A Review of Generative Adversarial Networks for Text to Image Tasks

Zihan Wo[a]
Faculty of Science, The University of Melbourne, Melbourne, 3010, Australia

Keywords: Text to Image Generation, GAN, Training and Testing Datasets.

Abstract: To deal with the task of text-to-image generation, many models have been created in the past decade. In these models, Generative Adversarial Network (GAN) is a widely used basic model. Many models are developed based on GAN or some models are developed based on GAN. With the development of the research, the performance of models is getting better. From the vague and unreal images generated by the primitive models, to clear and reasonable images generated by newer models, the modeling of this task is gradually becoming more refined, and people's understanding of this task is also being more completed. This paper will discuss the development process of the models by comparing several models with representative structures as a reference for subsequent researchers. Through this exploration, this paper aims to highlight the major developments and difficulties in text-to-image production, offering insights for future paths and possible enhancements in this quickly developing subject.

1 INTRODUCTION

Text to image generation has been a fast-developing region in the past decade. Its main research question is: how to generate an image as real as possible that fits the text semantics. Using a Generative Adversarial Network (GAN) for generative tasks is a frequently used generative approach. (Hong et al., 2018; De Rosa et al., 2021) The modeling has become progressively better in recent years, mainly in improving the quality of the generated images, rationalization, and the correlation of the text and the generated images. Through the development of the modeling, many kinds of model structures that have been optimized on the basis of previous research have been born. From the earliest model that just used a GAN baseline to complete this task, to adding more auxiliary models on top of GAN structures for image generation for better clarity, more reasonable, and better correlation with original text. The earlier models might just generate coarse images that are vague, and not much correlated to the text. However, the best performing models nowadays have been developed to produce very good quality images in a very short time.

The overall structure of GAN is to train two separate models, one for the generator and one for the discriminator. The generator initializes the input noise for image generation and improves the parameters based on the judge information given by the discriminator. After receiving the images produced by the generator, the discriminator calculates the likelihood that the image is authentic (Goodfellow et al., 2014). In GAN, the concept of competition is applied. In this game, competition pushes both teams to refine their strategies until the fakes are identical to the real ones (Goodfellow et al., 2014). The following models are all optimized based on GAN to get better generation. From the perspective of the development process, every model is representative of each time period, using different or partially the same auxiliary algorithms. They raise new questions that the previous models have and solve these problems.

This paper reviews the development of text image generation by listing a few model structures that are representative of the process of developing this field and discussing their advantages and disadvantages. This might serve as a reference for subsequent researchers.

[a] https://orcid.org/0009-0004-5335-0781

Wo, Z.
A Review of Generative Adversarial Networks for Text to Image Tasks.
DOI: 10.5220/0013699800004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 487-491
ISBN: 978-989-758-765-8

2 TEXT TO IMAGE GENERATION METHOD BASED ON GAN

2.1 GAN-INT-CLS Model

Introducing Matching-aware Discriminator on top of the original GAN. Original GAN uses image and text pairs as joint inputs to the discriminator, and the discriminator determines the probability of it to be a real image. This approach ignores whether the image is semantically matched to the text, causing a poor learning result. This model increases the correlation of text and generated image by adding input of real images with mismatched text while training the discriminator. This can help the model to learn judge whether the image is correlated to the text description. Furthermore, this model also introduces a manifold interpolation to further promote the performance. By interpolating the text embeddings in the training set, a large number of additional text embeddings are generated. This enhances the generator's ability to learn from the data distribution. Meanwhile, the discriminator further pushes the generator learning on the details of the data manifold by distinguishing between the interpolated text and the image matches. This can improve the variegation and quality of the generated image (Reed et al., 2016).

