# Text Generation Image about Gan

Sijia Ye*

*Chongqing Normal University, Chongqing, China*
**Corresponding author. Email: Ysjyesijia@163.com*

**ABSTRACT**

In recent years, the task of text to image generation has been an important research hotspot in the field of computer vision and natural language. The purpose of this task is to take a descriptive language text as the input, and then output an image with consistent text content. Due to the instability of training such as gradient disappearance and mode collapse in the model training of the generated countermeasure network, and may cause problems such as the inconsistency between the final generated result and the text semantics or the diversity of generated content, based on previous research, this paper proposes Gan's text generated image algorithm, which not only improves the stability of network training, but also improves the clarity of the image, Make the generated image more realistic. This paper studies the text to image generation based on Gan, discusses and analyzes the text to image generation network structure of GaN and the process of text to image generation; Combined with the function algorithm of text generated image of gan-int-cls, the is and vs scores of different models of gan-int-cls, gawwn, stackgan, stackgan + + and hdgan are tested. The results show that the is scores of the two data sets are greatly improved compared with the four methods of gan-int-cls, gawwn, stackgan and stackgan. Compared with hdgan, the is score on oxford-102 data set increased from 3.45 to 3.52. On the cub dataset, the is score increased from 4.18 to 4.33, and the score on the cub dataset reached 0.355, indicating that the hir-gan model has good effect and the generated image quality is higher. The hir-gan model has better image effect than the previous network model and matches the text description better. Through qualitative calculation, the is score and vs score of hir-gan have achieved the best results. Through the visualization experiment, it can also be seen that the image generated by hir-gan is the best and more in line with the semantic information described by the text.

***Keywords:*** *Gan, Text Generation Research, Image Research, Gan's Text Generation Image*

## 1. INTRODUCTION

With the development of Gan network, great breakthroughs have been made. Text image generation is a very complex problem, which requires not only to generate realistic images, but also to meet the description of the text. The current mainstream methods are based on GaN network and have achieved good results.

Many scholars at home and abroad have studied the research of text generated image based on GaN; Yanagi r proposed a new retrieval framework based on GaN from text to image, which greatly improves the performance of scene retrieval through a simple process. The image generated from text to image is used as the query of scene retrieval task. Different from many text image Gans studies that mainly focus on high-quality image generation, it is found that the generated image has reasonable visual features suitable for query [1]. The method developed by aecharungroj V can help other places better understand local brand identity, so as to effectively plan and manage these places. Content analysis was used to study 782 user generated images and 9633 user generated text comments in Google Maps local guide; In order to identify the positive and negative factors of local brand identification, 105 codes are generated by analyzing the content of the image, and the frequency of codes representing local brand identification is counted. It is the first time to use UGC to research from the perspective of text and image recognition. Using these two analysis methods can complement each other and alleviate their shortcomings [2].

Nowadays, more and more algorithms can solve the problem of conversion between multimodal information, especially the text generated image studied in this paper, which promotes the production and life efficiency and alleviates the labor cost. A stacked generation countermeasure network combined with category multi-target loss is proposed. Aiming at the problem that the image generated by Gan lacks detailed features, this paper proposes to combine the multi-objective loss of categories and introduce the category loss of images into the model, so that the model can learn the similar features of images of the same category, so as to improve the ability of model text to generate images [3,4].

 

## 2. RESEARCH ON TEXT IMAGE GENERATION BASED ON GAN

### *2.1. Generate Countermeasure Network Gan*

The generator is composed of two parts: one is the model discriminator and the other is the model discriminator. By learning the probability distribution of training samples, the generation countermeasure network continuously improves the proximity between the generated samples and the real samples, in order to generate data with the same or similar probability distribution as the real samples. The core idea of generating countermeasure network is similar to the Nash equilibrium in game theory. The generator and discriminator in the generated countermeasure network are regarded as both sides of the game. The generator continuously learns its distribution from the training data and generates data with the same distribution as the real data as possible to deceive the discriminator. The discriminator continuously improves its performance to identify whether the data is real data or generated data. Both sides produce a continuous game process, and constantly improve their performance, and finally achieve Nash equilibrium.

As a generative model of unsupervised learning, generative countermeasure network can not rely on any prior assumptions. Compared with other traditional generative models, its way of generating samples is simpler, and from the actual generation effect, generative countermeasure network can generate samples with better quality [5, 6].

