Journal of Computing and Electronic Information Management

ISSN: 2413-1660

A Complex Text Image Generation Method Based on Attention Mechanism

Shengan Zhou
Department of Electronic Information, Guangdong Vocational College of Administration, Guangzhou 510800, China

Abstract: In view of the difficulty of anti network text image generation, a method of anti network text image generation based on attention mechanism is proposed. By introducing the multimodality of image attention mechanism, the matching loss between fine-grained image and word level text can be calculated stably, so as to better optimize the training process of the generator. The test on cub bird dataset shows that the network can gradually generate realistic images with more fine details and has better application value.

Keywords: Attention mechanism; Countermeasure network; Text image generation; Multimodal similarity model.

1. Introduction

Text description is a difficult task to generate high-resolution and real images. It is helpful for computer vision system to obtain more in-depth visual understanding, and the method of generating realistic images has a wide range of application scenarios. When it is integrated into the automatic synthesis system, it can support photo editing and computer-aided design to assist the work of artists or graphic designers.

In recent years, the research on text to image synthesis has made exciting progress. Researchers combine recurrent neural network (RNN) with generative adversarial networks (GAN) [1,2] to create realistic images according to natural language description. These methods have been able to give surprising results in some fields, such as producing exquisite images of flowers or birds.

The proposal of attngan [3,4] makes people pay more attention to the detail factors affecting image quality. However, this method ignores the impact of fine-grained image information on high-quality image generation.

In order to further improve the quality of the generated image, this paper proposes a generation of anti network text image based on attention mechanism. This method

can recognize the importance of each image sub region based on word level details in the image generation task, and pay attention to the generation effect of important sub regions in the image.

2. Related work

In recent years, the depth generation model has attracted extensive attention, including variational automatic encoder (VAE) [6], autoregressive method [7] and generating countermeasure network [1,8]. Compared with the other two depth generation models, the generation countermeasure network (GAN) shows good performance in generating clearer samples [9]. It is a model proposed by goodflow et al. In 2014. There are many methods to study how to better apply Gan to different fields, such as medical application [10], pixel to pixel conversion [9], image synthesis [11], etc. Reed et al. Proposed a generation of antagonistic what where network (gawwn) [13]. The results show that by adding additional conditions (such as local key points or bounding boxes of objects), the model can obtain location and content instructions to control the location of objects. However, the model only improves the image resolution to 128 * 128. Some researchers believe that the instability of training is caused by the non joint support of implicit model distribution and data distribution [14].

In order to solve the problem of instability, many schemes have been proposed. One is to propose different training techniques. Berthelot et al. Proposed a new balance execution method to balance the generator and discriminator, and combined the loss derived from Wasserstein distance to train GaN based on automatic encoder [15]. Odena et al. Added an auxiliary classifier to the output part of the discriminator to improve the performance of conditional Gan [16].

3. iAttnGANmodel

Our proposed iatngan consists of two parts: image attention generation network and discriminant network with damm model.

a. Generation network with image attention mechanism .Gan model for text to image generation usually encodes the whole sentence into a single global sentence vector as the condition of image generation. The objective function of the whole model is defined as:

$$L = L_G + L_{CA} + \lambda L_{DAMSM}，其中 L_G = \sum_{i=0}^{m-1} L_{G_i} \quad (1)$$

Among them, λ Where, is the super parameter that determines the proportion of network loss and damm loss generated in the above formula. With the help of the discriminator after training, by minimizing $\mathcal{L}_G$ The generator can be optimized to

minimize the loss of attention generation against the network, so as to jointly approximate the conditional distribution and unconditional distribution. Antagonistic loss of generator$\mathcal{L}_{G_i}$Designed as：

$$\begin{aligned}\mathcal{L}_{G_i} = & -\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log D_i(\hat{x}_i,\overline{e})] \\ & -\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log D_i(\hat{x}_i)]\end{aligned} \tag{2}$$

The first item in the above formula represents conditional loss conditional on the sentence, and the second item represents unconditional loss. Whether the generated image matches the given text description is determined by the condition loss; Whether the image is true or false is determined by unconditional loss.

The second term in formula (5)$\mathrm{L_{CA}}$It is defined as the KL divergence between the standard Gaussian distribution and the conditional Gaussian distribution (i.e. the Gaussian distribution of the training data). At the same time, in order to prevent this item from having too much or too little impact on the loss function of the whole model, it is finally regularized and obtainedNI。The mathematical formulas involved are as follows.：

$$\mathcal{L}_{CA} = NI，\text{其中}$$

$$NI_i = \frac{KL(\mathcal{N}(\mu(s),\sigma^2(s))||\mathcal{N}(0,I))_i}{\frac{1}{D}\sum_{j=1}^{D} KL(\mathcal{N}(\mu(s),\sigma^2(s))||\mathcal{N}(0,I))_j} \tag{3}$$

The third term of formula (5)$\mathcal{L}_{DAMSM}$It represents the matching loss between fine-grained image information and word level text, which will be described in detail in Section 3.3.

