Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241099

A deep dive into generative modeling: Evaluating DF-GANs, DM-GANs, and AttnGAN

Haojing Tong

Department of Computer Science, The University of Manchester, M13 9PR, England

haojing.tong@student.manchester.ac.uk

Abstract. Generative Adversarial Networks (GANs) have become pivotal for generating synthetic data. This paper conducts a comprehensive comparison of three cutting-edge GAN models. In particular, this study delved deep into the architectural intricacies, strengths, and limitations of each model, emphasizing their distinct features and mechanisms. DF-GANs focus on producing natural images with a single-stage backbone, DM-GANs leverage memory structures to enhance model performance, while AttnGAN employs attention-driven, multi-stage refinement for precise text-to-image generation. Through a series of literature search, this study evaluates the applicability of these models in various scenarios, offering insights into their practical implications and potential areas of improvement. This comparative study aims to serve as a reference point for researchers and practitioners alike, shedding light on the contemporary advancements in GAN technology and guiding future developments in the domain.

Keywords: GAN, AttnGAN, DM-GAN.

1. Introduction

Generative Adversarial Networks (GANs) have become a significant part of the deep learning community, particularly in the generation of images [1]. GANs are made up of two networks: the first one is the generator which creates samples, and the other one is the discriminator which tries to distinguish between real and generated samples. The adversarial competition between two networks mentioned above leads to the generation of highly realistic data. This is the foundational works of GANs.

With the development of GANs in nearly one decade, it has gained immense popularity in the deep learning community. There are several reasons for this: High-Quality Generations [2], the output of GANs model is of high quality, which has high value in market. This capability was unprecedented before GANs reached maturity. At the same time, GANs are versatility. Not only it could be used in text-to-image synthesis which would be researched by this paper, also, it could be used in the fields of data augmentation, style transfer, and even generating music. Applications in various domains, GANs have many applications in diverse areas. For instance, face generation and art creation in the field of computer vision, voice generation in audio processing, generating medical scan images in medical imaging and creating realistic game environments in gaming field.

For basic GAN model, it has many drawbacks. For instance, the existing text-to-images models stack multiple generators acquiring high-resolution outputs [3]. A paragraph of text should be withdrawn as several image features, each feature matched to one generator. Each generator would equip with one output, which would satisfy the feature. After all the generators completed, the model would produce a final output. The problem of this final output is the image generated is regarded as the combination of each feature mentioned in the text. In a single image, it looks like a lot of segmentation. Resulting in the image is not a great generation with high market value.

In this case, many new GAN models make an improvement based on the basic GAN model according to the baseline. In this paper, three main models would be researched in the direction of text to image fields, DF-GANs model [3], DM-GANs model, and AttnGAN, respectively. DF-GAN [3] is a novel text-to-image synthesis model that addresses limitations in existing Generative Adversarial Networks (GANs). Unlike traditional methods which use a stacked architecture that might introduce entanglements between different image scales, DF-GAN uses a one-stage backbone to directly synthesize high-resolution images. To enhance text-image semantic consistency, it introduces a Target-Aware Discriminator with features like a simplified One- Way Output. Moreover, the model deepens text-image fusion using a newly proposed Deep text-image Fusion Block (DFBlock) to thoroughly integrate text and visual features. Experiments indicate DF- GAN's superior performance over other modern models for producing realistic visuals matching textual descriptions.

The DM-GAN is a newly introduced Generative Adversarial Network designed for text-to-image synthesis. It integrates a dynamic memory module to address shortcomings in initially generated images, ensuring a more refined output. Through these enhancements, DM-GAN consistently produces higher-quality images from textual descriptions compared to existing methods.

