Vol. 06, No. 4 (2024) 1829-1836, doi: 10.24874/PES.SI.25.03b.015

Proceedings on Engineering Sciences

www.pesjournal.net

TEXT DESCRIPTION TO IMAGE GENERATION USING GENERATIVE ADVERSARIAL NETWORK

Kayal Padmanandam[1]
Yeshasvi Mogula
Nikitha Pitla

Received 16.11.2023.
Received in revised form 25.12.2023.
Accepted 20.01.2024.
UDC – 004.9

Keywords:

Generative Adversarial Network (GAN), Attention Generative Adversarial Network, Generator, Discriminator, Style based Generator, Text to image.

A B S T R A C T

*The progression of translating text into images has been an imperative topic of research. The significant challenges arise from translating visual to textual information and vice versa. High-quality images can be generated from text using a Generative Adversarial Network (GAN), however, there are challenges associated with accurately portraying the content of the sentence provided to the model. Text-to-image conversion strategies can produce examples that closely reflect the descriptions' intended meaning. The user descriptions may however lack crucial details. To create the conditioned text descriptions, this study employs an Attention-Generative Adversarial Network to generate 256*256-pixel images that are image-sensitive. In the initial phase of GAN sketches, the input text descriptions serve solely to inform the basic form and color scheme of the devices. The information gleaned from the first stage, along with the textual descriptions, is fed into a GAN which generates images with high resolution and realistic detail. The conditional GAN training may be stabilized using conditioning augmentation, and the generated samples can have higher quality. Using Style based Generator, samples for each style of the image can be drawn. The proposed system can generate photorealistic visuals of an object when the user inputs the textual descriptions in the application's GUI.*

1. INTRODUCTION

Image Caption generator is one of the most prevalent topics in Natural Language processing and computer Vision. Similarly, generating images from natural language as text is the vice versa process. These are research problems in a broad sense, identical to those that arise during linguistic translation. Like different languages can convey the same meaning in different ways, images and text can encode similar information. Because of the multimodal characteristics issues, text-image and image-text conversions are not the same. The text description stating, "This is a magnificent red flower", has few possible translations into English. A broad variety of mental images could correspond to this description if one tried to create one. This multimodal behaviour is not exclusive to image captioning challenges, nonetheless, the sequential nature of language simplifies the difficulty there. The text-to-image structure is more tedious than image captioning.

[1] Corresponding author: Kayal Padmanandam
Email: kayalpaddu@gmail.com

Potential uses for image synthesis based on natural language are vast, once the technology is ready for widespread use. By using the existing textual descriptions **(**words) as a basis for generating new ones, the proposed system can be put to good use. Instead of spending a lot of time looking for relevant images, users could tell a computer what they want and the proposed system will generate it. Creators and machines could work more closely together to make content using natural language (Raja et al., 2022).

Creating high-quality images from a word-based description prevails to be a complex research obstacle. The resulting synthetic image must be accurate to the given description with high quality. If the process is automated, it can be used in the fashion industry to prototype new designs which is invaluable to designers. In addition, the ability to generate various designs of apparel items based on human-written descriptions would facilitate the discovery and intuitive selection of one's favourite designs. Since the generated images are crisper than those of the previous deep generative representation, GAN has gained a lot of attention recently. It has been shown that GANs can be trained with additional knowledge to guide the data generation process like Class labels, attributes, images, texts and other condition variables. Researching text-to-image structured algorithms, predominantly GAN-based approach implementing cutting-edge methods, allows researchers to compare the effectiveness of various methods and gain insights into this field. Also, Photo editing, CAD, and other fields could benefit greatly from this system. Creating synthetic versions of real-world images using GANs has shown encouraging results. Figure 1 shows the sample generation text to an image.

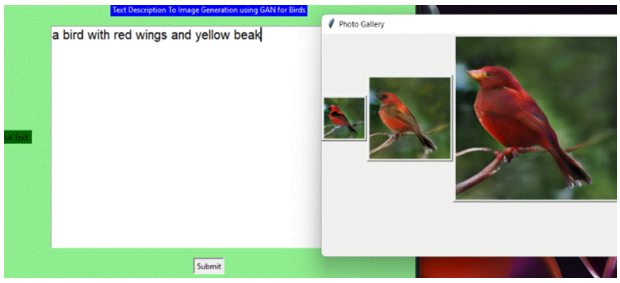

Figure 1. Sample image of text description to image generation

There's a discussion of related works in Section II, a description of the proposed Text Description to Image Generation system in Section III, and an explanation on methodology and experimental results are discussed in Sections IV &V respectively.

