BA-GAN: Bidirectional Attention Generation Adversarial Network for Text-to-Image Synthesis

Ting Yang, Xiaolin Tian, Nan Jia, Yuan Gao, Licheng Jiao

BA-GAN: Bidirectional attention generation adversarial network for text-to-image synthesis*

Ting Yang, Xiaolin Tian, Nan Jia, Yuan Gao, and Licheng Jiao

School of Artificial Intelligence, Xidian University, Xi'an 710071, China
{xltian, lchjiao}@mail.xidian.edu.cn
{20171213676,*ygao*_5}@stu.xidian.edu.cn, jiananxidian@163.com

Abstract. It is difficult for the generated image to maintain semantic consistency with the text descriptions of natural language, which is a challenge of text-to-image generation. A bidirectional attention generation adversarial network (BA-GAN) is proposed in this paper. The network achieves bidirectional attention multi-modal similarity model, which establishes the one-to-one correspondence between text and image through mutual learning. The mutual learning involves the relationship between sentences and images, and between words in the sentences and sub-regions in images. Meanwhile, a deep attention fusion structure is constructed to generate a more real and reliable image. The structure uses multi branch to obtain the fused deep features and improves the generator's ability to extract text semantic features. A large number of experiments show that the performance of our model has been significantly improved.

Keywords: text-to-image generation · BA-GAN · mutual learning.

1 Introduction

In recent years, text-to-image synthesis is a hot research topic. It covers two major areas, Natural Language Processing [13] and Computer Vision [1], which can be used in the interaction of art generation and entertainment.

At present, the research of text-to-image synthesis based on GAN [2] has become the mainstream trend. The research shows that the adversarial training of generator and discriminator in GAN can promote the matching relationship between the generated image and the text semantics.

Attention mechanism has also been applied to text - to - image synthesis in previous studies. AttnGAN [14] introduced the attention mechanism for the first time. The mechanism guides the generator to focus on the words in the sentence related to the different sub regions of the image. But AttnGAN only considers the context vector of sub-regions base on sentences. And this ignores some fine-grained information between two modalities and leads to inaccuracy of text-image matching. Aiming at this problem, this paper proposes a bi-directional attention multimodal similarity model (BAMSM).

The quality of the generated image is still not satisfactory in the initial

* The work is supported by the National Natural Science Foundation of China (No. 61977052).

stage although it has made great progress to use multi-stage GAN for text-to-image synthesis. In this paper, the deep attention fusion structure is proposed to improve the feature extraction ability of generator and get more high-quality initial image.

The main contribution of this paper as follows:

We propose BAMSM, a bidirectional attention multimodal similarity model, which calculates the cross-modal similarity through mutual learning.

A deep attention fusion structure (DAFS) is proposed to improve the feature extraction capability of the generator and integrate more information to achieve the generation of high-quality images.

The channel perception adjustment module(CPAM) is proposed to promote the generation of high-quality initial images by extracting multi-level features.

2 Related Work

It is a basic challenge for text-to-image synthesis to determine high semantic consistency between text and image in the field of computer vision.

In recent years, a variety of generation models based on GAN have appeared successively with the development of deep learning. The original GAN-INT-CLS [9] generated images with a resolution of only 64*64 and low matching degree with text. GAWWN [10] is proposed to get the generated images with a resolution of 128*128. StackGAN [15] uses two stages as basic network architecture to generate images that meet the requirements. In the first stage, rough images similar to the texture, color and content of text description are generated. Then in the second stage, the initial image is refined continuously to synthesize the final image. StackGAN++ [16] is an improvement of the end-to-end model that generates higher quality images. However, the above methods always use global sentence vectors when selecting conditional constraints, resulting in the omission of word information in sentences in the process of image generation. Furthermore, some details of text semantics are lost.