The Attentional Generative Adversarial Network (AttnGAN) model is designed for fine-grained text-to-image synthesis, utilizing attention-driven, multi-stage refinement. Unlike traditional meth- ods that condition image generation on a global sentence vector, AttnGAN employs both global and word-level vectors. This enables the generator to focus on specific words, drawing relevant parts of the image in successive refinements. In addition, the model introduces a Deep Attentional Multimodal Similarity Model (DAMSM) that computes image-text similarity at both the global sentence and de- tailed word levels. This results in a more precise image-text matching loss for training. Empirical evaluations suggest that AttnGAN surpasses previous models in performance, and visual analyses further confirm its ability to automatically attend to pertinent words during image creation.

In the main body of paper, this study would introduce the three models briefly in architecture, mathematical theory and workflow. And the discussion about the applications, limitations and future development of each model.

2. Method

2.1. Introduction of DF-GANs

DF-GAN shown in Figure 1 is a novel text-to-image synthesis model using a one-stage backbone, enhances text-image semantic consistency with a Target-Aware Discriminator, and deepens feature fusion with a Deep text-image Fusion Block, outperforming existing methods.

The main idea of DF-GANs can be summarized as follows: 1) Changing the stacked architecture [4, 5] which is the common models in recent research as the backbone to generate the high-resolution images to the Deep Fusion models [3] which generates high resolution images. This architecture replaces the original multiple-stage backbone as one-stage back- bone [3]. 2) The progress of traditional studies fixing extra networks [6,7] during the adversarial training [3]. These studies need to be improved because it is easily cheated by the generator in the process of synthesizing adversarial networks. Thus, the DF-GANs improved the text- to-image semantic consistency. 3) The cross-modal attention [8] has inability to fully utilize the whole paragraph of text information due to high-cost computational cost. In DFGANs model [3], the DFBlock [3] includes several Affine Transformations [9]. An affine transformation is a specific type of linear transformation that can change the size, shape, position, and orientation of an object.

The principle of DF-GANs can be summarized as follows:

(1) The formula [3] of one-stage method with hinge loss [10] is shown as follows:

$$L_{D} = -E_{x\sim Pr}[min(0, -1 + D(x, e))] \tag{1}$$

$$-(1/2)\ E_{G(z)\sim Pg}[min(0, -1 - D(G(z), e))]$$

$$-(1/2)\ E_{x\sim Pmis}[min(0, -1 + D(x, e))]$$

$$L_G = E_{G(z)\sim Pg}\ [D(G(z),e)]$$

For the formula above, the parameter z is the noise vector; the parameter e is the sentence vector; the parameter ¶r, ¶g, ¶mis represented real data distribution, synthetic data distribution, and mismatching data distribution respectively.

(2) The formula [3] of one-stage method with MA-GP is shown as follows:

$$-LD = -\ Ex\sim Pr[min(0, -1 + D(x, e))] \tag{2}$$

$$-(1/2)\ EG(z)\sim Pg[min(0, -1 - D(G(z),e))]$$

$$-(1/2)\ Ex\sim Pmis[min(0, -1 + D(x, e))]$$

$$-+\ kEx\sim Pr[(||\Delta xD(x,e)|| + ||\Delta eD(x,e)||)p]$$

$$-$$

$$-LG = -\ EG(z)\sim Pg[\ D(G(z),e)]$$

For the formula above, two hyper-parameters k and p are used for balancing the effectiveness of gradient penalty.

(3) The formula [3] of with efficient text-image fusion is shown as follows:

$$Y = MLP_1(e),\ \theta = MLP_2(e) \tag{3}$$

For the formula above, the first which is the language-conditioned channel-wise scaling parameter Y is predicted by MLP-one, the second which is the shifting parameters θ is predicted by MLP-two, the is the sentence vector mentioned before.

$$-AFF\ (xi|e) = \upsilon ixi + \theta i \tag{4}$$

For the formula above, which represented the Affine Transformation. Where the x_i is the i_{th} channel of visual feature maps.