2. RELATED WORKS

High-resolution image generation from text descriptors is essential for several real-world applications, including art creation and computer-aided architecture. Significantly, the advent of deep generative models has led to progress in this area (Goodfellow et al., 2014), (Kingma et al., 2014) and (van den Oord et al., 2016) (Mansimov et al., 2016) extending AlignDRAW led to the discovery of a model Deep Recurrent Attention Writer (DRAW) (Gregor et al.,, 2015) to sketch texture features iteratively while paying close attention to the pertinent words in the caption. The Paper (Nguyen et al., 2017) proposed an estimated Lagrangian approach for generating images from annotations. (Reed et al., 2017) have been using conditional Pixel CNN (van den Oord et al, 2016) to summarise the images from text using a multi-scale model structure. In comparison to various deep generative models, generative adversarial networks (GANs) (Ian Goodfellow et al., 2014), have demonstrated outstanding performance to generate cleanser samples. (Denton et al., 2015) (Isola et al., 2017) (Ledig et al., 2017), (Szegedy et al., 2016), (Larabi, 2007), (Radford et al., 2016), **(**Onyema et al., 2023) and (Salimans et al., 2016). Their subsequent work illustrated that GAN can be capable of producing better sample data by integrating multiple conditions (e.g., objects and locations). (Zhang et al., 2017a,b) used distinctive GANs to generate pictures of varying dimensions after stacking numerous GANs for textual content-to-photo structure. Nevertheless, all of their GANs are constrained to the global sentence vector, so image generation is missing best-grained phrase-degree statistics. The attention mechanism has also been incorporated into models of sequence transmission. It has been used effectively in responding to questions (Xu et al., 2015) about image subtitling (Yang et al, 2016) and machine translation to model multilevel dependencies. (Vaswani et al., 2017) have also shown that, with the aid of an attention model, machine translation models produce state-of-the-art results. Despite these advances, the attention mechanism in GANs for textual content-to-picture structure has until now to be investigated. It should be noted that AlignDRAW (Mansimov et al., 2016) and (Russakovsky et al., 2015) pre-owned LAPGAN (Denton et al., 2015) to expand the copy of the photograph to a superior resolution. As a result, GANs were not given much consideration in their framework (Gal et al., 2022). In our opinion, the proposed AttnGAN is the GAN to employ multi-level (e.g., word and sentence level) conditioning to generate fine-grained greater images.

3. PROPOSED SYSTEM

A high-resolution, detailed image is obtained with the proposed method. There is also a Deep Attentional Multimodal Similarity Model (DAMSM) included with the AttnGAN. Based on global sentence-level data and fine-grained word-level information, the DAMSM calculates the similarity between the generated source image and the statement. To generate a training dataset

for the generator, DAMSM applies a fine-grained loss for image capture matching. The technique includes three obligations, i) An Attentional Generative Adversarial Network projected for structure and quality image for the text descriptions. There are two attentional generative networks: the DAMSM and the Attentional Generating Network which are two dissimilar elements introduced in the Attention GAN field. (ii) a complete and accurate study is conducted to analytically assess the projected ATTNGAN. The experiments demonstrate that AttnGAN outperforms original condition GAN's models significantly. (iii) A comprehensive scrutiny is carried out by depicting the AttnGAN's attention layers. It is demonstrated for the first time that the patterned conditional GAN can instantaneously attend to the appropriate terms from image generation.