Scott Reed et al. test the model using the CUB dataset and the Oxford-102 dataset. They mainly do the comparison of internal modules, comparing this model with GAN baseline, GAN-CLS including an image-text matching discriminator, GAN-INT including a text manifold interpolation, and GAN-INT-CLS including these two modules. Human assessments are mainly used in the testing. In CUB's test, GAN and GAN-CLS obtained some correct color information, but the image did not look real. GAN-INT and GAN-INT-CLS get reasonable images with all or at least some of the caption matches in most of the time. In Oxford-102’s test, all four methods can get reasonable images that fit the text. Original GAN is the most diverse in flower morphology. If the text does not indicate this part, the model will give very diverse flower morphology, while the other methods will give more regular images. The result shows that this model improves the authenticity compared with the original GAN. The images seem to be more like real images (Reed et al., 2016).

2.2 StackGAN Model

Based on the original GAN, this model is separated into two halves. The first stage is to outline the original shapes and colors to produce a low-resolution image. The next stage is to fix the flaws in the image created in the first phase, enhance the image's features, and produce a realistic, high-resolution image (Zhang et al., 2017). The generating stage is divided in this model. The quality of the generated images is enhanced by the addition of residual blocks in the second learning step for text and image attributes. Furthermore, a Matching-Aware Discriminator is added in the second stage to improve the consistency of text and image.

Han Zhang et al. test the model using three datasets, CUB, Oxford-102, and MS COCO. They compare this model with GAN-INT-CLS and GAWWN. They use IS as the indicator, and also conduct human assessments to compensate for the inability of IS to assess the consistency of generated images with text. StackGAN gets the best IS score and human assessment ranking. Compared with GAN-INT-CLS, StackGAN improves IS by 28.47% (from 2.88 to 3.70) on the CUB dataset and 20.30% (from 2.66 to 3.20) on Oxford-102. The human assessment ranking also shows that this model can generate images that are more real based on the text. The images generated by GAN-INT-CLS are lack of details and they are not realistic enough. For GAWWN, it cannot generate any reasonable image when the condition is only textual description (Zhang et al., 2017).

2.3 AttnGAN Model

This model consists of a deep attentional multimodal similarity model (DAMSM) and an attentional generative network (Xu et al., 2018). By adding attention methods to the GAN baseline, the model may respond to textual keywords by focusing on the relevant areas of the image. Higher-quality images are produced by improving the semantic alignment of text and image regions. A convolutional neural network (CNN) is used as the image encoder and a bi-directional long short-term memory (LSTM) as the text encoder in the DAMSM. The method calculates fine-grained losses in image production and assesses text-image similarity at the word level by mapping subregions of images and words into a single semantic space (Xu et al., 2018). This can help to improve the consistency.

Tao Xu et al. test the model using the CUB and MS COCO. They use IS as the indicator to assess the image quality and R-precision as a complementary indicator to assess how well the generated images are based on the original text. The model is compared with GAN-INT-CLS, GAWWN, StackGAN, StackGAN-v2 and PPGN. In CUB's test, AttnGAN achieved an IS of 4.36, which is significantly better than the best score of 3.82 for all previous methods, while COCO's best IS improved from 9.58 to 25.89. The results show that the AttnGAN generates higher resolution images compared to other models. Meanwhile, for the AttnGAN model itself, when the hyperparameter λ in the model objective function is increased, the IS as well as the R-precision of the model itself is improved. It shows that the proposed attention mechanism has a significant impact on model optimization (Xu et al., 2018).

2.4 DM-GAN Model

This model consists of two stages: a crude creation and an refinement stage based on dynamic memory. Memory Writing, Key Addressing, Value Reading, and Response are the four components that make up the refinement stage. This model's primary innovation is Memory Writing, which embeds word features into the memory feature space using a convolution operation (Zhu et al., 2019). This can calculate the importance of words, and highlight the important words' information. It enables the model to use related words to do the refinement, instead of using partial text information, and some sentence-level information. It does the refinement in word-level, more nuanced than the previous models.

