Generative Adverarial Metric

Daniel Jiwoong Im & Roland Memisevic
Montreal Institute for Learning Algorithms
University of Montreal
{imdaniel,memisevr}@iro.umontreal.ca

Chris Dongjoo Kim & Hui Jiang
Department of Engineering and Computer Science
York University
{kimdon20}@gmail.com, {hj}@cse.yorku.ca

Abstract

We introduced a new metric for comparing adversarial networks quantitatively.

1 Model Evaluation: Battle between GANs

A problem with generative adversarial models is that there is not a clear way to evaluate them quantitatively. In the past, Goodfellow et al. (2014) evaluated GANs by looking at the single nearest-neighbour data from the generated samples. LAPGAN was evaluated in the same way, as well as using human inspections (Denton et al., 2015). For human inspections, volunteers were asked to judge whether given images are drawn from the dataset or generated by LAPGAN. In that case, the discriminator can be viewed as a human, while the generator is a trained GAN. The problems with this approach are that human inspectors may have high variance, which makes it necessary to average over a large number of human inspectors, and the experimental setup is both expensive and cumbersome. A third evaluation scheme, used recently by (Radford et al., 2015) is classification performance. However, this approach is rather indirect and relies heavily on the choice of classifier. For example, in mentioned work they used the nearest neighbour classifier, which suffers from the problem that Euclidean distance is not a good dissimilarity measure for images.

Here, we propose an alternative way to evaluate generative adversarial models. Our approach is to directly compare two generative adversarial models by having them engage in a "battle" against each other. The naive intuition is that because every generative adversarial models consists of a discriminator and a generator in pairs, we can exchange the pairs and have them play the generative adversarial game againts each other.

The training and test stage are as follows. Consider two generative adversarial models, M_1 and M_2. Each model consists of a generator and a discriminator,

$$M_1 = \{(G_1, D_1)\} \text{ and } M_2 = \{(G_2, D_2)\}. \tag{1}$$

During the training stage, both models are being trained to prepare them for the battle with one another. Thus, in the training phase, G_1 competes with D_1 in order to be equipped for the battle in the test phase. Likewise for G_2 and D_2. In the test phase, model M_1 plays against model M_2 by having G_1 try to fool D_2 and vice-versa.

Table 1: Model Comparison Metric for GANs

	$\boldsymbol{M_1}$	$\boldsymbol{M_2}$
$\boldsymbol{M_1}$	$D_1(G_1(\mathbf{z}))$, $D_1(\mathbf{x}_{train})$	$D_1(G_1(\mathbf{z}))$, $D_1(\mathbf{x}_{test})$
$\boldsymbol{M_2}$	$D_2(G_2(\mathbf{z}))$, $D_2(\mathbf{x}_{test})$	$D_2(G_2(\mathbf{z}))$, $D_2(\mathbf{x}_{train})$

Accordingly, we end up with the combinations shown in Table 1. Each entry in the table contains two scores, one from discriminating training or test data points and the other from discriminating

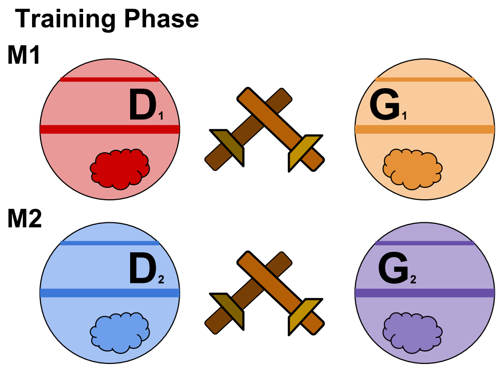

Figure 1: Training Phase of Generative Adversarial Networks.

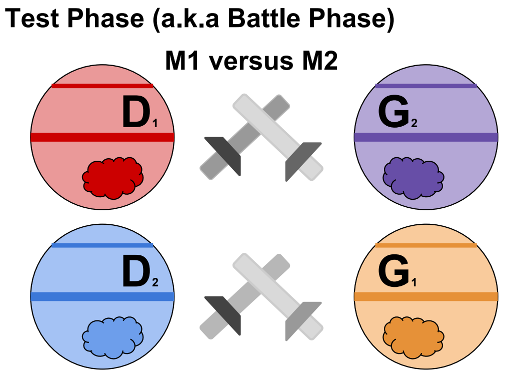

Figure 2: Training Phase and Test Phase of Generative Adversarial Networks.

generated samples. At test time, we can look at the following ratios between the discriminative scores of the two models:

$$r_{test} \overset{\text{def}}{=} \frac{\epsilon\big(D_1(\mathbf{x}_{test})\big)}{\epsilon\big(D_2(\mathbf{x}_{test})\big)} \text{ and} \tag{2}$$

$$r_{samples} \overset{\text{def}}{=} \frac{\epsilon\big(D_1(G_1(\mathbf{z}))\big)}{\epsilon\big(D_2(G_2(\mathbf{z}))\big)}, \tag{3}$$

where $\epsilon(\cdot)$ outputs the classification error rate. These ratios allow us to compare the model performance.

The test ratio, r_{test}, tells us which model generalizes better since it is based on discriminating the test data. Note that when the discriminator is overfit to the training data, the generator will also be affected by this. This would increase the chance of producing biased samples towards the training data, for example.

The sample ratio, r_{sample}, tells us which model can fool the other model more easily, since the discriminators are classifying over the samples generated by their opponents. Strictly speaking, as our goal is to generate good samples, the sample ratio determines which model is better at generating good ("data like") samples. We suggest using the sample ratio to determine the winning model, and to use the test ratio to determine the validity of the outcome as outlined below.

The reason for using the latter is that we cannot decide which model is better solely based on the sample ratio. Consider as a counter example the case where the discriminator of M_1 only outputs false and the generator of M_1 is trained against the discriminator of M_1. On the other hand, M_2 is a model that is trained based on generative adversarial objective.Then, the error rate for D_1 on samples generated by M_2 will be zero. So, M_1 wins since the error rate of M_1 is lower than error rate of M_2. However, M_1 should lose to M_2 since M_2 is obviously not a good model. This problem arises because we have not accounted for the test ratio. To remedy this, our proposed evaluation metric qualifies the sample ratio using the test ratio by defining the winning model as follows:

$$\text{winner} = \begin{cases} \text{M1} & \text{if } r_{sample} < 1 \text{ and } r_{test} \simeq 1 \\ \text{M2} & \text{if } r_{sample} > 1 \text{ and } r_{test} \simeq 1 \\ \text{Tie} & \text{otherwise} \end{cases} \tag{4}$$

We call this evaluation Generative Adversarial Metric (GAM). GAM is not only able to compare generative adversarial models againts each other, but also to partially compare other models, such as the VAE by observing the sample ratio r_{sample} as a evaluation criterion[1].

[1] The demonstration of GAM evaluations can be found at http://arxiv.org/pdf/1602.05110.pdf

REFERENCES

Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2015.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2014.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *http://arxiv.org/pdf/1511.06434.pdf*, 2015.