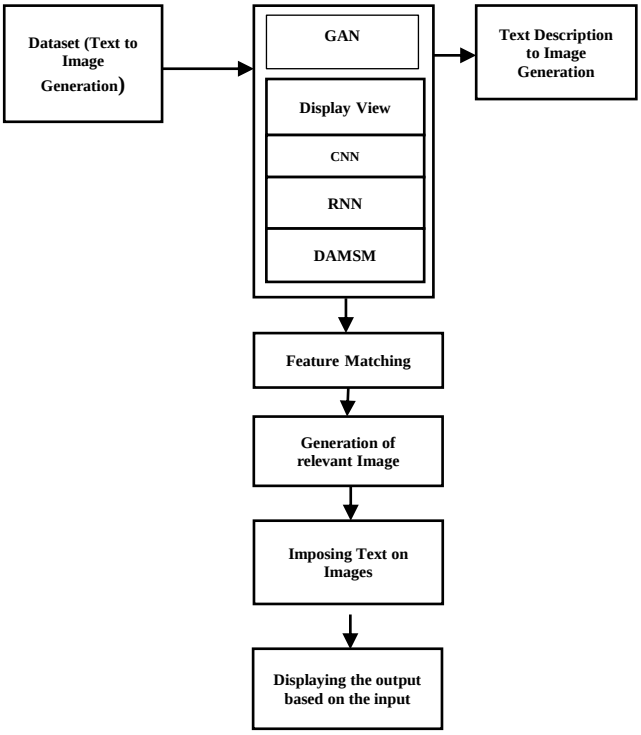

Figure 2. Progressive Diagram of GAN

Figure 2 shows the working of GAN with the given dataset. The algorithms used in the Text Description to Image Generation are AttnGAN (Attentional Generative Adversarial Network), and DAMSM (Deep attentional multimodal similarity model). The system first detects the Attn GAN from the context and considers the mapping between keywords and images. Then the trained output of DAMSM is sent to the discriminator to find the real and fake images. Style-based generator is used to reshape and resize the image for the clear visibility of the images.

3.1 System Architecture

Using Global Attention, we can design and build text to image generation system that automatically extracts key phrases and words from a written description. A low-resolution picture is generated from phrase attributes and the random noise vector is used as the backdrop environment in a generative neural network. Figure 3 is to resolve the issue of false positive caused by erroneously authenticating fake photos, a discriminative network is employed. The generators and discriminators for a high-resolution picture are arranged like a tree.

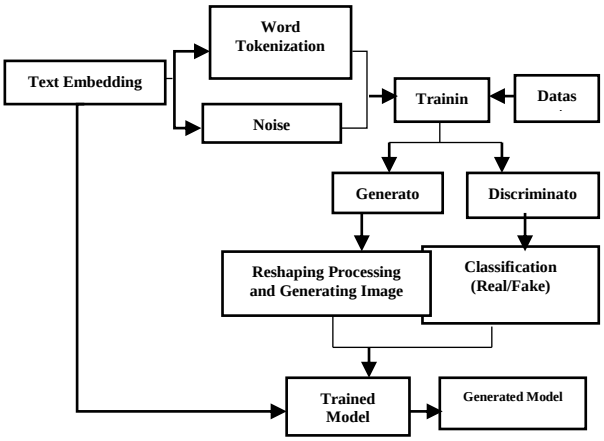

Figure 3. System architecture

3.2 Working model of Attention Generative Adversarial Network

Existing GAN-based text-to-image generation models (Zhang et al, 2017a,b) encode the entire sentence text description into a single word vector. This approach, however, creates an innovative model of attention that generates different graphics from the subregions according to words. The used attentional generative network has m Generators which will take hidden states of the generator and generate the smaller extent of images to the larger scale of images. As shown in Equation 1, the sampled standard normal distribution is represented by the noise vector Z. The vector ê is considered to be the sentence vector, while e is considered as the matrix of the sentence vector. At the i^{th} stage of the attention GAN as a neural network, f^{ca} represents the acclimatizing (Conditional) Augmentation for the sentence vector, from which f^{attn} forms the attention model

$$\boldsymbol{h_0} = \boldsymbol{f_0}\left(\boldsymbol{z}, \boldsymbol{f^{ca}}(\bar{\boldsymbol{e}})\right):$$
$$\boldsymbol{h_i} = \boldsymbol{f_i}\left(\boldsymbol{h_i - 1}, \boldsymbol{f_i^{attn}}(\boldsymbol{e}, \boldsymbol{h_i - 1})\right) \quad (1)$$

$$\hat{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{G_i}(\boldsymbol{h_i})$$

The two features word and image features from the preceding hidden layer are taken as inputs for the following single-layer perceptron. The word features are initially transformed into the image features shared semantic space by creating a new feed-forward neural layer, e′ = Ue, where U € R. To find the right word for each j^{th} sub-region of the picture, a word-context vector is created using hidden features. Each column of h indicates a feature representation of an image sub-

region. The word context vector for the jth sub-region is a dynamic description of word vectors pertinent to h_j derived by s´ j, i = hTj e´i, where j,i is the weight given to the i^{th} word by the model while creating the j^{th} sub-region of the picture. Finally, the image feature and the equivalent word-context features are combined to generate images at the later level of the discriminator. To generate realistic images, one can use the generative attentional network, which has as its objective function the generation of images at various levels (Xu et al, 2017).