AttnGAN [14] is proposed to solve this problem. It can find the word that is most relevant to the image sub-region by considering the context vector of sub-regions base on sentences, and then calculates the similarity between image and text. A text-image-text circular structure is proposed by MirrorGAN [8], which transformed the generated image into text through encoding, and then compared it with the given text description. Obj-GAN [6] proposed a new object-driven attentional image generator and a target recognizer based on Faster RCNN. Compared with the above methods, the BAMSM proposed in this paper has its advantages. It mainly points at the mutual learning between the word features of the text and the sub region features of the image, so as to improve the semantic consistency between the generated image and the input text.

3 Our model

In this paper, we propose a bidirectional attention generation adversarial network for text-to-image synthesis. The network is always used to extract information from text and generate corresponding image. The network structure is shown in Figure 1.

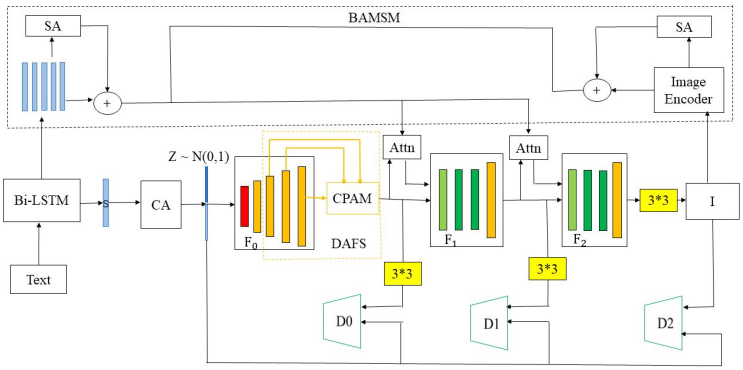

Fig. 1. The architecture of the proposed BA-GAN. I indicates the generated image.

3.1 Text encoder and Image encoder

Text encoder Bi-LSTM [4] is used as a text encoder to encode the input sentences. The output of the encoder is the word feature matrix e^{D*T}, where D is the dimension of the word vector and T is the number of words in the sentence. In addition, the last hidden state of Bi-LSTM is taken as the global vector of the sentence, representing $\overline{e} \in R^D$, which is used for similarity comparison with the global feature of the image.

Image encoder Inspired by previous studies, we use the pretrained network Inception-v3 [11] to extract image features. Firstly, we scale the image to 299*299 resolution and input into the encoder. Then local feature $f \in R^{768*N}$ is extracted from the "$mixed\_6e$" layer in Inception-v3, where 768 is the dimension of local feature vector, N is the number of image sub-regions. Meanwhile, global features $\bar{f} \in R^{2048}$ are extracted from the last layer of Inception-v3. Finally, image features are mapped to the same semantic space of text features, and two new feature vectors $v \in R^{D*N}$ and $\bar{v} \in R^{2048}$ are obtained.

Bidirectional attention multi-modality similarity model The principle of BAMSM is to calculate the similarity between the words and the sub-regions through mutual learning. The first, we input e and v into the self-attention module, and extract the self-attention weights w_e and w_v respectively. The main purpose of this operation is to find the context weight inside the modality. Then the context proportion between text and image is calculated by bidirectional attention. Finally, we calculate the image-text matching score.

We multiply weights with the vector itself to obtain the weighted word vector and image vector respectively, we compute them with:

$$e' = e + e * SA(e) \tag{1}$$

where SA represents the self-attention extraction module. e' is the weighted word vector. We define the formula (1) as $v' = v + v * SA(v)$ to obtain the weighted sub-region vector. Then we calculate the similarity matrix s between the word vector and the sub-region vector :

$$s = e'^T v' \tag{2}$$

where $s \in R^{T*N}$, $s_{i,j}$ represents the similarity between the i-th word and the j-th subregion. We take $w = s^T$, and then normalize the s and w matrices respectively, they are computed by:

$$\overline{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1} \exp(s_{k,j})},\ \overline{w}_{j,i} = \frac{\exp(w_{j,i})}{\sum_{k=0}^{N-1} \exp(s_{j,k})} \tag{3}$$

