ON THE TRADEOFF BETWEEN MODE COLLAPSE AND SAMPLE QUALITY IN GENERATIVE ADVERSARIAL NETWORKS

Sudarshan Adiga, Mohamed Adel Attia, Wei-Ting Chang and Ravi Tandon

Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, USA
E-mail:{*adiga, madel, wchang, tandonr*}@email.arizona.edu

ABSTRACT

Generative Adversarial Networks (GAN) are used to generate synthetic samples while closely following the underlying distribution of a real data set. While GANs have recently gained significant popularity, they often suffer from the mode collapse problem, where the generated samples lack diversity. Moreover, some approaches that attempt to resolve the model collapse problem do not necessarily yield high quality synthetic samples. In this paper, we propose two novel performance metrics, namely mode-collapse divergence (MCD) which quantifies the extent of mode collapse for a GAN architecture. Second, we propose the metric Generative Quality Score (GQS), which measures the quality of generated samples. We present a comprehensive study of the performance of various GAN architectures proposed in the literature through the lens of MCD and GQS, for three different data sets, namely MNIST, Fashion MNIST and CIFAR-10.

***Index Terms*—** Generative Adversarial Networks, Mode Collapse, Performance Metrics.

1. INTRODUCTION

Generative Adversarial Networks (GAN) is a recently proposed framework to generate synthetic samples [1] from a given data set. GANs have found numerous recent applications, such as video generation [2], image synthesis [3], and generation of synthetic medical records without violating patient privacy [4, 5].

GANs fall in the category of generative models, and comprise of two competing learning networks namely a generator G, and a discriminator D as shown in Figure 1. During the training phase of GAN, the generator samples from the latent noise space ($z \sim P_Z(z)$) and generates a synthetic sample $x_g = G(z)$. The samples from the real data set, $x_r \sim P_{X_r}(x)$ along with the generated samples, x_g are fed to the discriminator. The discriminator determines whether the input samples are real or fake, through the discrimination function $D(x)$. Typically, deep neural networks (and variations) are used to model the functions $G(\cdot)$, and $D(\cdot)$. The parameters of generator and discriminator networks are updated according to a game-theoretic minimax loss function as

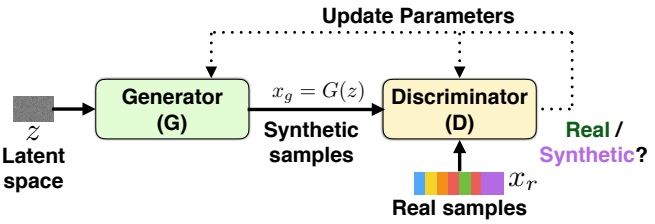

Fig. 1: Generative Adversarial Network (GAN) architecture.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{X_r}(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim P_z(z)}\left[\log(1-D(G(z)))\right] \quad (1)$$

The first term in the loss function is the expected logarithm of $D(x)$, the probability of correctly accepting a real sample, whereas the second term is the expected logarithm of $(1 - D(G(z))$, the probability of correctly rejecting a synthetic sample. This training loss function forces the discriminator to improve its synthetic/real discrimination capability. The feedback from the discriminator is then used by the generator to improve the quality of synthetic samples.

Despite numerous applications, the original formulation of GANs [1] is prone to problems such as mode collapse and mode missing. Mode collapse is defined as a phenomenon in which the generator tends to produce more samples corresponding to a particular mode/class. Mode missing problem is an extension of mode collapse in which some modes are not present in the generated samples. To combat this problem, many alternatives have been proposed, which either involve the modification of the optimization function (e.g. [6, 7]) or modification of the GAN architectures (e.g. [8, 9]). Salimans et al. [10] present the mechanism of mini-batch discrimination as a technique to encourage diversity at the generator's output thereby ensuring convergence during the training of GAN. The mechanism of mini-batch discrimination is also achieved in PacGAN architecture. In PacGAN, multiple generated samples or multiple samples from the real distribution are fed to the discriminator in every iteration of the training phase [9]. In addition to the above methods, an improvement

in the stability during training was observed by the replacement of the loss function with Wasserstein distance (WGAN) and Least Squares loss (LSGAN) [6], [7]. The authors of Conditional GAN (CGAN) propose the generation of samples conditioned on class labels to achieve more control over the modes of the samples being generated [8]. Various algorithms are compared against multiple evaluation metrics like Frechet Inception Distance and Inception Score in [11]. However, despite the advent of numerous algorithms proposed to solve the mode collapse, the quantitative evaluation of the performance of these algorithms with respect to mode collapse as well as sample quality still remains a challenging problem. The inception score [10] attempts to quantify image quality through KL divergence between the conditional distribution (labels given images) and the marginal label distribution for the synthetic samples. Other scores are also proposed in [11], however, the main drawback of existing metrics is that they do not explicitly take into account the mode diversity and the image quality of the underlying data set.