$$L = L_G + L_{DAMSM,\ Where\ L_G} = \sum_{i=0}^{z-1} L_{G_i} \qquad (2)$$

In this ƛ defined as the hyperparameter used to equilibrium the two GAN losses that estimated the conditional and unconditional distributions jointly. The generator Gi has an identical discriminator I, the i^{th} phase of AttnGAN. Figure 4 explains the flow diagram of GAN. The adversarial forfeiture of G_i decides whether the photos are real or fake, whereas the Conditional loss distinguishes between images and phrases that match and those that do not. In addition to being trained on the generative model, each discriminator is taught to classify inputs as real or false. ATTN: GAN discriminators are fundamentally discontinuous, allowing them to be trained to focus on one picture only.

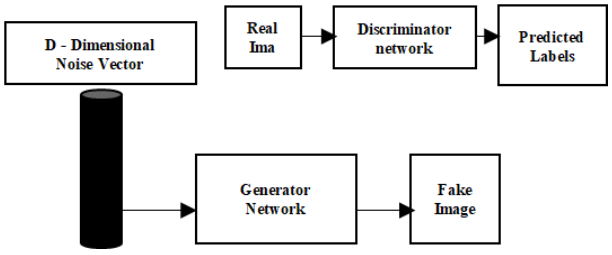

Figure 4. Flow Diagram of ATTN GAN

3.3 Working process of Deep Attentional Multimodal Similarity Model (DAMSM) for visualizing and comparing the similarity of multiple object categories

The DAMSM get trained to deuce computational models that maps imageries segments & sentence phrases to a common latent space, assessing the researchers sought to estimate a fine-grained loss for image generation by comparing the similarity of images with their text counterparts, only at the word level. Image-Text similarity is a matching algorithm used to measure the correspondence between the image descriptions and the corresponding text segments. The results can be used to calculate a comparison matrix for all possible pairings of arguments in the sentences and images. Finally, the cosine comparison between the corresponding words c_i and e_i is used to determine whether or not the i^{th} word is related to the image. The attention-consumed per-character, per-word correlation between the full image (Q) and the entire text description (D) is defined as the function of the lowest error classification.

3.4 Working of Style based Generator for the images

In a feedforward network, the latent code is typically sent from the input layer to the generator. It deviates from this format by substituting a learned constant for the input stage in constructing the model's initial conditions. A non-linear mapping network (f: Z W) will return "W" as a starting point if a latent code z is given as input from the input latent space Z. For consistency and ease of use, to the proposed system standardize the measures. Instead of just passing the latent code through the input layer as is done in a traditional generator, it instead transfer to a latent space W that acts as a controller for the generator through adaptive instance normalization (AdaIN) at each convolution layer. After each convolution, the outcomes of the nonlinear analysis and the addition of Gaussian noise are shown. A learned affine transform and scaling parameters specific to the B channel are then applied to the noise input. There is a total of eight layers in the mapping network f (two for each resolution) and eighteen in the synthesis network g. A separate 1 x 1 convolution technique is employed to transform the final layer's output to RGB. Compared to the standard in the industry of 23.1M trainable parameters, the proposed generator has 26.2M. Based on the findings of the research conducted, it was determined that a Multi-Layer Perceptron (MLP) with 8 layers would be the optimal architecture for realizing the 512-dimensional mapping f. Affine transformations learned at the end of each convolution layer in the synthesis network regulate the AdaIN procedure, which specializes w to styles y = (ys, yb).

4. METHODOLOGY

Initially, the model was presented as a system of trained modules, each of which had a brief description of its purpose and function. as the proposed system has a standardized web application to let users' query rationalization go as smoothly as possible. An appealing user interface has been developed so that users may quickly grasp the information presented and quickly apply it to their personalized imagination. The web application design has a user-friendly GUI allowing for simple navigation and users' text-based image generation.

4.1. Text Feature Extraction

Semantic vectors, which represent the meaning of words, are extracted from the text description using the bidirectional LSTM, allowing for the extraction of text features. Here we have two hidden states for each word, one for each conceivable orientation generated using a two-way LSTM. Therefore, we

represent the importance of a word by combining its two hidden states. The feature vector for each word is stored in a separate column of the feature matrix, and the final hidden states of the bidirectional LSTM are then put together to generate the sentence vector. Figure 5 shows how the features of the given input are extracted.