where $w_{j,i}$ is similarity between the j-th sub-region and the i-th word. Then, the bidirectional attention mechanism is used to get the word context vector and sub-region context vector. They are calculated respectively as follows:

$$c_i = \sum_{j=0}^{N-1} \alpha_j {v_j}',\ c_j = \sum_{i=0}^{T-1} \alpha_i {e_i}', \tag{4}$$

where $\alpha_i = \frac{\exp(\gamma \overline{w}_{j,i})}{\sum_{k=0}^{T-1} \exp(\gamma \overline{w}_{j,k})}$, $\alpha_j = \frac{\exp(\gamma \overline{s}_{i,j})}{\sum_{k=0}^{N-1} \exp(\gamma \overline{s}_{i,k})}$, γ is a hyperparameter.

Finally, we calculate the matching score of text vector and image vector using cosine similarity theorem. Inspired by the minimum classification error formulation [5], the match score of text and image based on bidirectional attention mechanism is computed as follows:

$$R(Q, D) = \log\left(\sum \exp(\gamma_0 R(c_i, {e_i}'))\right)^{\frac{1}{\gamma_0}} \tag{5}$$

$$R(D, Q) = \log\left(\sum \exp(\gamma_0 R(c_j, {v_j}'))\right)^{\frac{1}{\gamma_0}} \tag{6}$$

where $R(c_i, {e_i}') = \frac{c_i {e_i}'}{\|c_i\| \|{e_i}'\|}$, $R(c_j, {v_j}') = \frac{c_j {v_j}'}{\|c_j\| \|{v_j}'\|}$. D is the text and Q is the image corresponding to the text. γ_0 is a hyperparameter.

Loss Function Text and image encoder aims to make text and image learn from each other and achieve better image-text matching. The calculation process of text-image matching score is different from DAMSM.

$$L_{BAMSM} = {L_1}^w + {L_2}^w + {L_1}^s + {L_2}^s \tag{7}$$

Where, ${L_1}^w$ and ${L_1}^s$ indicates the word loss and sentence loss when the image matches with the given text, and ${L_2}^w$ and ${L_2}^s$ represents the word loss and sentence loss when the text matches with the given image.

$${L_1}^w = -\sum_{i=1}^{M} \log P(D_i|Q_i),\ {L_2}^w = -\sum_{i=1}^{M} \log P(Q_i|D_i) \tag{8}$$

where $P(D_i|Q_i) = \frac{\exp(\gamma_1 R(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_1 R(Q_j, D_i))}$, and it is the probability that the generated image can correspond to the text description; $P(Q_i|D_i) = \frac{\exp(\gamma_1 R(D_i, Q_i))}{\sum_{j=1}^{M} \exp(\gamma_1 R(D_j, Q_i))}$, and it denotes the probability of text corresponding to image. In addition, M refers to the batch size at the time of training. γ_1 is a hyperparameter.

L_s is similar to L_w. We just define formula (5) as $R(Q, D) = \frac{\bar{v}^T \bar{e}}{\|\bar{v}\| \|\bar{e}\|}$, and then substitute it into formula (8) to get the sentence loss in the process of text-image matching.

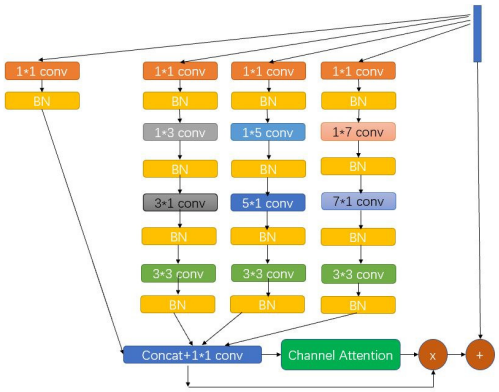

Fig. 2. The architecture of the proposed CPAM.