Main Contributions: In this paper, we propose two novel performance metrics, namely mode-collapse divergence (MCD) which quantifies the extent of mode collapse for a GAN architecture. Second, we propose the metric Generative Quality Score (GQS), which measures the quality of generated samples. In contrast to existing approaches, MCD and GQS *explicitly account for the intrinsic* mode diversity and sample quality of the real data set. Our metrics present a systematic approach to readily quantify and compare the performance of various GAN architectures. We present a comprehensive study of the performance of various GAN architectures proposed in the literature through the lens of MCD and GQS, for three different data sets, namely MNIST, Fashion MNIST and CIFAR-10.

2. PERFORMANCE METRICS FOR GAN

As discussed in the introduction, GAN architectures suffer from the problem of mode collapse and may not yield high quality samples. As an example, consider Fig. 3 (a), where we can observe the generated samples from MNIST data set using five different GAN architectures. We can observe that samples from VanillaGAN have high quality, however they suffer from mode collapse (only the digit 1 appears in all generated images). On the other hand, PacGAN/CGAN do not suffer from mode collapse, however, the sample quality of PacGAN is worse compared to the samples from CGAN. We next present the two metrics, namely mode collapse divergence (MCD), and generative quality score (GCS) in order to capture these two fundamental performance aspects of GANs.

2.1. Mode Collapse Divergence (MCD)

In order to define MCD, let us consider P_{Y_r} as the probability distribution of the labels (modes) in the real data set. Similarly, if we consider the set of generated samples, we can estimate P_{Y_g} as the probability distribution of the labels (modes) in the synthetic data set. Mode Collapse divergence (MCD) is defined as the symmetric Kullback-Leibler (KL) divergence[1] between P_{Y_r} and P_{Y_g}:

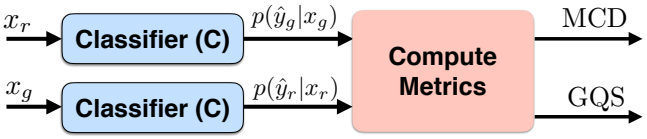

Fig. 2: Computation of the evaluation metrics using a pre-trained classifier.

$$\text{MCD} = \frac{1}{2}\left(\text{KL}(P_{Y_r}||P_{Y_g}) + \text{KL}(P_{Y_g}||P_{Y_r})\right), \quad (2)$$

where the KL divergence between two pmfs $P_1(\cdot)$, and $P_2(\cdot)$ is defined as $\text{KL}(P_1||P_2) = \sum_y P_1(y)\log(P_1(y)/P_2(y))$. The value of MCD is expected to be high with the onset of mode collapse and mode missing, and is expected to be close to zero when the mode distributions of the generated samples and the underlying data set are close to each other.

2.2. Generative Quality Score (GQS)

In order to define GQS, let us consider the real data set, and focus on the joint distribution of the data samples, and labels, i.e., $P_{X_r,Y_r}(x,y)$. The conditional entropy $H(Y_r|X_r)$ then captures the uncertainty in the underlying labels given the data samples. Similarly, if $P_{X_g,Y_g}(x,y)$ is the joint distribution of the generated samples and labels, then $H(Y_g|X_g)$ is the corresponding uncertainty for labels given the synthetic data samples. We define GQS as the following:

$$\text{GQS} = e^{H(Y_r|X_r) - H(Y_g|X_g)}, \quad (3)$$

The justification for defining GQS as above follows from the observation that for a high quality synthetic sample, we should expect $H(Y_g|X_g)$ to be as close to zero as possible, and must also be compared to the real data samples as a baseline. A GQS score in the range $[0,1]$ implies that synthetic images are of lower quality than the real data set. GQS score of greater than 1 implies that the synthetic images are of even higher quality on average than the real data samples, a phenomenon we observe during severe mode collapse.

