 128. And then FID score is calculated for each of the 64 GANs.

5 Results

5.1 Cloth & model

	UGAN	**CGAN$_2$**	**CGAN$_1$**	**TrueImages**
UGAN	0	223.725	158.855	2283.79
CGAN$_2$	223.725	0	468.738	2077.349
CGAN$_1$	158.855	468.738	0	2489.150
True Images	2283.79	2077.349	2489.150	0

Table 1: FID scores for cloth & model GANs. UGAN is the unconditional GAN, CGAN_2 is the CGAN with two dimensional noise, CGAN_1 is the CGAN with only one dimensional noise.

As shown in the table1, by adding one dimensional label, the generated images have higher quality, and the FID score is decreased by 206. If we change the second an the third dimension to softmax label, and only let one degree of freedom to represent rest of the variety, the quality of generated images then greatly decreased. And also, the latent distribution visualization of the three GAN models are provided in the appendix.

5.2 MNIST

As shown in Fig1, by adding the dimension of latent vectors, generated images by CGANs have better quality. However, starting from some point, the FID score starts to oscillate and stops decreasing. Here in my experiment, this specific point is around 8. This may suggest that except for 10 label, the rest of the information in MNIST dataset can be fully represented by a span of 8 basis vectors.

The red line shows the average FID scores between different batches of true images. Which means the best FID score that generated images can have. The gap between the red line and the green line is caused by architecture of generative and discriminator models. Since the model are used in this study is relatively simple, the gap is around 40, if we can further improve the complexity of GAN model, we can get a better FID score.

6 Conclusion and discussion

6.1 potentially a new way to evaluate information

From above experiment we found that except for 10 label, the rest of the information in MNIST dataset can be fully represented by a span of 8 basis vectors. If we train a same amount of unconditional

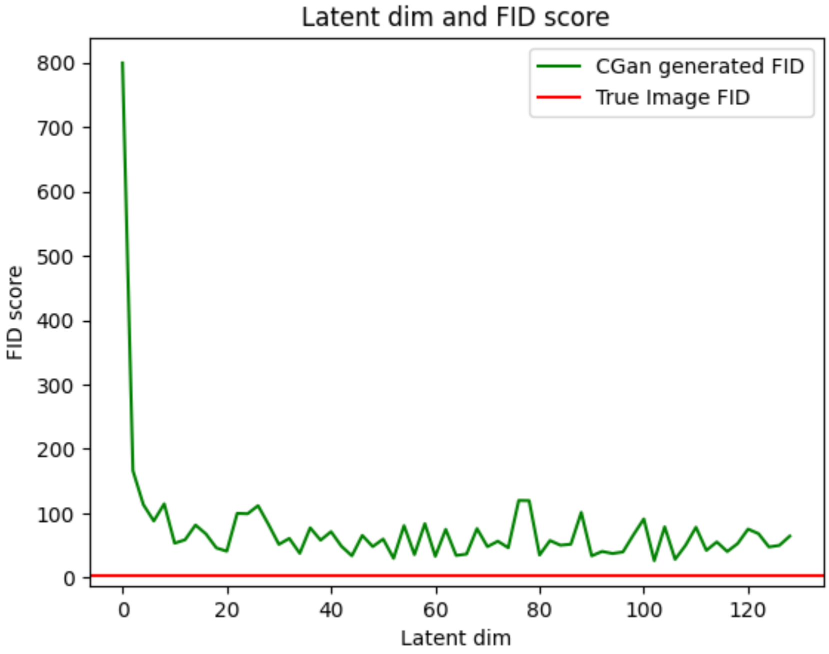

Figure 1: FID scores for 64 CGANs trained on MNIST with different latent dimension, The red line in the plot shows the average FID score between different batchs of true images.

GANs combining the result of the 64 CGANs in this study, we may also find out an optimal degree of freedom to represent the labels.

6.2 Limitation and future work

1, Training a GAN model is hard and not always stable, the result above only based on one trail of experiment. To have a more stable result, we might need run more trails of experiment and use average FID scores. 2, Training GAN model is time-consuming, get such a value is computational intensive. 3, The value get from this process is discrete and I haven't came up with a valid metric to pick such a value.

6.3 Future work

In the future, I want to run more experiment and to see if changing model architecture will affect the trend of FID plot. Since lack of knowledge background, I don't know if there is already a metric to evaluate information contained in a set of data. I want to do more reading on information theory and related theories and comapre it's different form the process I'm using.

References

[1] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567. 4

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661. 4

[3] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. URL https://arxiv.org/abs/1411.1784. 4

7 Appendix

Codes: All codes are available in this repo: DCGAN_latent_vis.