B discriminant network with image attention mechanism

The loss of damms can be expressed as:

$$\mathcal{L}_{DAMSM} = \alpha_1\mathcal{L}_1^w + \alpha_1\mathcal{L}_2^w + \alpha_2\mathcal{L}_1^s + \alpha_2\mathcal{L}_2^s \tag{4}$$

Among $\mathcal{L}_1^w$and $\mathcal{L}_2^w$It is a loss function calculated according to the sub region of the image and the words in the sentence, and$\mathcal{L}_1^s$and$\mathcal{L}_2^s$Global image vector$\overline{v}$Global sentence vector$\overline{e}$ Calculated。Here, weights are added to the two loss functions respectivelyα_1 and α_2。When calculating the matching loss between fine-grained image and text, during the training of sentence related loss function and words, the discriminator takes the real image and the corresponding text as a positive sample pair, while the negative sample pair includes the real image and mismatched text, and the generated image and the corresponding text. The super parameters in this section are set as follows: $\alpha_1 = 1.1$，$\alpha_2 = 0.9$，$\lambda = 5$，The learning rate is set to 0.0002. Cosine similarity and Euclidean distance are used to calculate the importance of the image in the generation network and discrimination network respectively. At the same time, the real image and text pairs are used to minimize$\mathcal{L}_{DAMSM}$To realize the pre training of damms. Since the size of the input image used to train damms is not limited, we use the real image

with the size of 299 * 299 for training. The whole training process has trained 200 cycles on a fixed data set.

4. Experiment

It is verified on the cub dataset [3]. The dataset contains 11788 images covering 200 species of birds. We use the method in reference [34] to preprocess the data set. Table 1 shows the statistics of the cub dataset.

Table 1 Statistics of cub dataset

	Train data	Test Data
DataSets	8,855	2,933
Words/Pic	10	10

In this section, we conduct quantitative and qualitative analysis of our method. First, iatngan and its variants were quantitatively evaluated (the results are shown in Figure 3 and table 2). "Iattngan1" in Table 2 represents the structure of stacking two generators, two discriminators and one attention model; "Iattngan2" refers to the structure of three stacked generators, discriminators and two attention models.

Table 2 Initial scores of different iattngan models in Cub test set

model	IS
λ=0.1, α_1=1, α_2=1	4.09±.05
λ=5, α_1=1, α_2=1	4.20±.04
λ=10, α_1=1, α_2=1	4.17±.05
λ=5, α_1=1, α_2=1	4.25±.04
λ=5, α_1=1.1, α_2=0.9	4.28±.03
λ=5, α_1=1.5, α_2=0.5	4.15±.04
λ=5, α_1=1.1, α_2=0.9	4.33±.03

We generate images based on the text description on the cub test set to compare the iattngan with the previous Gans model. As shown in Table 3, the optimal is value obtained by our method is higher, 4.33, compared with the previous optimal concept score value of 4.21. Experimental results show that iatngan with image attention mechanism can generate high-resolution images with fine-grained details more effectively than previous advanced methods..

5. Conclusion

In this paper, we propose an anti network text image generation based on attention mechanism, which realizes the image attention mechanism to synthesize exquisite images based on text description. Firstly, the image attention mechanism is implemented in the attention generation network, and high-precision images can be generated through multi-stage process. Secondly, we add the image attention mechanism to the discrimination model and damsm to evaluate the matching between

fine-grained image and word level text and the authenticity of the image. Practice shows that our method is better than the previous model. In the future, we can also try to use other methods of calculating distance to explore images with higher is value.

References

[1] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," arXiv preprint arXiv: 1611. 01673, 2016.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, pp. 2672–2680, 2014.

[3] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "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.

[4] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1511–1520, 2017.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[6] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1511–1520, 2017.

[7] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "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.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[9] S. Li, M. Tang, Q. Guo, J. Lei, and J. Zhang, "Deep neural network with attention model for scene text recognition," Iet Computer Vision, vol. 11, no. 7, pp. 605–612, 2017.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134, 2017.

[11] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9242–9251, 2018.

[12] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2107–2116, 2017.

[13] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," arXiv preprint arXiv:1605.05396, 2016.

[14] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in Advances in Neural Information Processing Systems, pp. 217–225, 2016.

[15] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks[J]. arXiv preprint arXiv:1701.04862, 2017.