2.3. Computing MCD and GQS

In order to compute the metrics MCD and GQS, we propose the framework in Fig. 2. First, we input both the real samples X_r and generated samples X_g to a pre-trained classifier (C) and obtain the conditional distributions $P_{\hat{Y}_r|X_r}$ and $P_{\hat{Y}_g|X_g}$.

[1] We remark here that instead of using symmetric KL divergence, any other measure of divergence (or distance) between probability distributions can be used to define the MCD.

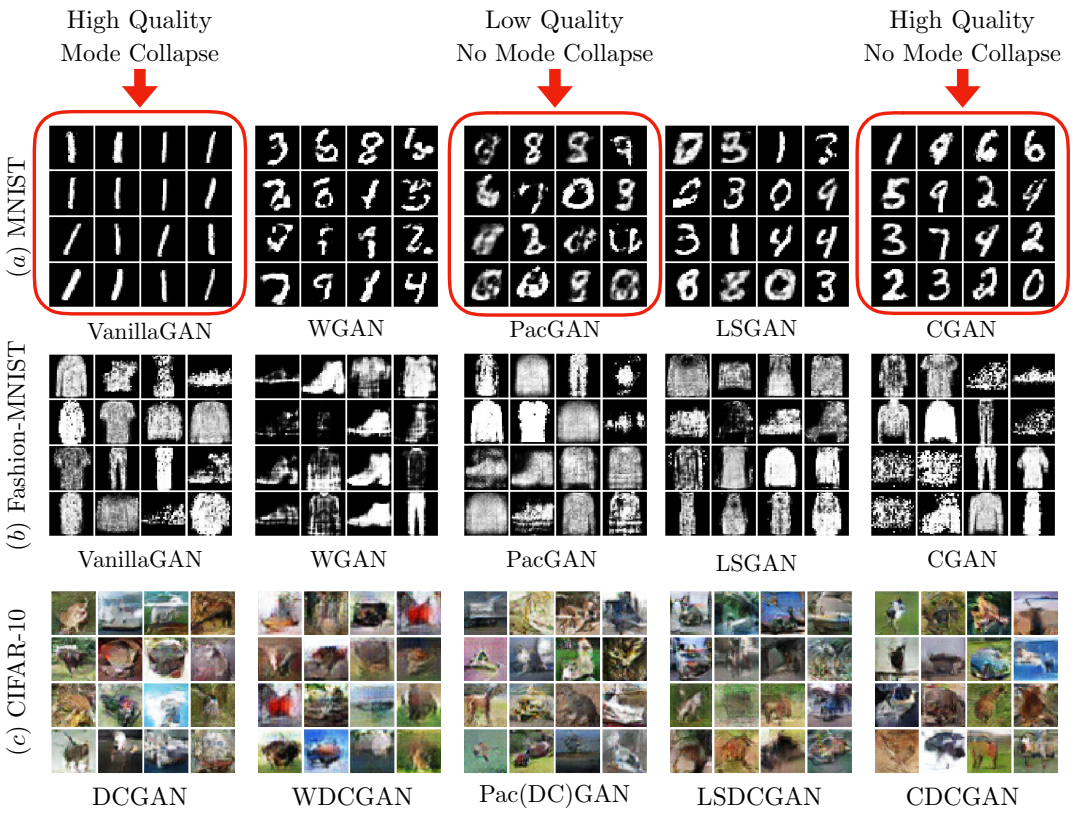

Fig. 3: MNIST, Fashion-MNIST, and CIFAR-10 samples generated by various GAN architechtures after 500,000 iterations of training. By visual assessment of the generated samples, Vanilla GAN suffers from significant mode collapse in the MNIST dataset. We can also assess that the LSGAN and CGAN produces high-quality samples across all the datasets.

From these conditional distributions, we estimate the labels $\hat{Y}_r$ and $\hat{Y}_g$, as well as marginal distributions $P_{\hat{Y}_r}$ and $P_{\hat{Y}_g}$. These distributions are used to compute the MCD using (2), and GQS using (3) (see Algorithm 1). We note here that if the real data set is not labeled, we can then use a classifier, trained in an unsupervised manner (e.g., clustering) for evaluation of the metrics.

