International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7248
***Article Received**: 28 April 2023 **Revised:** 18 June 2023 **Accepted:** 02 July 2023*

Evaluating Text-to-Image GANs Performance: A Comparative Analysis of Evaluation Metrics

[1]K. Dinesh Kumar, [2]Dr. Sarot Srang, [3]Dr. Dona Valy
[1]Mechatronics and Information Technology, Institute of Technology of Cambodia, Phnom Penh, Cambodia.
vkjdinesh@gmail.com
[2]Mechatronics and Information Technology, Institute of Technology of Cambodia, Phnom Penh, Cambodia.
srangsarot@itc.edu.kh
[3]Department of Information and Communication Engineering, Institute of Technology of Cambodia, Phnom Penh, Cambodia.
dona@itc.edu.kh

Abstract— Generative Adversarial Networks (GANs) have emerged as powerful techniques for generating high-quality images in various domains but assessing how realistic the generated images are is a challenging task. To address this issue, researchers have proposed a variety of evaluation metrics for GANs, each with its own strengths and limitations. This paper presents a comprehensive analysis of popular GAN evaluation metrics, including FID, Mode Score, Inception Score, MMD, PSNR, and SSIM. The strengths, weaknesses, and calculation processes of these metrics are discussed, focusing on assessing image fidelity and diversity. Two approaches, pixel distance, and feature distance, are employed to measure image similarity, while the importance of evaluating individual objects using input captions is emphasized. Experimental results on a basic GAN trained on the MNIST dataset demonstrate improvement in various metrics across different epochs. The FID score decreases from 497.54594 at Epoch 0 to 136.91156 at Epoch 100, indicating improved differentiation between real and generated images. In addition, the Inception Score increases from 1.1533 to 1.6408, reflecting enhanced image quality and diversity. These findings highlight the effectiveness of the GAN model in generating more realistic and diverse images with training progression. However, when it comes to evaluating GANs on complex datasets, challenges arise, highlighting the need to combine evaluation metrics with visual inspection and subjective measures of image quality. By adopting a comprehensive evaluation approach, researchers can gain a deeper understanding of GAN performance and guide the development of advanced models.

Keywords - GAN, Evaluation Metrics, IS, FID, LPIPS, GMSD, MSE, PSNR, MNIST.

I. INTRODUCTION

Generative Adversarial Networks (GANs) are one of the best generative models for producing high-quality output in various domains, including images, videos, and audio [1]. GANs consist of a generator and a discriminator, both are trained to learn a mapping between a low-dimensional noise vector and the high-dimensional space of the target data. While GANs have shown impressive results in generating realistic-looking data, evaluating their performance and comparing different models can be challenging.

Various metrics have been proposed for GANs to quantify the quality and diversity of generated data. However, no specific evaluation metric captures all aspects of GAN performance, and different metrics may prioritize different aspects of image quality or have different sensitivities to specific types of distortions or artifacts. Therefore, it is important to use a combination of metrics to get a more comprehensive and accurate evaluation of GAN performance [3][8][14].

The quality of an image synthesis technique is determined by its ability to adhere to user input and produce photorealistic and structurally coherent output, while also generating a diverse set of images which meet the requirements. To evaluate the generated image quality and diversity, various general metrics have been developed. These widely used metrics employ different approaches to extract features vectors/code from images, compute scores or distances, including Peak Signal-to-Noise Ratio (PSNR), Inception Score (IS), Fréchet inception distance (FID), structural similarity index measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), Maximum Mean Discrepancy.

In this study, we compare and analyze several popular GAN evaluation metrics. We discuss the pros and cons of each metric and the process of calculating the score of each. Our goal is to provide a clear overview of GAN evaluation metrics and their suitability for different applications. By doing so, we aim to help researchers and practitioners choose appropriate evaluation metrics for their specific use cases and avoid common pitfalls in GAN evaluation.

II. METHODOLOGY

Developing generative models with GANs is one of the most popular areas of research in the field of deep learning. Various metrics have been proposed to evaluate the effectiveness of GAN models. However, currently, there is no

***Article Received**: 28 April 2023 **Revised:** 18 June 2023 **Accepted:** 02 July 2023*

globally accepted metric for evaluating GANs. Evaluation of GANs considers two key properties: fidelity and diversity.

High fidelity indicates that the generated images closely resemble real images, while low fidelity results in distorted and blurry images.

High diversity generates a wide range of images with varying styles and appearances, whereas low diversity produces repetitive and similar images. Achieving a balance between fidelity and diversity is a significant challenge in GAN research, and various approaches have been proposed to address this challenge, such as regularization methods, loss functions, and architecture designs.

The two approaches Pixel Distance and Feature Distance are used to compute the image similarity between the actual images and generated synthetic images of GAN.

Pixel Distance: Pixel-wise Mean Square Error (MSE) is utilized in measuring the difference of pixel values of generated and real images. It is used to process the images and does not take the perceptual quality of the generated images.

Feature Distance: The feature extraction method is used to extract higher-level information such as shapes of objects, image structure, and texture by using the pre-trained neural network of the generated synthetic images and real images. It is evaluating the perceptual quality of the generated synthetic images.