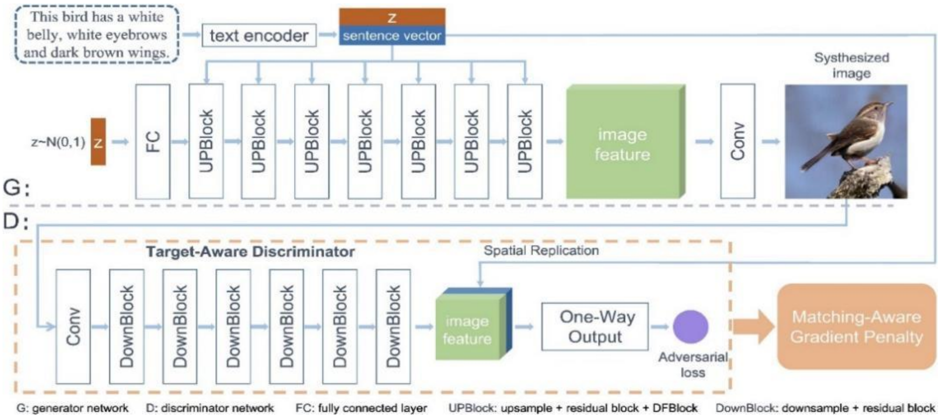

Figure 1. The workflow of DF-GANs [3].

2.2. Introduction of DM-GANs

The DM-GAN shown in Figure 2 is a novel Generative Adversarial Network that incorporates a dynamic memory mod- ule to refine initially generated images and utilizes memory writing and response gates to select and fuse relevant textual information, achieving superior performance in synthesizing high-quality images from text descriptions.

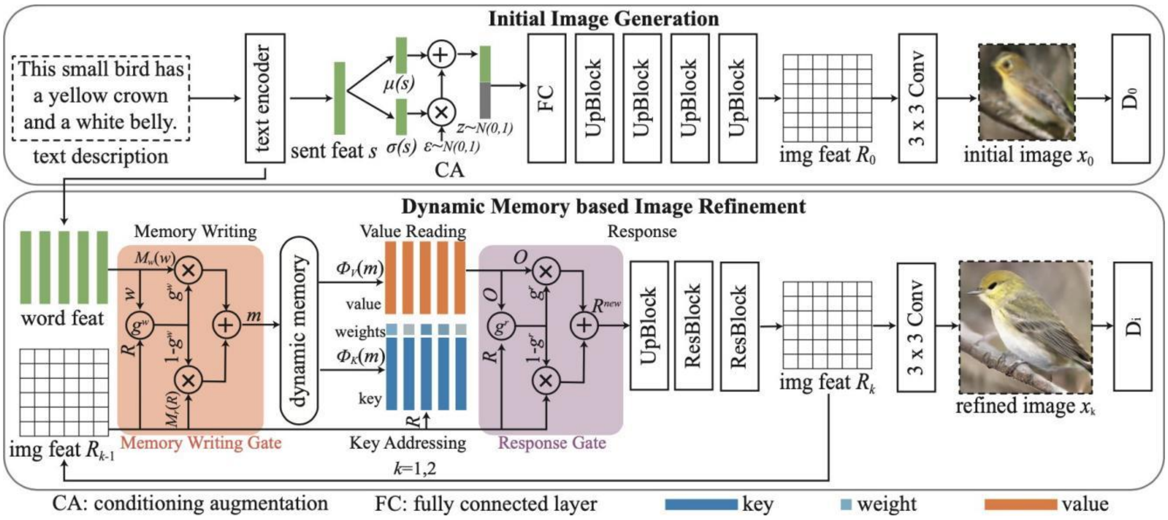

Figure 2. The workflow of DM-GANs [11].

The main idea of DM-GANs [11] can be summarized as follows:

1) To address issues with poorly generated initial images, researchers have integrated a memory mechanism inspired by recent advancements in memory network capabilities into DM-GANs. This innovation involves the incorporation of a key-value memory structure into the existing framework. By integrating this dynamic memory component, the model becomes capable of producing high-quality images, even when the initial image quality is suboptimal.