Minfeng Zhu et al. test the model using CUB and MS COCO. They use IS as the indicator to assess the image quality and R-precision as a complementary indicator to assess the consistency. A lower FID indicates that there is less separation between the generated and actual image distributions. The model is compared with GAN-INT-CLS, GAWWN, StackGAN, StackGAN-v2, PPGN, and AttnGAN. The IS of the DM-GAN model improves from 25.89 to 30.49 (17.77%) on the COCO dataset and from 4.36 to 4.75 (8.94%) on the CUB dataset, both of which are noticeably better than the other methods. The outcomes demonstrate that the DM-GAN model produces images of superior quality in comparison to alternative techniques. As DM-GAN improves its comprehension of the data distribution, FID decreases from 23.98 to 16.09 on CUB and from 35.49 to 32.64 on MS COCO. The CUB and COCO have seen improvements in R-precision of 4.49% and 3.09%, respectively. A higher R-precision means that the images generated by DM-GAN are more accurate in relation to the textual description. This further demonstrates the efficacy of the dynamic memorization technique (Zhu et al., 2019).

2.5 SD-GAN Model

This model uses Siamese to extract common semantics from the text. This enables the model to deal with generation bias due to expression differences, and solve the semantic consistency problem brought by different expressions. Meanwhile, semantic diversity and details are kept to get a more detailed generation. The core module of the model is divided into a text encoder and a hierarchical GAN (Yin et al., 2019). The text encoder uses a bi-directional LSTM to extract semantic features. The hierarchical GAN uses several generators to progressively generate images from low resolution to high resolution. Semantic-Conditioned Batch Normalization (SCBN) is also introduced in this model to enhance the embedding relationship between visual features and textual semantics. It enables the linguistic embedding to manipulate the visual feature maps by scaling them up or down, negating them, or shutting them off (Yin et al., 2019).

Guojun Yin et al. test the model using the CUB and MS COCO datasets. They use IS as the indicator to assess the image quality. They evaluate the image quality using IS as the indicator. To determine whether the produced images match the written description, they also employ a human evaluation procedure. GAN-INT-CLS, GAWWN, StackGAN, StackGAN++, PPGN, AttnGAN, HDGAN, Cascaded C4Synth, Recurrent C4Synth, LayoutSynthesis, and SceneGraph are the models that are compared with IS. The previous best IS on CUB was 4.36 for AttnGAN and 4.67 for SD-GAN. The previous best IS on MS COCO was 25.89 for AttnGAN and 35.69 for SD-GAN. The outcome demonstrates that SD-GAN produces the best-quality images.

For human evaluation, SD-GAN is contrasted with StackGAN and AttnGAN. When evaluating the images produced by these three models on CUB, the testers select the SD-GAN image as the best 68.76% of the time. Additionally, this figure is 75.78% for MS COCO. This demonstrates how well the images produced by SD-GAN match the original textual description. In general, SD-GAN produces images that are more consistent and of higher quality than those produced by earlier models (Yin et al., 2019).

2.6 MirrorGAN Model

In this model, the attentional generative network is augmented with an additional stage. After image generation, MirrorGAN will recreate the image's textual description, ensuring that the underlying semantics match the provided text. This paradigm introduces a Semantic Text Regeneration and Alignment Module (STREAM). A popular image caption system with encoder and decoder serves as the foundation for the basic STREAM design (Qiao et al., 2019). The encoder is a convolutional neural network (CNN), while the decoder is a recurrent neural network (RNN). The RNN is given the image and text encoding in order to generate the word distribution probabilities and accomplish the alignment of the image and text information.

Tingting Qiao et al. test the model using the CUB and MS COCO datasets. GAN-INT-CLS, GAWWN, StackGAN, StackGAN++, PPGN, and AttnGAN are used to compare the model. IS serves as a gauge for evaluating the caliber of the images produced. Images and original text are evaluated for consistency using R-precision. Additionally, the outputs of image production are evaluated overall using human evaluations. The highest IS is obtained by MirrorGAN on the COCO and CUB datasets. The IS of MirrorGAN in the CUB test is 4.56, which is higher than the 4.36 of the previously optimal AttnGAN. And on the COCO dataset, MirrorGAN gets 26.47, better than 25.89 for AttnGAN. The result shows that MirrorGAN can generate a wider variety of images with better quality. Meanwhile, MirrorGAN's R-precision scores on CUB and COCO datasets are much better than AttnGAN. For human assessment, the image generated by AttnGAN sometimes loses details. The colors do not match text descriptions, and sometimes shapes look strange. MirrorGAN obtained better results compared to AttnGAN, with more details and consistent colors and shapes. MirrorGAN is better than AttnGAN for semantic consistency and truthfulness (Qiao et al., 2019).