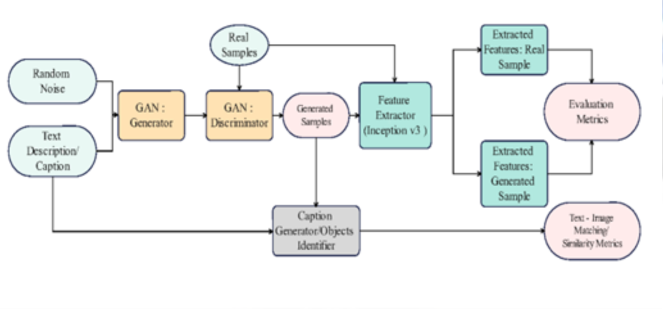
Real Samples
Random Noise
Text Description/ Caption
GAN : Generator
GAN : Discriminator
Generated Samples
Feature Extractor (Inception v3)
Extracted Features: Real Sample
Extracted Features: Generated Sample
Evaluation Metrics
Caption Generator/Objects Identifier
Text - Image Matching/ Similarity Metrics

Figure 1. The overall process of finding evaluation metrics score

The evaluation metrics assess both the quality and diversity of generated synthetic images. Moreover, to achieve higher accuracy, it is crucial to measure individual objects in the generated images according to the input caption. As a result, a pre-trained caption or object detector is utilized to compare the input caption with the objects/caption in the generated image. This comparison is performed using evaluation metrics to enhance the semantic comprehension of generated images.

In the overall process flow (Figure 1), a GAN is trained to generate images based on random noise and text description as inputs. The generated images are then fed into a discriminator, along with real images, to assess their quality. To evaluate the performance of the GAN, the generated samples and real samples are also passed through an Inception v3 model to extract features that can be used with various evaluation metrics. Additionally, the objects in both the generated and real samples/text descriptions are identified using an object identifier. Finally, the generated images are compared to the text descriptions using text-to-image matching or text similarity measures to further evaluate the GAN performance.

III. GAN EVALUATION METRICS

A. Inception Score (IS)

Inception Score (IS) is a single floating-point number that evaluates the list of generated images, providing a measure of their quality or how realistic they appear. This metric is known to be correlated with human evaluation [2]. However, it is worth noting that IS only assesses image quality and does not capture how well the generated synthetic image matches the actual image [3]. IS measures the generated synthetic image and actual image quality.

The Inception Score (IS) evaluates the generated images as follows:

1. Diversity of generated images
2. How clearly each image resembles a particular object or scene.

If both conditions are met, the Inception Score (IS) will be high, but if either one is not satisfied, the score will be low. The range of Inception Scores is infinite, with a score of 0 being the lowest possible and higher scores indicating better quality and more diverse images. An Inception v3 model pre-trained on the "ImageNet" dataset, which consists of 1000 distinct objects for image classification, can be utilized to calculate the Inception Score. To determine the quality and set of image diversity, the score is computed by taking into account the confidence of conditional class prediction for each generated image (image quality), the integral of the marginal distribution of the predicted classes (diversity). In essence, the Inception Score provides an estimation of the quality and image diversity.

The Inception Score processes the desirable properties of generated images, such as high classifiability and diversity, using an Inception v3 model which is a pre-trained image classification model. The score range is from 0 to 1000, which is the number of known object categories in Inception v3. If a generated image falls into one of these categories, the highest score is returned. However, the Inception Score value depends on both the number of classes in the generated image and the known 1000 classes in Inception v3. CIFAR-10 dataset consists of 10 classes of objects which contain 50,000 images in total and the Inception score for the actual CIFAR-10 is 11.24 +/- 0.12 [5].

The Kullback-Leibler (KL) divergence distinguishes between two probability distributions. For a conditional label distribution p(y|x), an Inception model is applied to each generated image. The entropy of conditional label distribution should be low for images containing objects with meaning.

Marginal distribution has high entropy to generate diverse images. The Inception score metric combines these two distributions to assess the generated image quality.

$$\text{IS} = e^{\mathbb{E}(KL(\text{p (y|x)||p (y)}))}$$

The Inception score computes by taking input image x and label y predicted using the inception model. It indicates the diversity and semantic meaning of generated synthetic images. This metric measures the KL divergence of all samples between the p(y/x) is the conditional distribution over labels given an image and the p(y) is the marginal distribution over labels averaged across all images, $\mathbb{E}$ is the expectation that is calculated by the average across multiple generated images. To measure diversity, [4] used a large number of samples (such as 50,000). However, the Inception score does not consider the artistic value or contextual information of the image. Instead, it evaluates how distinctive each object is in the image (low entropy), and the total how many numbers of different objects the GAN can generate (high entropy).

Inception Score will not consider object similarity of the same class, which means that a network can generate a "perfect" sample for every class that can get a high IS, even if it exhibits the intra-class mode dropping behavior.

B. *Mode Score(MS)*

Mode score is the extended version of Inception Score.

$$\text{MS} = e^{(\mathbb{E}[KL(\text{p (y|x)||p (y}*\text{))}])} - KL(\text{p (y)||p (y} *\text{))}$$

$$\text{p (y} *\text{)} = \int_x \text{p (y|x)} \; d\mathbb{P}_r$$

p(y $*$) is the empirical distribution of labels computed from training data and $\mathbb{P}_r$ is the real distribution. Mode score which is measured by $KL(\text{p (y)||p (y} *\text{))}$ [6], provides a more effective method for comparing the generated synthetic image to the actual image compared to Inception Score. This is because the mode score measures the divergence between the actual distribution and generated distribution.