2) Introducing a memory writing gate [11] to select the words that have tight relationship with the generated images dynamically. English is ambiguous, for traditional text information obtaining architecture, it is difficult to get the key value accurately. Thus, the memory writing gate help figuring this problem.

The principle of DF-GANs can be summarized as follows:

(1) The formula [11] of dynamic memory is shown as follows:

$$W = \{w1, w2, \ldots, wT\}, wi \in RNw, \tag{5}$$

$$R_i = \{r_1, r_2, \ldots, r_N\}, r_i \in \mathrm{R}^{N_r}$$

where W is the given word representations, and the R_i is the image features, T is the number of words, N_w is the dimension of word features, N is the number of image pixels and image pixel features are N_r dimensional vector.

$$m_i = M(w_i), m_i \in \mathrm{R}^{N_m} \tag{6}$$

where the formula [11] above is memory writing and $M(.)$ is the 1x1 convolution operation.

$$\alpha i, j = \frac{\exp\ (\phi_K(m_i)^T r_j}{\sum_{l=1}^{T} \exp\ (\phi_K(m_l)^T r_j}, \tag{7}$$

where $\alpha_{i,j}$ is the similarity probability between the i_{th} memory and the j_{th} image feature and and ϕ_K is the key memory. The above formula represented the key addressing.

$$o_j = \sum_{l=1}^{T} \alpha i, j \phi V\ (mi), \tag{8}$$

where the ϕ_V is the value memory. The above formula is the value reading part.

(2) The formula [11] of gated memory writing is shown as follows:

$$g_i^w(R, w_i) = \sigma\,(A * w_i + B * \frac{1}{N} \sum_{i=1}^{N} r_i), \tag{9}$$

$$m^i = M_w(w_i) * g_i^w + M_r(\frac{1}{N} \sum_{i=1}^{N} (1 - g_i^w)), \tag{10}$$

where the memory writing gate g^w combines image features R_i, and the W is the word features, σ is the sigmoid function, A is a 1 x N_w matrix, B is a 1 x N_r matrix. For the second formula, the M_w and M_r denote the 1x1 convolution operation.

2.3 Introduction of AttnGAN

The AttnGAN model shown in Figure 3 uses attention-driven mechanisms to generate fine-grained images from text by focusing on specific words, refining the image in stages, and employing a Deep Attentional Multimodal Similarity Model for precise image-text matching.

The main idea of AttnGAN [12] can be summarized as follows:

1) Attentional generative network: The generative network employs a two-stage process. Initially, it utilizes the global sentence vector to generate a low-resolution image. In subsequent stages, an attention layer queries word vectors using the image vector from each sub-region to create a word-context vector. This word-context vector is then combined with the regional image vector to form a multimodal context vector. This process effectively enhances the resolution, resulting in a higher resolution image.

2) Deep Attentional Multimodal Similarity Model (DAMSM). This model is created to calculate the similarity between a generated image and textual information, leveraging both global sentence-level and fine-grained word-level details.