Latent visualization for Kaggle cloth model:

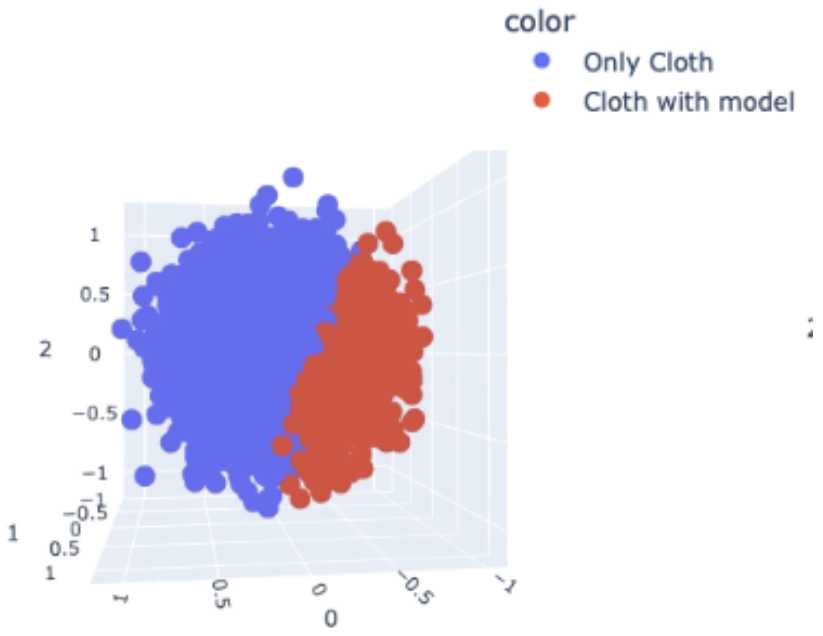

Figure 2: Latent distribution for unconditional GAN

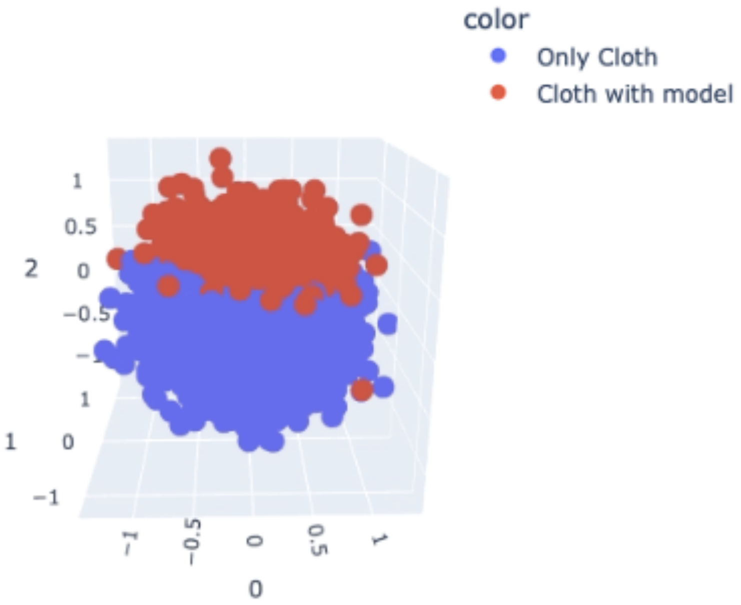

Figure 3: Latent distribution for conditional GAN with 1 dim binary label

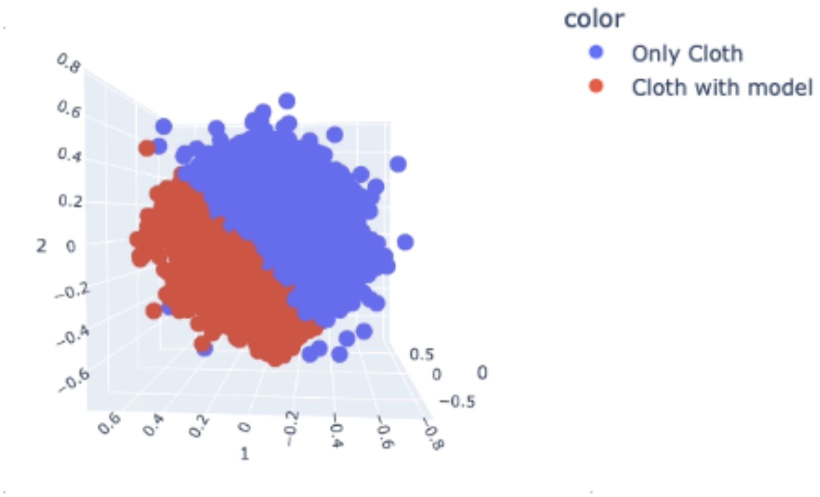

Figure 4: Latent distribution for conditional GAN with 2 dim softmax label