C. *Frechet Inception Distance (FID)*

Aim of FID score, is to evaluate the generated images quality by differentiating the stat of generated image set with the real image set from the target domain. FID is a metric measures the distance between extracted features of actual and generated synthetic image. It measures the similarity between the extracted feature statistics of actual and generated synthetic images based on computer vision features of the raw input mages extracted using the Inception v3 network. A lower FID score denotes higher similarity between two image sets, with a perfect score of 0 indicating identical image groups. Lower FID scores correspond to better image quality, and the relationship may be linear, with higher scores indicating lower image quality.

The step-by-step process of calculating the FID

Step 1: Load the Inception v3 which is pretrained model

○ The last layer of Inception v3 is removed and the activation output is taken as the last pooling layer.

○ Since the last layer has 2048 activations, each image is prediction has 2048 activation features. This is considered as the feature/coding vector of the images [3].

Step 2: Calculate feature vectors for images

○ The given input images need to be preprocessed accordingly.

○ A 2048 feature vectors is predicted for a group of actual images and the generated synthetic images

Step 3: Calculate the FID score

○ For “univariate” normal distribution FID [3] is given as

$$FID_u = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2$$

▪ μ *and* σ: mean and standard deviation of the normal distribution

○ For “multivariate” normal distribution FID [3] is given as

$$FID_m = \|\mu_x - \mu_y\|^2 + Tr\left(\Sigma_x - \Sigma_y - 2\sqrt{\Sigma_x \Sigma_y}\right)$$

▪ x *and* y : The generated and real embedding (activation from Inception Model)

▪ $\mu_x - \mu_y$: Magnitude of x and y

▪ Tr : Trace of the matrix

▪ $\Sigma_x \Sigma_y$: Covariance of the matrix of the vectors

The FID calculates the distance in feature extraction between generated synthetic images and a reference set of images, which is often the validation set, to determine the degree of similarity between them.

D. *Kernel Maximum Mean Discrepancy (K-MMD)*

The K-MMD metric is utilized for quantifying between two probability distributions, including the actual data distribution and the generated samples from a GAN.

To calculate K-MMD, the data is transformed into higher dimension future space using kernel function. Next, distance between the mean embeddings of the distributions in that space is computed. The mean embedding is the average value of a kernel function assessed at all samples taken from the distribution

The K-MMD distance is then given by [6]:

$$\text{K-MMD}^2 = \mathbb{E}[k(x,x')] + \mathbb{E}[k(y,y')] - 2\mathbb{E}[k(x,y]$$

where the kernel function k(x, x') is evaluated at all pairs of samples x and x', the kernel function k(y, y') is evaluated at all pairs of samples y and y', and the kernel function k(x, y) is evaluated at all pairs of samples x and y[6].

In order to calculate K-MMD, the mean embeddings of the distributions must be estimated from a finite set of samples, which can be accomplished using the kernel trick. This allows

***Article Received**: 28 April 2023 **Revised:** 18 June 2023 **Accepted:** 02 July 2023*

the kernel function will be computed for all pairs of samples without needing to explicitly compute the higher-dimensional feature space.

K-MMD is a valuable evaluation metric for GANs as it provides a quantitative measure of the dissimilarity between the generated samples and the actual data distribution, without necessitating labeled data.

E. Peak Signal to Noise Ratio (PSNR)

Peak Signal to Noise Ratio (PSNR) metric for GAN evaluates the generated images quality by comparing them to the original images. It is considered a robust metric that is capable of capturing information loss and contrast changes in the image [7].

$$PSNR = 10.\log_{10}\left(\frac{MAX^2}{MSE}\right)$$

MAX is the maximum possible pixel value in the original image. It is usually calculated as $2^n - 1$, n denotes the how many bits per pixel.

$$PSNR = 10.\log_{10}\left(\frac{(2^n-1)^2}{MSE}\right)$$

$$MSE = \frac{1}{N^2}\sum_{i-1}^{N}\sum_{j-1}^{N}(x(i,j) - y(i,j))^2$$

$$MSE = \frac{1}{N^2}\sum\sum(x-y)^2$$

x- Generated Image, y- Ground Truth Image,
N – Image Size, i – index of the row,
j- index of the column

The PSNR metric in GANs is evaluate the generated image quality by calculating the amount of information loss or contrast change with respect to the original images. It quantifies the relative differences between pixels in the original and reconstructed images, and is expressed as the maximum signal power to noise power ratio, typically measured in decibels (dB). A higher PSNR score indicates better contrast representation and less information loss. The PSNR score typically ranges from 20 dB to 40 dB, with higher values indicating more accurate content capture of the original source text. To obtain a more stable result, increase the number of samples.

F. Structural Similarity Index Measure (SSIM)

Structural Similarity Index Measure (SSIM) compares luminance, contrast, and structure of two images to determine their similarity. This metric can be used to evaluate how well a GAN captures the content of a real image and to ensure that the generated image appears realistic [8].