3.2 Multi-stage Generative Adversarial Networks

Basic framework Three-stage GAN is adopt to synthesize realistic and high-dimensional images in our model. We feed the acquired sentence vector $\bar{e}$ into the initial stage of GAN. The process of generating the final image I is as follows:

$$f_0 = F_0(z, (\bar{e})^{CA}),\ f_i = F_i(f_{i-1}, Attn\,(f_{i-1}, e')),\ I = G_i(f_i) \tag{9}$$

in which z is a random noise vector satisfying normal distribution. $(\bar{e})^{CA}$ represents global sentence features augmented by conditions, F_0, F_i, G_i are neural network model. $Attn$ indicates attention model.

The illustration of the proposed DAFS is shown in Figure 1. There are five layers in F_0. Since the shallow layer has little influence on the performance, the last three layers are cascaded. And then the multi-level and multi-scale features with different information are extracted by utilizing the proposed CPAM.

channel perception adjustment module Inspired by RFB network [7], this module can obtain features of different depths by using multi-branch structure. Figure 2 depicts the structure of CPAM. There are five branches in this structure. Four of which are mainly for extracting multi-level and multi-scale features from the input feature graph. 1*1 convolution is to fuse multi-channel information from the input characteristic graph.

Loss function In order to generate a more realistic and clearer image, we define the generator loss function as:

$$L = L_G + \alpha L_{BAMSM},\ L_G = \sum L_{G_i} \tag{10}$$

where $L_{G_i} = -\frac{1}{2}E_{\widehat{I}_i \sim P_{G_i}}\left[\log D_i(\widehat{I}_i)\right] - \frac{1}{2}E_{\widehat{I}_i \sim P_{G_i}}\left[\log D_i(\widehat{I}_i, \bar{e})\right]$. α is the hyper-parameter, which is used to indicate the importance of BAMSM. The generator and discriminator carry on adversarial training in the GAN network. The loss function of the discriminator is computed as:

$$\begin{aligned} L_{D_i} = &-\frac{1}{2}E_{I_i \sim P_{real_i}}\left[\log D_i\left(I_i\right)\right] - \frac{1}{2}E_{\widehat{I}_i \sim P_{G_i}}\left[\log(1 - D_i\left(\widehat{I}_i\right)\right] \\ &-\frac{1}{2}E_{I_i \sim P_{real_i}}\left[\log D_i\left(I_i, \bar{e}\right)\right] - \frac{1}{2}E_{\widehat{I}_i \sim\, P_{G_i}}\left[\log(1 - D_i\left(\widehat{I}_i, \bar{e}\right)\right] \end{aligned} \tag{11}$$

where I_i is real image from sampling the distribution P_{real_i} of stage i, $\hat{I}_i$ is the generate image from sampling the distribution P_{G_i} of stage i.

4 Experiments

Dataset and Evaluation Metrics As in previous studies on text-to-image synthesis, we evaluated the model on the CUB and COCO datasets. In order to evaluate the performance of the model and measure whether the generated image is true or false, Inception score(IS) [12], R-precision [14] and Fréchet inception distance(FID) [3] are used for evaluation. In addition, we use some generated images as examples to specify the practical effects of the models as shown in Figure 3.

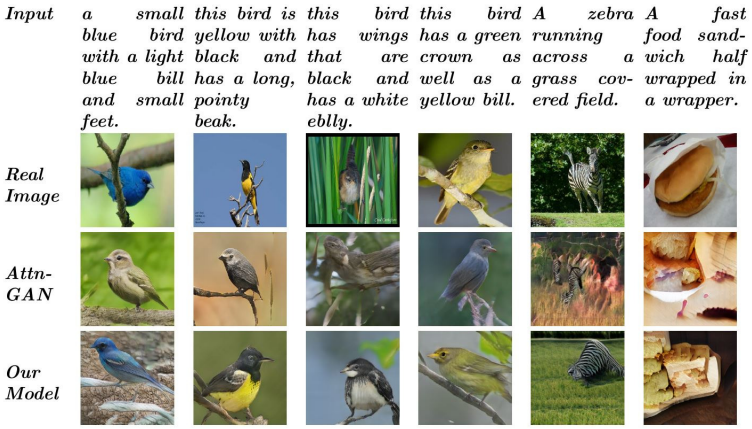

Fig. 3. Comparison of images generated by AttnGAN and Our Model.

