Evaluation of Generative Adversarial Network Performance Based on Direct Analysis of Generated Images

Shuyue Guan
Department of Biomedical Engineering
The George Washington University Medical Center
Washington DC, USA
frankshuyueguan@gwu.edu

Murray Loew
Department of Biomedical Engineering
The George Washington University Medical Center
Washington DC, USA
loew@gwu.edu

Abstract — Recently, a number of papers have addressed the theory and applications of the Generative Adversarial Network (GAN) in various fields of image processing. Fewer studies, however, have directly evaluated GAN outputs. Those that have been conducted focused on using classification performance and statistical metrics. In this paper, we consider a fundamental way to evaluate GANs by directly analyzing the images they generate, instead of using them as inputs to other classifiers. We consider an ideal GAN according to three aspects: 1) Creativity: non-duplication of the real images. 2) Inheritance: generated images should have the same style, which retains key features of the real images. 3) Diversity: generated images are different from each other. Based on the three aspects, we have designed the Creativity-Inheritance-Diversity (CID) index to evaluate GAN performance. We compared our proposed measures with three commonly used GAN evaluation methods: Inception Score (IS), Fréchet Inception Distance (FID) and 1-Nearest Neighbor classifier (1NNC). In addition, we discuss how the evaluation could help us deepen our understanding of GANs and improve their performance.

Keywords—generative adversarial network; GAN performance measure; GAN evaluation; deep learning; image generation; image analysis; structural similarity index

I. INTRODUCTION

In 2014, Goodfellow et al. [1] introduced the Generative Adversarial Networks (GANs), and it has become a state-of-the-art technique in the field of deep learning [2]. Recently, there have been about 500 types of GANs proposed [3] and a substantial number of studies are about the theory and applications of GANs in various fields of image processing [4]–[7]. Compared to the theoretical progress and applications of GANs, however, fewer studies focus on evaluating or measuring GANs’ performance [8]. Most existing GANs’ measures have been conducted using classification performance (e.g., Inception Score) and statistical metrics (e.g., Fréchet Inception Distance). Our previous study showed that adding generated images from a GAN to training data for a Convolutional Neural Network (CNN) classifier improved its classification performance [9]. Such result indicates that the generated images retain the main features of the real images. It also verifies that using classification is a proper method to evaluate GANs. Alternatively, a more fundamental approach to evaluate a GAN is to directly analyze the images it generated, instead of using them as inputs to other classifiers (e.g. Inception Network) and then analyzing the outcomes. In this study, we try to establish fundamental ways to quantitatively and qualitatively analyze GAN-generated images.

We evaluate the performance of a GAN as an image generator according to these three aspects:

- **Creativity**: non-duplication of the real images. It checks for overfitting by GANs.
- **Inheritance**: generated images should have the same style, which retains key features of the real (input) images. And this is traded off with the creativity property because generated images should not be too similar nor too dissimilar to the real ones.
- **Diversity**: generated images are different from each other. A GAN should not generate a few dissimilar images repeatedly.

A variety of indexes exist in digital image analysis to measure difference or similarity between images, such as image moment, Gray-level Co-occurrence Matrix (GLCM), and Structural Similarity (SSIM) [10] index. Corresponding to the three expectations of ideally generated images, we have designed three measurements based on these image analysis indexes to evaluate GAN performance, and have applied them to three typical GANs: DCGAN [11], WGAN-GP [12] and SNGAN [13].

We briefly summarize three commonly used GAN evaluation methods: Inception Score (IS) [14], Fréchet Inception Distance (FID) [15], and 1-Nearest Neighbor classifier (1NNC) [16], and compare those results with ours. In addition, we discuss how these evaluations could help us to deepen our understanding of GANs and to improve their performance. We define the combination of the three proposed indexes, **CID index**, which is the multiplication of...
Creativity, Inheritance and Diversity indexes. This measure is applied to directly analyze the generated images without using pre-trained classifiers. Results show that CID index reflects the performance of GAN better than the three compared measures. And, it is stable with respect to the number of images and provides explanation of results in three main respects of ideal GANs.

II. METHODS

A. GAN Evaluation Metrics

The optimal GAN for images can generate images that have the same distribution as real samples (used for training), are different from real ones (not duplication), and have variety. Expectations of generated images could be described by three aspects: 1) non-duplication of the real images, 2) generated images should have the same style, which means their distribution is close to that of the real images, and 3) generated images are different from each other. Fig. 1 displays four counterexamples of ideal generated images.