Algorithm 1 Computation of MCD and GQS

Input: x_r (real data samples),
x_g (synthetic data samples from the trained GAN)
Output: MCD, GQS
1. Input x_r to classifier to get $P_{\hat{Y}_r|X_r}$; Estimate $\hat{y}_r$ and $P_{\hat{Y}_r}$
2. Input x_g to classifier to get $P_{\hat{Y}_g|X_g}$; Estimate $\hat{y}_g$ and $P_{\hat{Y}_g}$
3. Calculate MCD and GQS using (2) and (3)
return MCD, GQS

3. EXPERIMENTS AND DISCUSSION

In this section, we present our experimental results and compare the following GAN architectures: Vanilla GAN [1], PacGAN [9], Least Squares GAN [7] , WGAN [6], and CGAN [8] through the lens of mode collapse (MCD) and sample quality (GQS) for three data sets, MNIST, Fashion MNIST and CIFAR-10. In particular, for MNIST and Fashion MNIST, we use multi-layer perceptron (MLP) as the learning network for both generator and discriminator. For the CIFAR-10 data set, we use deep convolutional (DC) networks for both generator and discriminator [12]. Fig. 3 shows the samples generated by the various GAN architectures for the three data sets. By visual assessment of the generated samples, Vanilla GAN suffers from significant mode collapse in the MNIST dataset. We can also assess that LS-GAN and Conditional GAN produce high-quality samples across all the data sets.

It is clear that the proposed metrics, MCD and GQS, are dependent on the quality of the underlying classifier used. For the scope of this work, we trained separate RNN classifiers for MNIST and Fashion MNIST data sets (with 97.97% and 97.94% accuracy respectively), and a CNN classifier (83.83% accuracy) for the CIFAR-10 dataset. However, it will be interesting to see how the performance trends change if a better classifier is used to compute the divergence and quality score. This is part of our planned future work.

In Figure 4, we plot MCD and GQS as the number of training iterations increase for the three data sets using the