The main idea of generative countermeasure network is to complete the learning process through two neural networks, in which the generative network G and the discriminant network d finally train the generative network G which can produce analog data close to the real data through joint learning and interaction. Different from the traditional single neural network structure, Gan applies two seemingly independent neural networks: generation network G and discrimination network D, and "links" the two networks into a whole through the optimization function. For generating network G, the output results need to be optimized by discriminating network D; For discriminating network D, we need to combine the output of generated network G and real data as a "supervised" input to complete our own training. In the whole process, the goal of generating network is to generate results closer to real data, which is reflected in the output results of discriminating network and applied to the optimization process of generating network G. The goal of the discrimination network is to improve the discrimination ability. As a two classifier, it distinguishes the generated data from the real data, and also follows the training rules of the two classifiers in the optimization training process [7, 8]. Ideally, with the continuous improvement of the generation ability of the generation network G, the discrimination network D can not distinguish the difference between the generated data and the real data. Of course, this also implies a condition that the discriminant network D must be strong enough, because for a weak discriminant network, although there are still great differences between the generated data and the real data, the discriminant network may not be able to distinguish. Therefore, in the learning and training process of generative countermeasure network, there is a process of minimization and maximization, that is, for the generative network, the difference between minimization and real data; For discriminant networks, the difference between generated data and real data is maximized. Through such completely opposite purposes, the generation network G and the discrimination network D "confront" each other to complete the training.

### *2.2. Text Image Generation Process Based on GaN*

The generation of scene graph is to extract the entities from the input text by using the scene graph analyzer, obtain the entities and the relationship between entities, and finally draw the generated scene graph. Scene graph analyzer is a python toolkit for parsing natural language into scene graphs. The toolkit has an easy-to-use user interface and an easy to configure design. Nouns in natural language are taken as entities, prepositions and adjectives are analyzed into the relationship between entities, and finally the generation scene is drawn [9, 10].

Firstly, the Gan model needs to process the scene map, convert the objects in the scene map and the positional relationship text between objects into the index in the thesaurus of the data set. After reading all the scene maps in a single image, the model represents all the represented objects and object relationship texts with 64 dimensional word embedding vectors, inputs the word embedding vectors into the graph convolution network, and finally calculates the 128 dimensional output vector corresponding to the objects. After getting the scene map embedding vector, Gan model needs to process the embedding vector through the object layout network to generate the initial object layout image. The object layout network consists of two parts, one is the bounding box regression network, and the other is the mask regression network. The function of bounding box regression network is prediction. The function of mask regression network is to obtain the object information of the target area gradient from the embedding vector of the input object, multiply the prefabricated target area mask with the image to be processed, retain the object image of the target area, and discard the values outside the area.

The data dimension after full connection layer processing becomes smaller. At this time, up sampling is required to restore the image to the original size to realize the mapping of the image from small resolution to large resolution. There are three commonly used up sampling methods: interpolation, deconvolution and de pooling. The up sampling method used in this paper is deconvolution. Deconvolution, also known as transpose convolution, first increases the size of the input image by a certain proportion, and then rotates the convolution kernel for forward convolution [11].

The second stage model adds attention mechanism, in which the first attention mechanism is the entity in the text and the second attention mechanism is the connection between entities. The attention mechanism and implicit features are used as the input for connection operation, which is used as the input of the residual network module. Finally, the output of the residual network module is directly up sampled to obtain the final high-resolution image. The difference between the second stage and the first stage is that the attention network is introduced to take the previously obtained implicit features as the input, so as to add the text description details ignored in the first stage network when generating the image, so as to better generate the detailed rich image in line with the text description.

Residual network is a deep neural network formed by many residual blocks. Ordinary network training algorithms will make more and more training errors with the deepening of network depth, but the difference of residual network is that the training effect is good even if the network is deeper. The emergence of residual network is helpful to solve the problems of gradient disappearance and network degradation. At the same time, when training deeper networks, it can ensure good training performance [12].

## 3. FUNCTION ALGORITHM OF TEXT GENERATED IMAGE BASED ON GAN-INT-CLS

Firstly, the text is processed, and the English text description of the image in the data set is transformed into a vector that can be recognized by the computer $\xi(t)$ Learn the function JT by optimizing the following loss function, such as formula (1):

$$\frac{1}{Z}\sum_{z=1}^{Z}\Delta(b_z.j_v(v_z)) = \Delta(b_z, j_v(t_z)) \tag{1}$$

Where VZ represents the image, TZ represents the text description, BZ represents the category label and represents the 0-1 loss. The formulas of classifiers JZ and JT are as follows (2) (3):

$$\mathrm{j}_v(v) = \arg\max_{b\in B} E_{t-\delta(b)}\left[\varsigma(v)^T \xi(t)\right] \tag{2}$$

$$\mathrm{j}_t(t) = \arg\max_{b\in B} E_{v=v(b)}\left[\varsigma(v)^T \xi(t)\right] \tag{3}$$

Where, $\varsigma$ is the image encoder, $\xi$ is the text encoder, $\delta(b)$ is the text description of category B, and V (b) is the image.