B. Measures for Metrics

Consider a set of real images: \( R = \{ I_r \} \). A GAN has been trained by this real image set and a set of synthetic images were generated from the GAN: \( G = \{ I_g \} \).

The Creativity measure is to find duplications of real images. For \( I_g \), the duplication occurs when the SSIM between \( I_g \) and any \( I_r \) is equal to or greater than 0.8. After removing duplications of real samples from \( G \), \( G^{rem} \) is the set of remaining generated images:

\[
G^{rem} = \{ I_g | \text{SSIM}(I_g, I_r) < 0.8 \}
\]

Creativity index is the percentage of remaining images:

\[
Creativity = \frac{|G^{rem}|}{|G|}
\]

The main idea to measure Inheritance is that same-type images have similar textures. We applied the GLCM-contrast to all real and generated images, and Inheritance index is the difference between average GLCM-contrast value of sets \( R \) and \( G^{rem} \). Specifically, the average GLCM-contrast of \( R = gc_r \); the average GLCM-contrast of \( G^{rem} = gc_g \). To normalize it to the range [0, 1], we define:

\[
Inheritance = 1 - \frac{|gc_r - gc_g|}{\max\{gc_r, gc_g\}}
\]

To compute the Diversity index, SSIM is used again to group similar generated images to the same cluster. Then, we calculate the entropy of these clusters. More clusters and fewer images in each cluster indicate that generated images are more diverse. The optimal condition is that every cluster has only one image; in this case, the entropy value is maximum. Suppose after clustering, we have \( m \) clusters: \( C_1 \cdots C_m \); referring to the Entropy:

\[
Diversity = -\sum_{i=1}^{m} p_i \log p_i; \quad p_i = \frac{|C_i|}{|G^{rem}|}
\]

The product of the three measures is the comprehensive measure (CID index):

\[
CID = Creativity \times Inheritance \times Diversity
\]

C. Related Measures

The Inception Score (IS) [14] is a commonly applied index to evaluate GANs’ performance. To compute the IS, we submit generated images to the Inception network [17] that was pre-trained on the ImageNet [18] dataset. From the perspective of the three aspects for ideal GANs, the IS focuses on measuring the Inheritance and Diversity. Specifically, \( x \in G \) is a generated image; \( y = \text{InceptionNet}(x) \) is the label obtained from the pre-trained Inception network by inputting image \( x \). For all generated images, we have the label set \( Y \). The IS index is defined:

\[
IS(G) = \exp(\mathbb{E}_y[D_{KL}(p(y|x)||p(y))])
\]

Where \( D_{KL} \) is the Kullback–Leibler (KL) divergence of two distributions [19].

The IS reflects Inheritance and Diversity; a larger value of IS indicates that a GAN’s performance is better. The limitations of IS are obvious: it depends on classification of images by the Inception network, which is by trained ImageNet. Thus, it may not be proper to use it on other applications.

1 See Sec.IV for a discussion of this value.
images or for non-classification tasks. Also, since Creativity is not considered by the IS, it has no ability to detect overfitting.

Fréchet Inception Distance (FID) [15] also uses the pre-trained Inception network. Instead of output labels it uses feature vectors from the final pooling layers of InceptionNet. All real and generated images are input to the network to extract their feature vectors.

Let \( \phi() = \text{InceptionNet}_\text{lastPooling}(\cdot) \) be the feature extractor. \( F_r = \phi(R), \quad F_g = \phi(G) \) are two groups of feature vectors extracted from real and generated image sets. Consider the distributions of \( F_r, F_g \) are multivariate Gaussian:

\[
F_r \sim \mathcal{N}(\mu_r, \Sigma_r); \quad F_g \sim \mathcal{N}(\mu_g, \Sigma_g)
\]

The difference of two Gaussians is measured by the Fréchet distance:

\[
\text{FID}(R, G) = \|\mu_r - \mu_g\|_2^2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right)
\]

In fact, FID measures the difference between distributions of real and generated images; that agrees with the goal of GAN training – to minimize the difference between the two distributions. But the Gaussian distribution assumption of feature vectors cannot be guaranteed. And, as with IS, it depends on the pre-trained Inception network.

The measure: 1-Nearest Neighbor classifier (1NNC) [16] does not require an additional classifier. Instead, it uses a two-sample test with the 1-Nearest Neighbor method on real and generated image sets. Similar to FID, 1NNC examines whether two distributions of real and generated image are identical, but it requires the numbers of real and generated images to be equal.