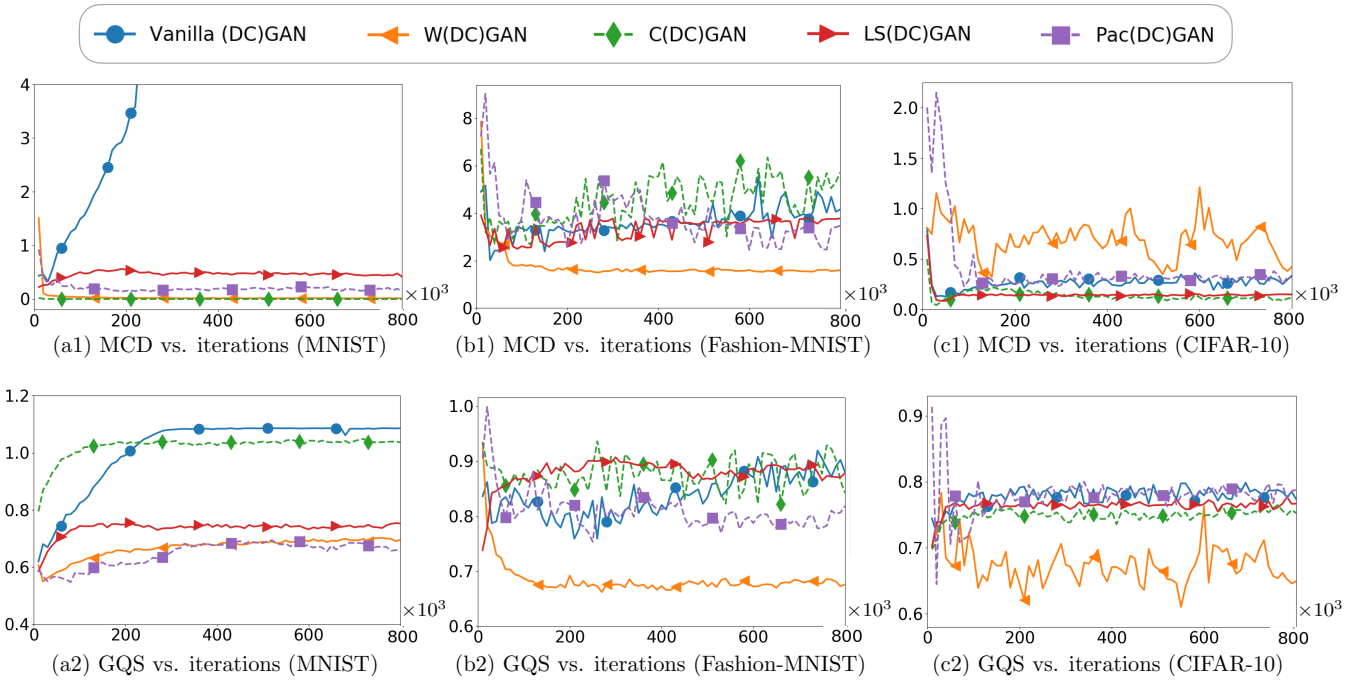

Fig. 4: Mode Collapse Divergence (MCD) and Generative Quality Score (GQS) are plotted versus the number of training iterations for different data-sets and different GAN architectures. Lower MCD and higher GQS values indicate more diverse and higher quality generated samples and vice versa.

five GAN architectures. As seen in Figure 4($a1$), Vanilla GAN undergoes severe mode collapse as reflected by the increasing value of MCD. However, the quality of samples generated by Vanilla GAN is the best compared to other architectures attributing to the highest value of GQS as shown in Figure 4($a2$). Among the remaining architectures, the performance of CGAN for MNIST dataset is found to be the best in terms of both sample quality and mode collapse as the value of MCD is low, and the value of GQS is close to 1.

Similarly from Fig. 4($b1$), we observe the performance of WGAN to be the best for Fashion MNIST dataset in terms of mode collapse as it has a relatively lower value of MCD. However, the performance of LSGAN is found to be better compared to the other algorithms as it undergoes minimal mode collapse and has a higher value of GQS as seen in Figs. 4($b1$) and 4($b2$).

In the experiments with CIFAR-10 data set, we observe that CDCGAN, and LSDCGAN perform better in terms of mode collapse and most of the algorithms perform relatively similar when compared against the quality of samples generated. The performance of the WDCGAN is low for both the mode collapse as well as the quality of samples due to the high value of the MCD and a low value for the GQS score.

4. CONCLUSION

In this paper, we propose two new evaluation metrics to quantitatively measure the performance of Generative Adversarial Networks. In particular, we presented mode collapse divergence (MCD) and generative quality score (GQS) as two principled metrics to capture mode collapse and sample quality. Several GAN architectures, namely, vanilla GAN, WGAN, LSGAN, PacGAN, and CGAN were compared over multiple datasets using the proposed evaluation metrics. We observe that the performance of different algorithms varies significantly with the underlying data set. The metrics to measure the mode collapse and the sample quality agree with the human visual assessment of the generated images. The metrics proposed can thus be useful in selecting an appropriate GAN framework to strike a tradeoff between sample quality and mode collapse. As part of our future work, we plan to do a comprehensive comparison of the proposed metrics with other existing approaches (e.g., Inception Score (IS) [11], Maximum Mean Discrepancies (MMD) [13]).

5. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo, "Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks," *CoRR*, vol. abs/1709.07592, 2017.

[3] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *IEEE Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 5907–5915.

[4] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun, "Generating multi-label discrete electronic health records using generative adversarial networks," *arXiv preprint arXiv:1703.06490*, 2017.

[5] Aleksei Triastcyn and Boi Faltings, "Generating differentially private datasets using gans," *CoRR*, vol. abs/1803.03148, 2018.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[7] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2813–2821.

[8] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[9] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh, "Pacgan: The power of two samples in generative adversarial networks," *arXiv preprint arXiv:1712.04086*, 2017.

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[11] Daniel Jiwoong Im, He Ma, Graham W. Taylor, and Kristin Branson, "Quantitatively evaluating gans with divergences proposed for training," *CoRR*, vol. abs/1803.01045, 2018.

[12] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[13] Eugene Belilovsky, Wacha Bounliphone, Matthew Blaschko, Ioannis Antonoglou, and Arthur Gretton, "A test of relative similarity for model selection in generative models," in *International Conference on Learning Representations*, 2016.