Luminance:

The arithmetic mean of pixels is denoted by µ[8].

$$\mu_x = \frac{1}{N}\sum_{i-1}^{N} x_i$$

N: Number of pixels in the image x

xi: What is the pixel value in ith position in image x

Contrast:

The contrast is measured by calculating the standard deviation of pixels and is denoted by σ[8].

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2\right)^{\frac{1}{2}}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1) + (2\sigma_{xy} + C_2)}{(\mu_x{}^2 + \mu_y{}^2 + C_1)(\sigma_x{}^2 + \sigma_y{}^2 + C_2)}$$

x- Generated Image,
y- Ground Truth Image,
µ: mean, and σ: variance
$\sigma_{xy}$is the covariance of the two images;
The constants (C_1, C_2) used to stabilize the division.

$$C_1 = (0.01 - MAX)^2 \quad C_2 = (0.03 - MAX)^2$$

The evaluation metric SSIM determines similarity between two input images and generates a value ranging from -1 to +1. If the SSIM score is +1, it implies that the two input images are almost identical, whereas a score of -1 indicates that they are substantially different [9]. Nevertheless, metrics like Peak Signal-to-Noise Ratio (PSNR) and Mean Squared Error (MSE) only consider the image intensity and do not always align well with subjective fidelity ratings.

G. Feature Similarity Index Measure(FSIM)

FSIM is used for assessing the generated image quality in GANs. Unlike SSIM, which computes the similarity between two images based on pixel values, FSIM compares the feature representation of two images. It calculates the image similarity between two images in terms of their feature maps, which are obtained by applying a bank of filters to the image and computing the response of each filter. FSIM is robust to changes in illumination, contrast, and noise. FSIM measures the ratio of the joint statistics of feature maps and individual statistics of the feature maps. FSIM uses a bank of filters, that is to capture the structural information in an image, while SSIM uses Gaussian filters. FSIM is calculated based on two properties, such as “Phase Congruency (PC) and Gradient Magnitude (GM)”.

Phase Congruency (PC) is a technique used to detect image features, and it is invariant to variation of light in the image. One of the key characteristics of PC is that it highlights image features in the frequency domain. Additionally, PC is invariant to the contrast changes.

Gradient Magnitude (GM): The image gradient computation is performed using Gradient Magnitude (GM) in image processing. Various convolutional masks used to express gradient operators. Various convolutional masks available to calculate the gradients, with Gx and Gy representing horizontal and vertical gradients, respectively, for an image f(x).

Gradient magnitude of f(x) is defined as

$$G = \sqrt{{G_x}^2 + {G_y}^2}$$

The phase congruency maps of f1 (test image) and f2 (reference image) denoted by PC1 and PC2, respectively. In addition, the magnitude maps G1 and G2 were extracted from these images. FSIM is then calculated using PC1, PC2, G1, and G2. The image similarity is calculated as

$$S_{PC} = \frac{2PC_1PC_2 + T_1}{{PC_1}^2 + {PC_2}^2 + T_1}$$

The constant T1 is introduced to increasing S_{PC} stability. T1 calculated based on the dynamic range of values of PC.

$G_1(x)$ represents the gradient magnitude value at position x in the magnitude map G1, and $G_2(x)$ represents the gradient magnitude value at position x in the magnitude map G2. These values are used in the similarity measures calculation to assess the similarity between the test image and the reference image based on their gradient magnitude.GM values G1(x) and G2(x) are compared and the similarity measure is [12]

$$S_G(x) = \frac{2G_1(x).G_2(x) + T_2}{{G_1}^2(x) + {G_2}^2(x) + T_2}$$

T_2 is a constant depending on the range of values of GM.

$S_{PC}(x)$, $S_G(x)$are used to measure $S_L(x)$of f1(x) and f2(x):

$$S_L(x) = S_{PC}(x).S_G(x)$$

The locations where edges occur are more important in conveying visual information compared to smooth areas. Phase Congruency (PC) is used to measure the importance of a local structure. Basically, if the PC value is high for either f1(x) or f2(x) at a given location x, it means that the HVS (Human Visual System) will consider that location to be a significant factor in determining the similarity between f1 and f2.

$$PC_m(x) = max(PC_1(x), PC_2(x))$$

FSIM index is defined as

$$FSIM = \frac{\sum_{x\epsilon\Omega} S_L(x).PC_m(x)}{\sum_{x\epsilon\Omega} PC_m(x)}$$

Here, Ω refers to the spatial domain of the image [12].

H. R-Precision

The R-Precision evaluation metric is utilized to evaluate the correspondence between text descriptions and generated images. It ranks the similarity between the actual caption of a given generated image and randomly selected captions to assess visual semantic similarity. The image and 99 captions are encoded using text and image encoders, and calculate their cosine similarity.

To calculate R-Precision, captions are ranked based on their similarity to the generated images, and the top r most similar captions (usually r = 1) are considered. It will be arranged in the decreasing similarity. This model embeds the images and the text description. “Cosine distance between the matching image caption pair minimized while the cousin distance between the mismatching caption pair maximized”. A drawback of R-Precision, however, is that it does not assess the individual objects quality, but focuses on the global background and salient features [13].