Ablation studies In order to study the effect of each part of the proposed method on the quality of generated image in CUB dataset, we conducted some ablation experiments. The baseline model of this paper is AttnGAN. The experimental results are described in Table 1. Effective experiment based on data set on baseline network, we set the hyperparameters as: $\gamma = 5,\ \gamma_0 = 5,\ \alpha = 5,\ \gamma_1 = 10$.

We mainly evaluate the validity of BAMSM and DAFS. The experimental results are described in Table 1. It can be seen the proposed module can improve network performance significantly compared with baseline model.

Table 1. Performance of ablation study both in IS and FID on CUB dataset

Model	IS ↑	FID ↓
Baseline	4.36	23.98
Baseline + BAMSM	4.80	21.84
Baseline + DAFS(F_0)	4.73	23.82
Baseline + DAFS(F_0, F_1, F_2)	4.74	23.72
Baseline + BAMSM + DAFS(F_0)	4.88	20.82
Baseline + BAMSM + DAFS(F_0, F_1, F_2)	4.68	23.04

Function of the proposed BAMSM BAMSM is mainly proposed to better learn the semantic correspondence between texts and images. Text-image matching is realized by considering the relationship within the text and semantic relevance between text and image. In addition, the self-attention mechanism is used to extract the self-attention weights of word vectors and local features of images. The context relations within modes are represented by the weights.

Function of the proposed DAFS DAFS is proposed to generate high quality initial images in the initial stage of the network. It is applied to obtain multichannel and multi-level information. Then images containing more semantic information are generated by fusing them. In additon, the proposed CPAM plays a important role in this module. The experimental results show that DAFS is effective.

Comparative experiment We compared our method with the existing text-to-image synthesis method on CUB and COCO datasets, and the results are shown in Table 2.

Table 2. IS, R-precision and FID scores by existing method and our BA-GAN on CUB and COCO testsets.

Dataset	CUB			COCO		
Method	IS ↑	R-precision ↑	FID ↓	IS↑	R-precision ↑	FID↓
GAN-INT-CLS	2.88	-	68.79	7.88	-	60.62
GAWWN	3.62	-	53.51	-	-	-
stackGAN	3.70	-	35.11	8.45	-	33.88
stackGAN++	4.04	-	25.99	8.30	-	-
AttnGAN	4.36	53.82	23.98	25.89	82.98	35.49
MirrorGAN	4.56	57.51	18.34	26.47	82.44	34.71
Our model	4.88	58.64	20.82	27.79	83.21	31.08

5 Conclusion

In this paper, we propose a generative adversarial network based on bidirectional attention for text-to-image synthesis, abbreviated BA-GAN. Firstly, we build a bidirectional attention multi-modality similarity model to learn the semantic corresponding relationship between text and image. The text encoder containing the image information is obtained through the model. Secondly, we propose a deep attention fusion structure to generate high-quality initial image. Deeper feature and multi-channel information are extracted through multi branch structure to generate clearer initial image. A large number of experiments show the effectiveness of our proposed BA-GAN in the text-to-image synthesis.

References

1. Bissoto, A., Valle, E., Avila, S.: The six fronts of the generative adversarial networks. arXiv preprint arXiv:1910.13076 (2019)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in Neural Information Processing Systems **27** (2014)

3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems **30** (2017)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
5. Juang, B.H., Hou, W., Lee, C.H.: Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio Processing **5**(3), 257–265 (1997)
6. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12174–12182 (2019)
7. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 385–400 (2018)
8. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1505–1514 (2019)
9. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning. pp. 1060–1069. PMLR (2016)
10. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in Neural Information Processing Systems **29** (2016)
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
12. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in Neural Information Processing Systems **29** (2016)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems **27** (2014)
14. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: 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 (2018)
15. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: 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 (2017)
16. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(8), 1947–1962 (2018)