## 4. ANALYSIS OF EXPERIMENTAL TEST RESULTS

In this experiment, the level inversion residual generation countermeasure network is evaluated qualitatively and quantitatively, and the models in this experiment are compared with some existing models, including gan-int-cls, gawwn, stackgan, stackgan + +, hdgan. The quantitative evaluation is calculated by calculating the is and vs scores and randomly selecting 35000 images from the test set. Qualitative evaluation is to directly observe the generated images through visual experiments, and make intuitive comparison with some current models. Comparing the is scores of different models on oxford-102 dataset and cub dataset, the higher the is score, the better the effect of generating the model. The comparison results of is scores of different models are shown in Table 1 and figure 1.

**Table 1.** Comparison of is scores of different models

| Contrast model | Oxford-102 | CUB |
|---|---|---|
| GAN-INT-CLS | 2.68 | 2.90 |
| StackGAN | 3.21 | 3.11 |
| StackGAN++ | 3.26 | 4.07 |
| HDGAN | 3.52 | 4.18 |
| HIR-GAN | 3.57 | 4.33 |

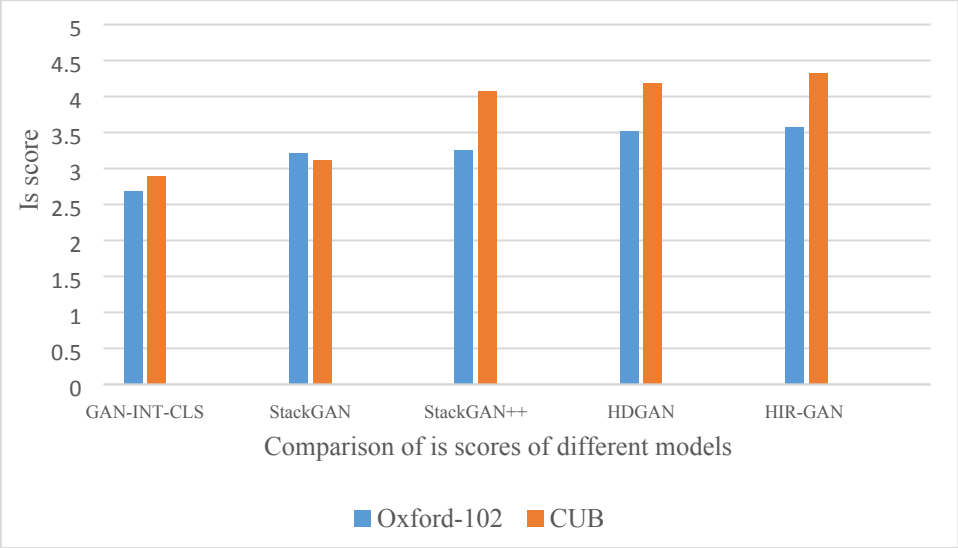


**Figure 1.** Comparison of is scores of different models

It can be seen from the above data that the is score of the two data sets is greatly improved compared with the four methods of gan-int-cls, gawwn, stackgan and stackgan. Compared with hdgan, the is score on oxford-102 data set increased from 3.45 to 3.52. On the cub dataset, the is score increased from 4.18 to 4.33, which has been greatly improved on both datasets, indicating that the hir-gan model has a good effect and the generated image quality is higher.

Next, the vs scores of different models are compared. The vs scores of different models on oxford-102 dataset and cub dataset are compared. The higher the vs score, the better the effect of the generated model, that is, the higher the consistency between the text and the generated image. The data in the figure represents the values described by the real image and the corresponding text. The test results are shown in Figure 2.