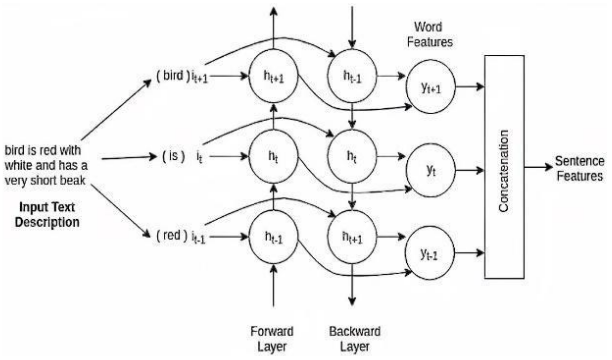

Figure 5. Architecture of text feature extraction

4.2. Priming- (Conditional Augmentation)

When it comes to text embedding, the latent space is typically quite high dimensional even above a hundred. When learning the generator with a small amount of data, it is not ideal to have gaps in the latent data manifold. Figure 6 is about the condition augmentation. The latent variables are drawn at random from a separate Gaussian distribution.

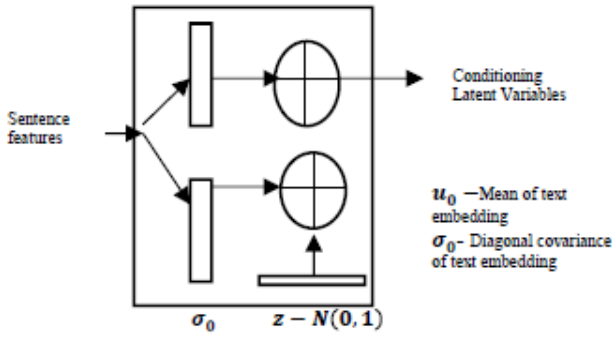

Figure 6. Architecture of Conditional Augmentation

4.3. Making Pictures - (Generative Network)

The picture creation module has a tree-like structure with numerous generators to produce images at varying scales, and it receives a conditional variable and noise vector generated by the pre-processing phase as input. Figure 7 shows the visuals produced at the first branch level are very simple in terms of colour and structure. When this is done, the generators at the subsequent branches can concentrate on filling in the remaining details in order to produce images with a better resolution.

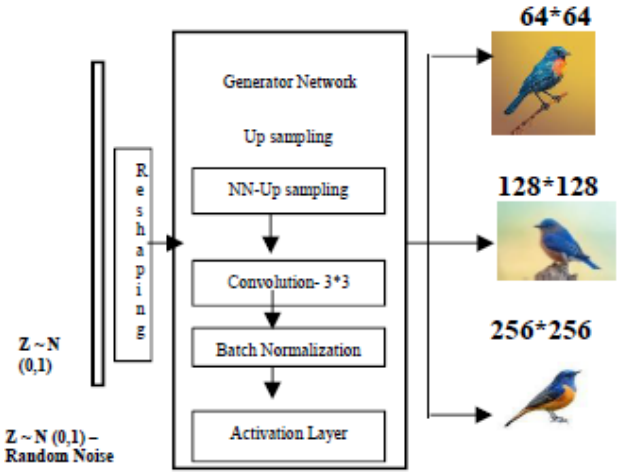

Figure 7. Architecture of image generation

4.4. A genuine method of identifying images- (Discriminative Network)

Figure 8 the image produced by the generative network is reduced in dimension by the discriminator, which comprises down-sampling, convolution, residual blocks, and the sentence characteristics are appended. A comparison is then made between the original image in the dataset and the created image. The discriminator determines the generator and discriminator losses based on the output.

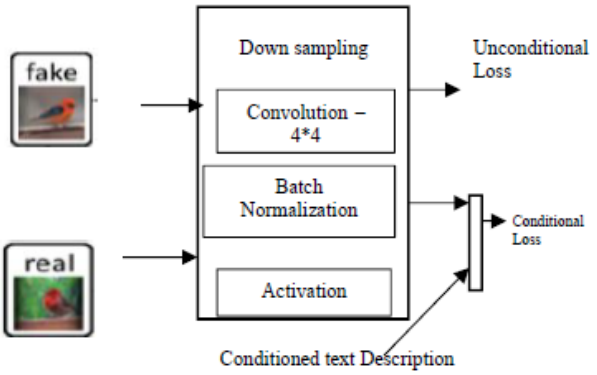

Figure 8. Architecture of image identification

This causes the generator's weights to change. There is a concurrent training process for both the generator and the discriminator.