I. The Visual Semantic (VS) similarity

The alignment between generated images and text can be measured using VS similarity metric, that calculates the distance of the two modalities through a trained VS embedding model [10]. This is achieved by learning two mapping functions that map text and images to a common representation space. VS similarity is calculates based on the image and text encoders.

$$VS = \frac{f_t(t).f_x(x)}{\|f_t(t)\|_2.\|f_x(x)\|_2}$$

$f_t(.) \rightarrow$ Text Encoder, $f_x(.) \rightarrow$ Image Encoder

The VS score has a high standard deviation even for real images, which limits its precision in evaluating the model performance [11].

J. Caption Generation

To assess the quality of text to image models, one commonly used approach is caption generation. In this method, the generated images are compared with their original captions to ensure their relevance. If the generated images accurately reflect their captions, it should be possible to deduce the actual captions from them. To achieve this, a pre-trained caption generator can be employed to generate captions for each generated image, and then standard language similarity metrics such as BLEU, METEOR, and CIDEr can be used to compare the generated captions with the original ones [10].

An inherent challenge of using caption generation for text to image model evaluation is that multiple valid captions can exist for a given image. The dissimilarity between two captions does not necessarily indicate that they do not describe the same image. To address this challenge, either the real or generated captions should emphasize specific objects in the image. Captions that describe the overall layout of a scene without explicitly mentioning specific objects can also exist. To account for semantic content and objects in the scene, Semantic Object Accuracy metric can be employed alongside language similarity metrics [10].

K. Structural Objective Analysis(SOA)

Evaluation metrics for generated images usually focus on the overall quality of the image, rather than on individual objects or areas within the image. Only R-Precision and Caption Generation considered the image caption while evaluating the generated images.

The SAO metric is an evaluation tool that uses an object detector which is pretrained to assess whether an image has the objects mentioned in its caption. It aims to measure the accuracy with which an image represents the objects referred to in its caption and is therefore called led SOA.

For computing the Semantic Object Accuracy (SOA) for the COCO dataset, first filter the validation set captions based on keywords that are related to the object label. Next, the process involves identifying all captions for each of the 80 labels in the COCO dataset that indicate the presence of the object, and generating three images for each caption. Pre-trained YOLOv3 network is then run on COCO dataset for each of the generated images to check whether the network can recognize the specific object

SOA-C (Class average) measures average number of images in each class that the YOLOv3 network can recognize a specific object. SOA-I (Image average) measures the average number of images that contain desired object that the YOLOv3 network correctly detects.

$$\text{SOA-C} = \frac{1}{|C|}\sum_{c\epsilon \mathrm{C}}\frac{1}{|I_c|}\sum_{i_c\epsilon I_c} YOLOv3(i_c)$$

$$\text{SOA-I} = \frac{1}{\sum_{c\epsilon C}|I_c|}\sum_{c\epsilon \mathrm{C}}\sum_{i_c\epsilon I_c} YOLOv3(i_c)$$

$$YOLOv3(i_c) = \begin{cases} 1\ if\ YOLOv3\ detected\ an\ object\ of\ class\ c \\ 0,\ \ otherwise \end{cases}$$

for classes of object c ∈ C. i ∈ Ic has an object of class c. Many images contain objects which are not exactly mentioned on caption and it did not measure the false negative's rate bit rather focus on the recall that is True Positive.

Certain evaluation metrics incorporate scene layouts and bounding boxes, and one such metric is the Semantic Object Accuracy (SOA) that computes the Intersection Over Union (IoU) between the real and predicted regions of interest. This is referred to as SOA-IoU. The IoU measures the overlap between the real and predicted bounding boxes by calculating the ratio of intersection to their union. In order to determine IoU, YOLOv3 networks detects the objects in each image. As multiple images include multiple instances of the same object, IOU calculated for the predicted bounding box for the input object and actual/original bounding box is calculated. IOU value of an image and object are the maximum of computed values which is mentioned as the upper bound of the actual IoU. It will be evaluating the image content as the individual objects and their features [11].

L. Learned Perceptual Image Patch Similarity(LPIPS)

LPIPS is an evaluation metric that calculates the perceptual similarity between real images and generated images. Unlike traditional pixel-based metrics such as Mean Squared Error (MSE) or Peak Signal-to-Noise ratio (PSNR), LPIPS is correlate well with human evaluation perception.

LPIPS is used a pretrained network which extracts features from a large dataset of images that are relevant to human perception. This network comprises multiple convolutional layers and a global pooling layer that summarizes the image features into a fixed-length vector. To measure the image similarity of two images, and the respective feature vectors of the images are into the network and the distance between the vectors is measured using a distance metric like L1 or L2 distance [13].

Here is a simplified algorithm for computing LPIPS for GAN evaluation:

1. Sample real images from real distribution P, set of generated images from the generated distribution Q.
2. Use a pretrained network to extract features vectors of each image in the sets.
3. Compute distance between the feature vectors of each pair of generated and real images using a distance metric, such as L1 or L2 distance.
4. Take the average of the distances to get the LPIPS value.