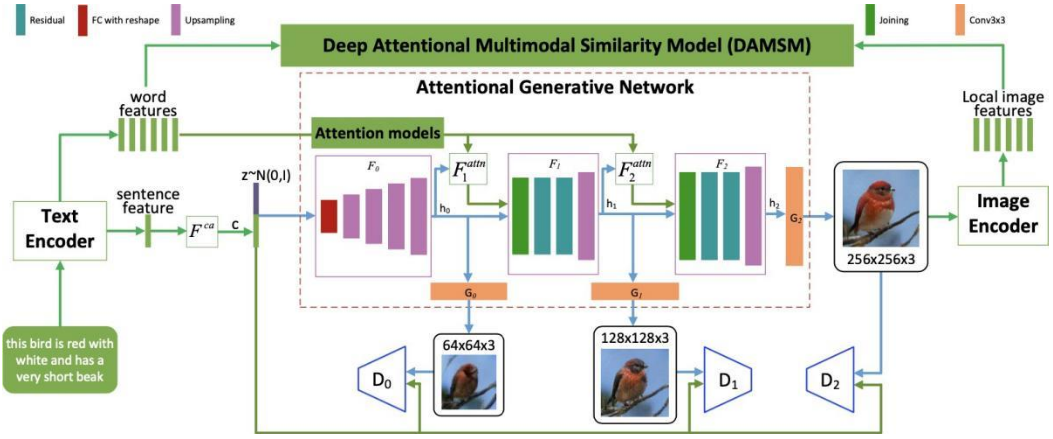

Figure 3. The workflow of AttnGAN [12].

The principle of AttnGAN can be summarized as follows:

1) The formula [12] of Attentional Generative Network is shown as follows:

$$L = LG + \lambda LDAMSM\ , \tag{11}$$

$$Where\ L_G = \sum_{i=0}^{m-1} LG_i,$$

where is the hyperparameter to balance two terms of the equation.

2) The formula [12] of Deep Attentional Multimodal Similarity Model is shown as follows:

The attention-based image-text matching score between the entire image (Q) and the complete text description (D) is defined as follows:

$$R(Q, D) = \log\left(\sum_{i=1}^{T-1} \exp\left(v_2 R(c_i e_i)\right)\right)^{\frac{1}{v_2}} \tag{12}$$

The DAMSM loss part formula:

$$P(D_i|Q_i) = \frac{\exp\left(v_3 R(Q_{i,} D_i)\right)}{\sum_{j=1}^{M} \exp\left(v_3 R(Q_{i,} D_j)\right)} \tag{13}$$

where υ_3 is a smoothing factor determined by experiments.

$$L_1^w = -\sum_{i=1}^{M} logP(D_i, Q_i), \tag{14}$$

$$L_2^w = -\sum_{i=1}^{M} logP(Q_i, D_i), \tag{15}$$

The following is the final DAMSM loss [12]:

$$L_{DAMSM} = L_1^w + L_2^w + L_1^s + L_2^s \tag{16}$$

3. Application and discussion

Based on the DF-GANs model [3] features, the main application related is computer graphics for movies and games. Production companies could use textual descriptions to help prototype scenes, characters, or objects in films or video games. Comparing with traditional text to image model, the DF-GANs is simpler on the architecture and generate text-matching images efficiently as well. Also, it could use the datasets sufficiently [3].

Meanwhile, the limitations should also be put forward: Firstly, the DF-GANs only can deal with the text information in sentence-level. If the text information given is really specific and detailed in paragraph, the DF-GANs do not have enough ability running. Secondly, many pre-trained large language model [13,14] is really popular, if the DF-GANs could combine the large language model in, it is a great improvement on the performance.

Based on the DM-GANs [11] model features, the suitable application related is refining the images in low quality to high quality. According to the application, the model has some limitation, it would be illustrated below: 1. Dependency on the quality of initial images: Existing methods for text-to-image synthesis heavily rely on the quality of the initial image that is generated. If this initial image is poorly generated, the subsequent refinement processes find it challenging to improve the image to a satisfactory quality. The DM-GAN [11] introduces a dynamic memory module to refine unclear or fuzzy image content when initial images are not well-generated. 2. Unchanged text representation in image refinement: Each word in a text description may contribute differently to the image's content. However, many existing methods use an unchanged text representation throughout the image refinement processes. DM-GAN addresses this by designing a memory writing gate that selects the vital text information based on the initial image content. This enables a more accurate generation of images that are conditioned on the text description.

For the future development on this model, here are some general directions: 1. Expand to more diverse datasets: While the model is evaluated on the Caltech-UCSD Birds 200 and Microsoft COCO datasets, testing and refining it on more diverse and complex datasets can be a direction for future work. 2. Further improvement of the memory module: While the dynamic memory module is a significant step forward, there's always room for refining its design or exploring other architectures to improve its efficiency and accuracy.