5. EXPERIMENTAL RESULTS

To assess the generative model's performance, two well-known measures, the Inception Score (IS) and the Frechet Inception Distance (FID), were employed. The IS uses a pre-trained Inception Network model to recognize patterns in sensory data to determine class probabilities for simulated data. Higher IS scores suggest more objective and diverse material. Our IS calculations are carried out using the same Inception

Networks as were used in our prior in-depth comparisons. There are networks available on StackGAN that are tailored to the CUB and Oxford-102 datasets. The IS is flawed since it does not take statistical information. An excellent IS score would be achieved by a generative model with minimal diversity that generates just a few high-quality examples for each class. To circumvent this issue, the Frechet Inception Distance was devised (FID). Using FID, one may determine if a generative model produced a distribution with statistical properties that are comparable to those of the training data. To be more specific, an Inception Network is used by FID to calculate activation characteristics of both training set photos and final images. Table 1 (Mogula et al., 2023) shows the measures of the images and scores of the input text. The Frechet Distance is then determined by comparing the genuine and fake image characteristics. When the FID score is low, the produced pictures statistically reflect the training set as nearly as possible.

Table 1. Metrics for the image measurement

Metric	The Image Quality	The Image Drivers	Object Fidelity	The Mentioned Objects	The Numerical Aligment	The Posiitonal Aligment	Paraphrase Robustness	Text Relevance	Automatic
IS	✓		✓						✓
FID	✓	✓							✓
Score FID			✓						✓
R-Prec					✓			✓	✓
SOA				✓				✓	✓
Caption Img								✓	✓

A. Evaluation Measures

i. Inception Score (IS):

The entropy of the distribution of classes over the sampled data is a measure of diversity; a high value indicates a lack of dominant classes and a more evenly distributed training set. Here is the scoring equation: where the probability distribution is a KL-divergence. Given the label and the generated image, the conditional probability is denoted by p(y/x), with p(y) standing for the marginal probability. Information loss due to the approximation of an empirical distribution is quantified by the KL-distance.

ii. Fractional Frechet-Inception Distance (FID):

It has been reported that scores for FID generated by the TensorFlow and Pytroch implementations are different, even though there are quantitative measures of quality. While style-attn GAN results may appear to be on the lower end compared to SOTA, a closer look at the generated images reveals otherwise. After looking at a few hundred examples, it is observed that the style-AttnGAN-generated photos are more consistent and appear more photorealistic. Table 2 shows the dataset performances comparison with different scores.

Table 2. Performance Comparison

Data set	**Frechet Inception distance**	**Inception score**	**Human Rank**
CUB-200	16.89	4.08 ±.04	1.29±.03
OXFORD-102	48.88	3.27±.05	1.30±.03

iii. R-Precision:

To calculate the R-precision, R-precision ranks potential text explanations for each image based on their process of similarity and relies on selecting the most relevant descriptions. The R^{th} place, accuracy equals recall. Table 3, the Best inception score is an indicator of how well a particular GAN model performs on the CUB dataset. R-precision rate is a measure of accuracy that reflects how closely each generated image matches the original photo (Xu et al, 2017).

Table 3. Calculated Scores of Inception Score and R-precision

Method	**Inception Score**	**R-Precision Score**
AttnGAN2, no StyleAttnGAN	3.99±.05	10.38 ±5.8
AttnGAN2, StyleAttn GAN=0.1s	4.20±.07	16.58 ±4.85
AttnGAN2, StyleAttn GAN=1	4.38±.08	34.98± 4.03
AttnGAN2,StyleAttn GAN=5	4.39±.05	58.69 ±5.42
AttnGAN2, StyleAttn GAN=10	4.30±.08	63.88± 4.86

6. CONCLUSION

In this paper, we present an updated version of the neural network with information flow, called "Captioner," which is an integral part of the GAN architecture. The proposed network architecture may be thought of as a chain: text > picture > text, with the central idea being to use the restored original text as input to the network. We base our comparison of

the existing StackGAN framework to four different variant implementations. Based on this study, it is observed that the Captioner module can generate images with more authentic expressions and preservations of primitive elements. To further enhance the image data quality, the embeddings are used in error calculation. Moreover, both synthesized image data and real picture data are useful for bolstering quality. We anticipate that integrating these diverse implementations will certainly elevate to cutting-edge GAN designs. It is a novel approach to create generations that approximate real-world situations with GANs that provide trustworthy outcomes.