M. Gradient Magnitude Similarity Deviation (GMSD)

GMSD is the evaluation metric and it is used to find the similarity by measuring the difference of gradient magnitude of the real and generated images. It needs the high-quality real images to compare with the generated images to find the quality. Gradient magnitude is calculated using the Root Mean Square (RMS) of image's directional gradient along with two orthogonal directions. Gradient measures using the classic Roberts, Sobel, Scharr and Prewitt filters, etc., Prewitt filter calculates gradient and it is simple among 3x3 template gradient filters. The horizontal (x) and vertical (y) directions are

$$h_x = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix} \quad h_y = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix}$$

Horizontal filter hx , vertical filter hy are applied to the reference and distorted images to obtain the corresponding horizontal and vertical gradient images of both images. At any given location i in the images, gradient magnitude of the reference image $m_r(i)$ and the distorted image $m_d(i)$ are calculated using the following formula:

$$m_r(i) = \sqrt{(r \otimes h_x)^2(i) + \left(r \otimes h_y\right)^2(i)}$$

$$m_d(i) = \sqrt{(d \otimes h_x)^2(i) + \left(d \otimes h_y\right)^2(i)}$$

"$\otimes$" is the convolution operation.

Gradient magnitude similarity (GMS) map is calculated as follows:

$$GMS(i) = \frac{2m_r(i)m_d(i) + c}{{m_r}^2(i) + \; {m_d}^2(i) + c}$$

c is a positive constant that supplies numerical stability.

Gradient Magnitude Similarity Mean (GMSM):

$$GMSM = \frac{1}{N}\sum_{i=1}^{N} GMS(i)$$

A higher GMSM score indicates better image quality. Here, N represents the total number of pixels in the image.

The Gradient Magnitude Similarity Deviation (GMSD):

$$GMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GMS(i) - GMSM)^2}$$

The score obtained from GMSD depends on the implementation and normalization method used, with scores ranging between 0 and 1. A higher score in GMSD indicates better image quality. [14].

IV. RESULT AND DISCUSSION

Evaluation metrics for Generative Adversarial Networks (GANs) provide a quantitative measure of the performance of the GAN model in generating realistic images. These metrics are important to evaluate the performance of the GAN model based on the generated images and optimize the hyperparameters of a model to enhance the result. Every metric has its own strength and weaknesses, and the choice of metric depends on the specific objectives of the evaluation. Our Research aims to provide an overview of the evaluation metrics and their experimental results for the basic GAN using the MNIST dataset.

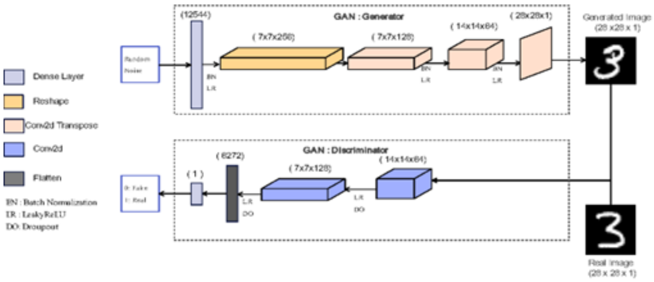

Figure 2. Basic GAN Architecture

The generator model accepts a random noise vector of size 100 as input and generates images of size 28x28x1.On the other hand, the discriminator model takes the generated image and a real image of size 28x28x1 as input and predicts whether it is real or fake. In our experiment, we utilized the InceptionV3 and VGG16 pre-trained models to extract features from both the generated and real images. To evaluate the performance of the GAN, we preprocessed the generated and real images and performed feature extraction. Based on our analysis, we conducted an experiment to evaluate the performance of the GAN across three different categories: Epoch 0, Epoch 50, and Epoch 100. We compared the scores obtained from each epoch as follows:

TABLE I. MODELSUMMARY OF BASIC GAN

GAN: Generator		**GAN: Discriminator**	
Layers	**Output Shape**	**Layers**	**Output Shape**
Dense Layer	(None, 12544)	Conv2D	(None, 14, 14, 64)
Batch normalization	(None, 12544)	LeakyReLU	(None, 14, 14, 64)
LeakyReLU	(None, 12544)	Dropout	(None, 14, 14, 64)
Reshape	(None, 7, 7, 256)	Conv2D	(None, 7, 7, 128)
Conv2d Transpose	(None, 7, 7, 128)	LeakyReLU	(None, 7, 7, 128)
Batch Normalization	(None, 7, 7, 128)	Dropout	(None, 7, 7, 128)
LeakyReLU	(None, 7, 7, 128)	Flatten	(None, 6272)
Conv2d Transpose	(None, 14, 14, 64)	Dense	(None, 1)
Batch Normalization	(None, 14, 14, 64)		
LeakyReLU	(None, 14, 14, 64)		
Conv2d Transpose	(None, 28, 28, 1)		

TABLE II. EXPERIMENTAL RESULT

Evaluation Metrics	**Range**	**Epoch 0**	**Epoch 50**	**Epoch 100**
FID	0 to infinity, lower is better	497.54	163.46	136.91
Mode Score	0 to 1, higher is better	0	1	1
Inception Score	>0, higher is better	1.1533	1.5388	1.6408
MMD	0 to infinity, lower is better	1147.54	863.64	823.85
PSNR	0 to infinity, higher is better	32.3508	33.7207	34.1515
Structural Similarity Index (SSIM)	-1 to 1, higher is better	0.3015	0.4806	0.4886