Suppose \(|R| = |G|\), and we wish to compute the leave-one-out (LOO) accuracy of a 1-NN classifier trained on \( R \) and \( G \) with labels “1” for \( R \) and “0” for \( G \). In the optimal situation, the LOO accuracy \( \approx 0.5 \) because the two distributions are very similar. If LOO accuracy \( < 0.5 \), the GAN is overfitting to \( R \) because the generated data are very close to the real samples. In an extreme case, if the GAN memorizes every sample in \( R \) and then generates them identically, i.e., \( G = R \), the accuracy would be 0 because every sample from \( R \) would have its nearest neighbor from \( G \) with zero distance. LOO accuracy \( > 0.5 \) means the two distributions are different (separable). If they are completely separable, the accuracy would be 1.0.

Compared to IS and FID, 1NNC seems a more independent measure. However, the \(|R| = |G|\) requirement limits its applications and the local conditions of distributions will greatly affect the 1-NN classifier.

For IS, higher values imply better performance of GANs; and for FID, lower is better. But for 1NNC, 0.5 is the best score. We regularize 1NNC by this function:

\[
r(x) = -|2x - 1| + 1
\]

\[
r_{1NNC} = r(1NNC)
\]

For r1NNC, the best score is 1.0.

III. EXPERIMENTS & RESULTS

A. One image type by DCGAN

To test the proposed measures, in the first experiment, we used one type of image (Plastics; 12 images) from the USPtex database [20] to train a DCGAN. Then, the trained GAN generated several groups containing different numbers of synthetic images. Finally, we compute our proposed measures, IS, FID, and r1NNC results by using these generated images and 12 real images.

Computations of FID and r1NNC require that the two image sets have the same number of images. We divided the generated images into many 12-image subsets to compute the indexes with 12 real images and then found their average values. Fig. 2 shows the plots of these indexes. FID is scaled by 0.01 to fit the axes. The result indicates that these indexes are stable to different numbers of testing images, especially when the number is greater than 1000.

B. Four image types by three GANs

In the second experiment, four types of image (Holes, Small leaves, Big leaves and Plastics; 12 images for each type) are used to train three GANs (DCGAN, WGAN-GP and SNGAN). Then, the trained GANs generated 1,200 synthetic images for each type. Twelve sets of synthetic images were generated; Fig. 3 shows samples from 4 real image sets and 12 generated image sets. Visual examination of these synthetic images indicates that the DCGAN seems to give the most images similar to the real ones, but many of its generated images are duplicates of real ones. Thus, the DCGAN overfitted the training data. The SNGAN’s generated images are most dissimilar to real images; it lacks Inheritance feature. The WGAN-GP, however, well balanced the Creativity and Inheritance features.

Finally, we applied these measures on the 12 generated image sets. Fig. 4 shows plots of results. To emphasize the order of each index for different generators and image type, values are normalized from 0 to 1 by columns for plotting.
To compare the three GANs, Table I shows summarized results and Fig. 4 gives more details. For the best generator, the proposed CID index agrees with IS and 1NNC and the visual appearance of generated images. Since DCGAN overfitted to training data, its Creativity score is the lowest. CID reflects such limitation of DCGAN but IS, FID and 1NNC do not. FID shows DCGAN is the best, however, its Creativity index is the lowest.

**TABLE I. MEASURE RESULTS AVERAGED BY GENERATORS**

<table>
<thead>
<tr>
<th>*</th>
<th>C</th>
<th>I</th>
<th>D</th>
<th>CID</th>
<th>IS</th>
<th>FID</th>
<th>r1NNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>0.135</td>
<td>0.869</td>
<td>2.920</td>
<td>0.330</td>
<td>1.285</td>
<td>143.223</td>
<td>0.553</td>
</tr>
<tr>
<td>W</td>
<td>1.000</td>
<td>0.820</td>
<td>6.524</td>
<td>5.341</td>
<td>1.406</td>
<td>380.642</td>
<td>0.897</td>
</tr>
<tr>
<td>SN</td>
<td>0.994</td>
<td>0.886</td>
<td>1.050</td>
<td>0.993</td>
<td>1.135</td>
<td>416.295</td>
<td>0.267</td>
</tr>
</tbody>
</table>

* Generator models: DC: DCGAN; W: WGAN-GP; SN: SNGAN. **Bold**: the highest score in column. **Underline**: the lowest score in column. *FID*: lower is better.
REFERENCES


