Evaluation of Generative Adversarial Network Performance Based on Direct Analysis of Generated Images
Shuyue Guan and Murray Loew
Department of Biomedical Engineering, George Washington University
Washington, D.C. USA

Motivation
The Generative Adversarial Network (GAN) is a state-of-the-art technique in the field of deep learning. GAN is a neural-network based generative model that learns the probability distribution of real data and creates simulated data with a similar distribution.

Recently, a substantial number of studies are about the theory and applications of GANs in various fields of image processing. Fewer studies, however, have directly evaluated GAN outputs. A fundamental way to evaluate GANs is to directly analyze the images they generate, instead of using them as inputs to other neural networks (CNN classifiers) and then analyzing the outcomes. In this study, therefore, we try to establish elemental ways to quantitatively and qualitatively analyze GAN-generated images.

Evaluation Metrics

Inheritance: keep the same style & key features of real images.

Creativity: not duplication of real images (no overfitting)

Diversity: more different generated images, no mode collapse and dropping

Methods

Creativity Measure
To find duplication of real images.

1. Generate N images from GAN: \( \mathcal{G} = \{g_i\} \)
2. Compare each generated image \( g_i \) with every real image \( r_j \) by Structural Similarity (SSIM)
3. If any \( \text{SSIM}(g_i, r_j) \geq 0.8 \), remove the \( g_i \)
4. Finally, \( m \) generated images remain: \( G'_{\text{rem}} \)

\[ \text{Creativity index} = \frac{G'_{\text{rem}}}{N} \]

Notations
Set of remaining generated images:

\[ G'_{\text{rem}} = \{I_g | \text{SSIM}(I_g, r_i) < 0.8\} \]

Inheritance Measure

Same-type images have similar textures.

1. Choose a texture measure for images:
   - Grey-Level Co-occurrence Matrix (GLCM)-contrast
   \[ \sum_{i,j} P_{i,j} (i-j)^2 \]

2. Compute the average GLCM-contrast for real and remaining generated images:
   - Average GLCM-contrast of real: \( \bar{g}g_r \)
   - Average GLCM-contrast of \( G'_{\text{rem}} \): \( g_{g_{\text{rem}}} \)

3. To make it range \([0, 1]\):

\[ \text{Inheritance index} = 1 - \frac{\bar{g}g_r - g_{g_{\text{rem}}}}{\max\{g_{g_r}, g_{g_{\text{rem}}}\}} \]

Diversity Measure

To measure clusters containing similar images.

1. In \( G'_{\text{rem}} \), select one image \( I_{g_1} \) and move to \( C_1 \)
2. If any \( \text{SSIM}(I_{g_2}, I_{g_1}) \geq 0.8 \), move to \( C_1 \)
3. Until \( G'_{\text{rem}} = \emptyset \) to have \( m \) clusters: \( C_1 \ldots C_m \)
4. Refer to Entropy

\[ \text{Diversity index} = -\sum_{i=1}^{m} p_i \log p_i \]

CID index
To multiply the 3 measures together to obtain the final measure:

\[ \text{CID} = \text{Creativity} \times \text{Inheritance} \times \text{Diversity} \]

Experiments & Results

4 training groups: 4 types real images from the USPtex1.0 database. Each group has 12 images.

3 types GAN were trained:
- DCGAN, WGAN-GP, SNGAN
- GAN could generate unlimited number of images but with replication. We generated 1200 images, which are 100 times of real ones, for each type.
- For comparison, besides the proposed 4 measures, we computed these measures:
  - Inception score (IS)
  - Fréchet Inception Distance (FID)
  - 1-Nearest Neighbour classifier (1NNC)

Conclusion
CID index directly analyzes the generated images without using pre-trained classifier and better reflects the performance of GAN than some other measures. It is stable to the number of images and provides explanation of results in 3 main respects of good GAN.