TABLE III. COMPARISON OF EVALUATION METRICS

Metric	**Type**	**Range of Result**	**Interpretation**	**Advantages**	**Limitations**
Inception Score (IS)	Generative	>0, higher is better	Measures quality and diversity of generated synthetic images based on image classification	Easy to compute, good for comparing models with similar image content	Ignores image quality, not good for comparing models with different image content
Mode Score (MS)	Generative	0 to 1, higher is better	Measures the quality of individual modes or clusters of generated images	Useful for detecting mode collapse, can identify poor image quality in specific modes	Computationally expensive, not good for models with continuous image distributions
Fréchet Inception Distance (FID)	Generative	0 to infinity, lower is better	Measures the similarity between feature representations of generated and real images	Good at capturing image quality and diversity, widely used in GAN evaluation	Requires the image feature extractor and large number of images, not good for small datasets
Maximum Mean Discrepancy (MMD)	Generative	0 to infinity, lower is better	Measures the difference between distributions of generated and real images in a feature space	Model-agnostic, useful for models with non-Gaussian image distributions	Computationally expensive, requires feature extractor
Peak Signal-to-Noise Ratio (PSNR)	Pixel-based	0 to infinity, higher is better	Measures the difference between pixel values of generated and real images	Widely used and easy to compute, good for comparing image quality	Not a good indicator of perceptual quality, not good for comparing models with different image content
Structural Similarity Index (SSIM)	Pixel-based	-1 to 1, higher is better	Measures the similarity between the structural information of real and generated images	Better at capturing perceptual image quality, good for comparing models with similar image content	Requires accurate image alignment, not good for comparing models with different image content
Semantic Object Accuracy (SOA)	Object-based	0 to 1, higher is better	Measures the accuracy of object detection in generated images	Good at capturing object-level quality and semantic content, useful for evaluating object detection models	Requires object annotations, not good for comparing overall image quality
Learned Perceptual Image Patch Similarity (LPIPS)	Perceptual	0 to 1, lower is better	Measures the perceptual similarity between generated and real images	Correlates well with human perception, good for evaluating image quality and diversity	Requires feature extractor, not good for comparing models with different image content

The evaluation metrics provide insights into the GAN model’s performance at different stages. The FID score measures the similarity between real and generated images, which improves over time as the score decreases. The Mode Score indicates the GAN’s ability to capture various image variations, progressing from 0 to 1 as the epochs increase. The Inception score reflects the image quality and diversity, showing improvement as it increases from 1.1533 to 1.6408. The MMD score measures the alignment between real and generated image distributions, decreasing as training progress. Both PSNR and SSIM scores increase, indicating enhanced image quality and structural similarity. Overall, these metrics demonstrate positive progress in the GAN model’s ability to generate realistic, diverse, and high-quality images as training advances.

All the metrics performed relatively well with the Basic GAN on the MNIST dataset, which is simple and straightforward. The Mode Score for the generated images from random noise starts at 0 and gradually increases to a maximum value of 1 in the later epochs. Mode Score is not highly effective

DOI: https://doi.org/10.17762/ijritcc.v11i8s.7248
***Article Received**: 28 April 2023 **Revised:** 18 June 2023 **Accepted:** 02 July 2023*

to evaluate the performance of our Basic GAN. However, for future work, it is challenging to evaluate the GAN model's performance on more complex datasets like MSCOCO. While evaluation metrics are essential, it is equally important to combine them with a visual inspection of the generated images and other subjective measures of image quality. By employing multiple evaluation methods, we can gain a comprehensive understanding of GAN model performance and effectively guide the development of new and enhanced models.