References:

Denton, E. L., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. doi:10.48550/arXiv.1506.05751

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. doi:10.48550/arXiv.2208.01618

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. doi:10.48550/arXiv.1406.2661

Gregor, K., Danihelka, I., Graves, A., Jimenez Rezende, D., & Wierstra, D. (2015). A recurrent neural network for image generation. In *International Conference on Machine Learning*. doi:10.48550/arXiv.1502.04623

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition Conference (CVPR)*. doi:10.1109/CVPR.2017.632

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations*. doi:10.48550/arXiv.1312.6114

Larabi, S. (2007). Textual description of images. *International Journal of Simulation Modelling*, *6*(2), 93–101. doi: 10.2507/IJSIMM06(2)S.04

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Computer Vision and Pattern Recognition Conference (CVPR)*. doi:10.1109/CVPR.2017.19

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). Generating images from captions with attention. In *International Conference on Learning Representations*. doi:10.48550/arXiv.1511.02793

Mogula, Y., Kayal, P., & Subetha, T. (2023). A survey on text description to image generation using GAN. In G. Ranganathan, X. Fernando, & S. Piramuthu (Eds.), *Soft computing for security applications* (Vol. 1428, pp. 52). Springer, Singapore. doi:10.1007/978-981-19-3590-9_52

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Computer Vision and Pattern Recognition Conference (CVPR)*. doi:10.48550/arXiv.1612.00005

Onyema, E. M., Balasubaramanian, S., Suguna, K. S., Iwendi, C., Prasad, B. V. V. S., & Edeh, C. D. (2023). Remote monitoring system using slow-fast deep convolution neural network model for identifying anti-social activities in surveillance applications. *Measurement: Sensors, 27*, 100718. doi:10.1016/j.measen.2023.100718

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*. doi:10.48550/arXiv.1511.06434

Raja, S. A., Sundarvadivazhagan, B., Vijayarangan, R., & Veeramani, S. (2022). Malicious webpage classification based on web content features using machine learning and deep learning. In *International Conference on Green Energy, Computing and Sustainable Technology (GECOST)* (pp. 314-319). doi:10.1109/GECOST55694.2022.10010386

Reed, S., Oord, A., Kalchbrenner, N., Gómez Colmenarejo, S., Wang, Z., Chen, Y., Belov, D., & Freitas, N. (2017). Parallel multiscale autoregressive density estimation. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 70, 2912-2921). https://proceedings.mlr.press/v70/reed17a.html

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115*(3), 211–252. doi:10.48550/arXiv.1409.0575

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In *NIPS for computer vision*. doi:10.48550/arXiv.1606.03498

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture. doi:10.1109/CVPR.2016.308

van den Oord, A., Kalchbrenner, N., Espeholt, L., Kavukcuoglu, K., Vinyals, O., & Graves, A. (2016). Conditional image generation with pixel CNN decoders. doi:10.48550/arXiv.1606.05328

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. doi:10.48550/arXiv.1706.03762

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. doi:10.48550/arXiv.1502.03044

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2017). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. doi:10.48550/arXiv.1711.10485

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Computer Vision and Pattern Recognition Conference (CVPR)*. doi:10.48550/arXiv.1511.02274

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017a). Text to photo-realistic image synthesis with stacked generative adversarial networks. In *International Conference on Computer Vision*. doi:10.1109/ICCV.2017.629

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. (2017b). Realistic image synthesis with stacked generative adversarial networks. doi:10.48550/arXiv.1710.10916

Kayal Padmanandam
BVRIT HYDERABAD College of Engineering for Women,
Hyderabad,
India
kayalpaddu@gmail.com
ORCID 0000-0002-3872-8422

Yeshasvi Mogula
BVRIT HYDERABAD College of Engineering for Women,
Hyderabad,
India
yeshasvireddy26@gmail.com
ORCID 0009-0007-2233-2320

Nikitha Pitla
BVRIT HYDERABAD College of Engineering for Women,
Hyderabad,
India
nikithap121@gmail.com
ORCID 0009-0000-7997-2547