2.7 ControllableGAN Model

This model solves a previous problem that modifying one attribute of a model might cause other attributes to change during generation as well. The model introduces Channel-Wise Attention on top of the AttnGAN to enhance the association of words with specific visual attributes. Also, the model uses a Word-Level Discriminator to provide fine-grained supervision when training the generator. This can ensure each subregion of the image is semantically consistent with the word description. Previous models have been devoted to optimization for image quality and text-to-image correlation, while this model addresses an aspect of visual attributes that has not been focused on in previous models.

Bowen Li et al. test the model using the CUB and MS COCO datasets. They compare this model with StackGAN++ and AttnGAN. IS is used as an indicator to assess the quality of the image generated. R-precision is used to assess the consistency. Also, to further assess whether the model can generate controlled results, L2 reconstruction error is calculated between images generated from original text and images generated from modified text. The model obtains higher IS and R-precision than the other models on CUB and is competitive on COCO, above StackGAN++ and only slightly below AttnGAN. For reconstruction error, ControllableGAN 's reconstruction error is significantly lower compared to the other models, which suggests that ControllableGAN can better preserve the content in the image generated from the original text. In the qualitative comparison, ControllableGAN can accurately manipulate specific visual attributes based on the modification of the given text. In contrast, the other two models are more likely to generate new content when text is modified or to change some visual attributes that are not relevant to the modification. This part of the test highlights the problem of manipulating specific visual attributes that the model primarily addresses (Li et al., 2019).

3 DATASETS AND ASSESSMENT INDICATOR

Text to image tasks mainly uses CUB, Oxford-102 and MS COCO as training and assessing datasets. CUB is a dataset of bird images, co-founded by Stanford University and Peking University. It includes 200 species of birds with about 60 images of each species. Oxford-102 is a dataset of flower images, including 102 species of flower from the UK. Each of the species has 40-258 images. MS COCO is a much bigger dataset created by Microsoft. It includes many kinds of objects, mainly from daily scenes, with complex backgrounds and a greater number of targets. The two main assessment indicators are inception score (IS) and R-precision. Inception score is mainly used to assess the quality of the generated image, while R-precision is mainly used to assess the semantic similarity between the textual description and the generated image. In addition, in many situations, human assessment is added to assess the image generation quality as a whole as well as the semantic similarity. This is to compensate for the lack of comprehensiveness and accuracy that may occur with data assessment.

4 CONCLUSIONS

This paper introduces and discusses several models for solving the text to image task. From the theoretical analysis of sections of different models, and the test data that the researchers raising the model provide, the performance of these models is progressively increasing from the past to the present with years of development. The most primitive model might just generate some vague and highly unreal images, such as the GAN baseline and GAN-INT-CLS model about ten years ago. They can give some basic shapes and colors, but these images do not seem to be real. Also, the original text is not well presented in the images the model generated. But the latest models can generate very clear and realistic images, like the MirrorGAN, which also fits the text very well. And for the ControllableGAN, even starts optimizing the visual attributes, not only focusing on quality and consistency like the previous models. Basically, each model is better than the previous models, which can also be supported by the test data. This paper might serve as a reference for subsequent researchers to study the advantages and disadvantages of these models and the development process of the solution of text to image task.

REFERENCES

De Rosa, G. H., & Papa, J. P. (2021). A survey on text generation using generative adversarial networks. Pattern Recognition, 119, 108098.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Hong, S., Yang, D., Choi, J., & Lee, H. (2018). Inferring semantic layout for hierarchical text-to-image synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7986-7994).

Li, B., Qi, X., Lukasiewicz, T., & Torr, P. (2019). Controllable text-to-image generation. Advances in neural information processing systems, 32.

Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1505-1514).

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).

Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., & Shao, J. (2019). Semantics disentangling for text-to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2327-2336).

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

Zhu, M., Pan, P., Chen, W., & Yang, Y. (2019). Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5802-5810).