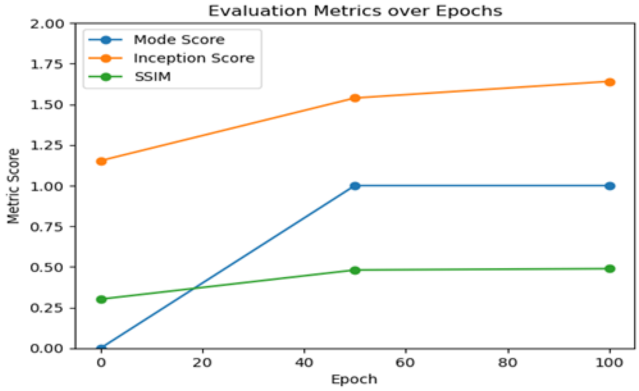

Figure 3. Evaluation Metrics(MS, IS,SSIM) : Higher score is better

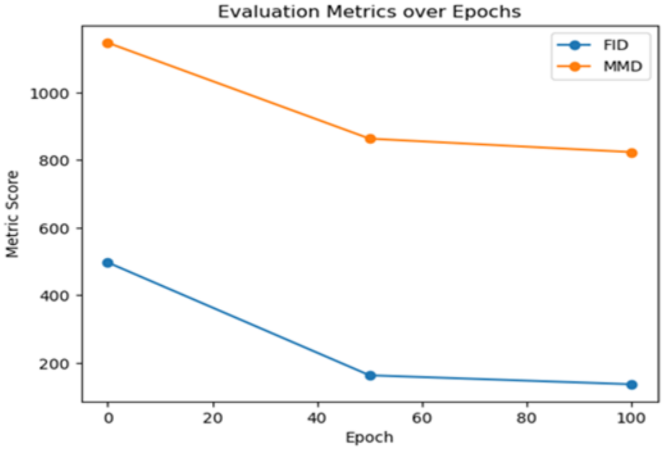

Figure 4. Evalution Metrics (FID, MMD) : Lower score is better

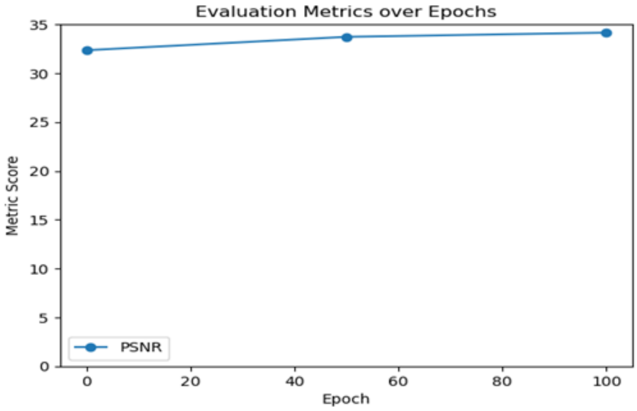

Figure 5. Evaluation Metrics (PSNR) : Higher Score is better

V. CONCLUSION

In this research article, we conducted a comparative experimental analysis of different evaluation metrics for text-to-image Generative Adversarial Networks (GANs). We evaluated the effectiveness of each metric in assessing the quality of generated images. Our findings suggest that there is no single metric that is universally effective, and different metrics perform differently depending on the dataset and GAN architecture. However, the metrics IS and FID consistently performed well in all experiments, indicating their reliability as indicators of image quality. Our study provides valuable insights into evaluating GANs for text-to-image generation, but there are still areas that require further research in the future. Future research can focus on exploring additional evaluation metrics, such as perceptual path length and Learned Perceptual Image Patch Similarity (LPIPS), with diverse datasets and GAN architectures. Additionally, combining multiple evaluation metrics could be investigated to achieve a more comprehensive assessment of image quality.

REFERENCES

[1] K. D. Kumar, S. Srang and D. Valy, "A Review of Generative Adversarial Networks (GANs) for Technology-Assisted Learning: Solving Teaching and Learning Challenges," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 820-826, doi: 10.1109/ICACRS55517.2022.10029021

[2] A simple explanation of the Inception Score | by David Mack | Octavian | Medium

[3] Yu, Yu & Zhang, Weibin & Deng, Yun. (2021). Frechet Inception Distance (FID) for Evaluating GANs.(PDF) Frechet Inception Distance (FID) for Evaluating GANs (researchgate.net).

[4] Tim Salimans, Ian Goodfellow, Wojciech Zaremba , Vicki Cheung , Alec Radford , Xi Chen , “Improved Techniques for Training GANs”, [1606.03498] Improved Techniques for Training GANs (arxiv.org), 2016.

[5] How to Implement the Inception Score (IS) for Evaluating GANs - MachineLearningMastery.com

[6] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Gu, Yu Sun, Felix Wu, Kilian Q. Weinberger , “An empirical study on evaluation metrics of generative adversarial networks”, 2018, 1806.07755.pdf (arxiv.org).

[7] Weizhi Du , Shihao Tian , “Transformer and GAN Based Super-Resolution Reconstruction Network for Medical Images”, [2212.13068] Transformer and GAN Based Super-Resolution Reconstruction Network for Medical Images (arxiv.org), 2022.

[8] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh and Eero P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity ”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 13, NO. 4, APRIL 2004.

[9] All about Structural Similarity Index (SSIM): Theory + Code in PyTorch | by Pranjal Datta | SRM MIC | Medium.

[10] Tobias Hinz, Stefan Heinrich, and Stefan Wermter,"Semantic object accuracy for generative text-to-image synthesis". arXiv preprint arXiv:1910.13321, 2019.

[11] Stanislav Frolova,b,∗ , Tobias Hinzc , Federico Raueb , J¨orn Heesb , Andreas Dengela,b, "Adversarial Text-to-Image Synthesis: A Review", Adversarial Text-to-Image Synthesis: A Review (arxiv.org), 2021.

[12] Zhang L, Zhang L, Mou X, Zhang D. FSIM: a feature similarity index for image quality assessment. IEEE Trans Image Process. 2011 Aug;20(8):2378-86. doi: 10.1109/TIP.2011.2109730. Epub 2011 Jan 31. PMID: 21292594.

[13] Chaudhary, D. S. ., & Sivakumar, D. S. A. . (2022). Detection Of Postpartum Hemorrhaged Using Fuzzy Deep Learning Architecture . Research Journal of Computer Systems and Engineering, 3(1), 29–34. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/38

[14] Leihao Wei, Yannan Lin, William Hsu, "USING A GENERATIVE ADVERSARIAL NETWORK FOR CT NORMALIZATION AND ITS IMPACT ON RADIOMIC FEATURES", 2001.08741.pdf (arxiv.org) 2020.

[15] W. Xue, L. Zhang, X. Mou and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," in IEEE Transactions on Image Processing, vol. 23, no. 2, pp. 684-695, Feb. 2014, doi: 10.1109/TIP.2013.2293423.