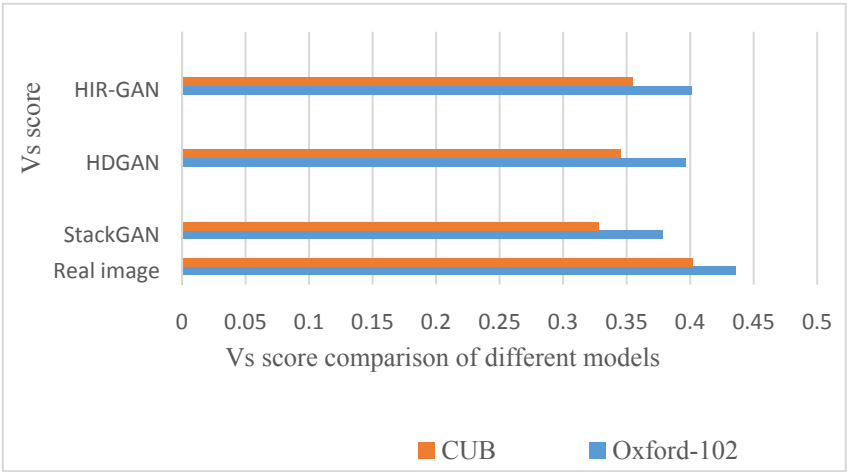


**Figure 2.** Vs score comparison of different models

It can be seen from the table that the hir-gan method has achieved the best results on both data sets, with a score of 0.401 on the oxford-102 data set and 0.355 on the cub data set, which shows that the image generated by the hir-gan model is better than the previous network model and better matches the text description.

## 5. CONCLUSIONS

As the variants derived from generation confrontation network become more and more abundant, several research directions have gradually formed. Text to image generation is one of the more popular and difficult research directions. After continuous innovation in recent years, the application of image generation algorithm is still mining, which makes the application field of image generation more and more extensive. It is a hot focus both in practical application and academic research. Although this paper proposes the research of text image generation based on Gan, there are still some problems that need further research and mining. The prospect of the future research of text to image generation. In the task of text semantic image generation, when image generation based on cgan model, the existing algorithm model is difficult to show the emotional characteristics of the text in the image,

resulting in poor emotional rendering effect of the generated image; In the research of image generation. Therefore, the text generated image based on GaN needs to be further studied.

## REFERENCES

[1] Yanagi R., Ren T., Ogawa T., et al. Query is GAN: Scene Retrieval with Attentional Text-to-image Generative Adversarial Network. IEEE Access, 2019, PP(99):1-1.

[2] Taecharungroj V., User-generated place brand identity: harnessing the power of content on social media platforms. Journal of place management and development, 2019, 12(1):39-70.

[3] Klostermann J., Plumeyer A., Bger D., et al. Extracting brand information from social networks: Integrating image, text, and social tagging data. International Journal of Research in Marketing, 2018, 35(4):538-556.

[4] [Mahwish B., Tariq S., Romana T., et al. Image reconstruction and text embedding using scan patterns with XOR in graph cut technique. Journal of Intelligent and Fuzzy Systems, 2017, 33(2):1-8.

[5] Khan M Z., Jabeen S., Khan M., et al. A Realistic Image Generation of Face from Text Description using the Fully Trained Generative Adversarial Networks. IEEE Access, 2020, PP(99):1-1.

[6] Lydia W., Ioana H., Paolo P S., et al. Knowledge-rich image gist understanding beyond literal meaning. Data & Knowledge Engineering, 2018, 117(SEP.):114-132.

[7] Yiming, Gan, Chunhai, et al. Research of Software Defect Prediction Based on GRA-SVM.. AIP Conference Proceedings, 2017, 1890(1):1-5.

[8] Gan Y., Zhang B., Ke C., et al. Research on Robot Motion Planning Based on RRT Algorithm with Nonholonomic Constraints. Neural Processing Letters, 2021, 53(4):3011-3029.

[9] Aris H., Ibrahim Z., Azman A., Simple Screen Locking Method Using Randomly Generated Number Grid on Image. International Journal of Mobile Human Computer Interaction, 2018, 10(4):42-71.

[10] Ramteke G D., Ramteke R J., Efficient Model for Numerical Text-To-Speech Synthesis System in Marathi, Hindi and English Languages. International Journal of Image Graphics & Signal Processing, 2017, 9(3):1-13.

[11] Srilakshmi P., Himabindu C., Chaitanya N., et al. Text embedding using image steganography in spatial domain. International Journal of Engineering & Technology, 2018, 7(3):1-4.

[12] Shiravale S S., Jayadevan R., Sannakki S S., Devanagari Text Detection From Natural Scene Images. International Journal of Computer Vision and Image Processing, 2020, 10(3):44-59.