Leveraging the features of the AttnGAN model [12], this model is tailored for precise text-to-image generation, emphasizing fine-grained details. It synthesizes images guided by natural language descriptions, with a particular emphasis on intricate image regions by attentively considering relevant words within the description. The AttnGAN accomplishes this feat through an innovative attention-driven generative network. This attention mechanism empowers the model to concentrate on specific portions of the text description when generating corresponding segments of the image, thereby ensuring that the generated images exhibit intricate details closely aligned with the textual descriptions.

For the future development about this model, here are some directions should be considered: 1. Wider fields of applications: While the primary application discussed is text-to-image synthesis, the architecture and principles of AttnGAN could be adapted to other tasks where fine-grained attention to specific details is crucial. This includes tasks like image-to-text synthesis, image captioning, or even video generation from textual descriptions. 2. Improvement on Resolution and Quality: As the model seems to already show promise in generating higher quality images by attending to word-level details, future iterations would better be focusing on generating even higher resolution images or more complex scenes, leveraging advancements in computational capabilities and neural network architectures.

4. Conclusion

In this work, the main goal is to analyze and compare mainstream GAN models based on their architecture, mathematical theory and workflow. It discussed the advantages of three models and their suitable applications, limitations and future development. Each model has many creative improvements on their own field based on traditional text to images GAN models. DF-GANs eliminates generator entanglements, enhances text-image semantic consistency without extra networks, and enables deeper, more effective text-image feature fusion, thereby efficiently synthesizing more realistic and text-coherent high-resolution images. DM-GANs significantly improves image generation from text descriptions by utilizing a dynamic memory module and adaptive gates to refine initial images and accurately incorporate varying levels of importance from text information, thereby achieving better qualitative and quantitative performance in generating high-quality images. AttnGAN significantly improves upon traditional models by enabling fine-grained text-to-image synthesis through attention mechanisms, which attend to relevant words for different parts of the image, leading to multi-stage refinement and better image-text matching. In the future, the planning is combining the advantages of three models and produce a refinement model, which is suitable for more complicated applications.

References

[1] Goodfellow I et al 2014 Generative adversarial nets, in Advances in neural information processing systems pp 2672–2680

[2] Huang G Jafari A H 2023 Enhanced balancing GAN: Minority-class image generation Neural computing and applications 35(7): 5145-5154

[3] Tao M Tang H Wu F et al. 2022 Df-gan: A simple and effective baseline for text-to-image synthesis Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16515-16525

[4] Zhang H Xu T Li H et al. 2017 Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks Proceedings of the IEEE international conference on computer vision 5907-5915

[5] Zhang H Xu T Li H et al. 2018 Stackgan++: Realistic image synthesis with stacked generative adversarial networks IEEE transactions on pattern analysis and machine intelligence 41(8) 1947-1962

[6] Park D H Azadi S Liu X et al. 2021 Benchmark for compositional text-to-image synthesis Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)

[7] Yu F Wang L Fang X et al. 2020 The defense of adversarial example with conditional generative adversarial networks Security and Communication Networks 2020: 1-12

[8] Xu T Zhang P Huang Q et al. 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks Proceedings of the IEEE conference on computer vision and pat- tern recognition 1316-1324.

[9] Perez E Strub F De Vries H et al. 2018 Film: Visual reasoning with a general conditioning layer Proceedings of the AAAI conference on artificial intelligence 32(1)

[10] Lim J H Ye J C 2017 Geometric gan arXiv preprint arXiv:1705.02894

[11] Zhu M Pan P Chen W et al. 2019 Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 5802-5810

[12] Xu T Zhang P Huang Q et al. 2018 Attngan: Fine-grained text to image generation with attentional generative adversarial networks Proceedings of the IEEE conference on computer vision and pat- tern recognition 1316-1324

[13] Devlin J Chang M W Lee K et al. 2018 Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805

[14] Radford A Wu J Child R et al. 2019 Language models are unsupervised multitask learners[J]. OpenAI blog, 1(8): 9