Project 4: AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

***Abstract*—This report is about assessing an Attentional Generative Adversarial Network (AttnGAN) method which generates images from text. This method can pay attention to the relevant words in the written natural language to synthesize details at different sub-regions of the image. Due to hardware limitations, this report just evaluates and extends the AttnDCGAN model. To evaluate this method, some text descriptions defined in the original paper are used to compare the generated images by the AttnDCGAN over the CUB dataset. For extension, different kinds of words are tested in the text description to see the behavior of the system in generating images.**

I. Introduction

One of the hot topics in multimodal learning and the interface between vision and language is the text-to-image synthesis method which is a key issue in art generation and computer-aided design. Most of the proposed methods in this area are based on Generative Adversarial Networks (GANs) and encoding the whole text into a sentence vector. The proposed method in [1], AttnGAN, has multi-stage refinement and is attention-driven. This approach has fine-grained not only at the sentence level but also at the word level leading to generating high-quality images.

AttnGAN has two components, an attentional generative network and a Deep Attentional Multimodal Similarity Model (DAMSM). The internal structure of this approach is illustrated in Figure 1.

A. Attentional Generative Network

The developed attention mechanism is used in the generator component to focus on words and draw different sub-regions relevant to those words. Thus, the text description encodes into the global sentence vector to generate a low-resolution image, and then an attention layer forms a word-context vector. A multimodal context vector is generated with the combination of the regional image vector and the corresponding word-context vector. In Figure 1, F^{ca} represents the Conditioning Augmentation converting the sentence vector to the conditioning vector. F_i^{attn} is the proposed attention model at the i^{th} stage of the AttnGAN. F^{ca}, F_i^{attn}, F_i and G_i are neural networks.

B. Deep Attentional Multimodal Similarity Model

The other component, DAMSM, computes the similarity between the generated image and the text description at the sentence and word level as well as fine-grained image-text matching loss for training the generator. The text encoder is a bi-directional Long Short-Term Memory (LSTM) [2] that extracts semantic vectors from the text description. In the bi-directional LSTM, each word corresponds to two hidden states, one for each direction. The image encoder is a Convolutional Neural Network (CNN) that maps images to semantic vectors. The inter- mediate layers of the CNN learn local features of different sub-regions of the image, while the later layers learn global features of the image.

The DAMSM loss is defined as

$$Ł_{DAMSM} = Ł_1^w + Ł_2^w + Ł_1^s + Ł_2^s \tag{1}$$

where "w" stands for "$word$" and "s" stands for "$sentence$". The loss functions of words are defined as follows:

$$\begin{aligned} Ł_1^w &= -\sum_{i=1}^{M} logP(D_i|Q_i) \\ Ł_2^w &= -\sum_{i=1}^{M} logP(Q_i|D_i) \end{aligned} \tag{2}$$

where the posterior probability of sentence D_i being matching with image Q_i is computed as

$$P(D_i|Q_i) = \frac{exp(\gamma R(Q_i, D_i))}{\sum_{j=1}^{M}(\gamma R(Q_i, D_i))} \tag{3}$$

where the attention-driven image-text matching score between the entire image (Q) and the whole text description (D) is defined as

$$R(Q, D) = log(\sum_{i=1}^{T-1} exp(\gamma_2 R(c_i, e_i))^{\frac{1}{\gamma_2}} \tag{4}$$

For computing the $Ł_1^s$ and $Ł_2^s$ the Eq. 5 should be defined as follows and then used in the equations of 2:

$$R(Q, D) = \frac{\bar{v}^T \bar{e}}{\|\bar{v}\|\|\bar{e}\|} \tag{5}$$

II. Validation

The aim of this section is to utilize the proposed network in [1] to generate high-resolution images from text descriptions. In [1], two networks are proposed for this purpose, AttnGAN and AttnDCGAN. AttnGAN can generate three images with different resolutions, 64x64x3, 128x128x3 and 256x256x3. However, the AttnDCGAN can just generate the 64x64x3 image. Here, AttnGAN is built over the CUB [3] and COCO dataset and AttnDCGAN is just built over the CUB dataset.

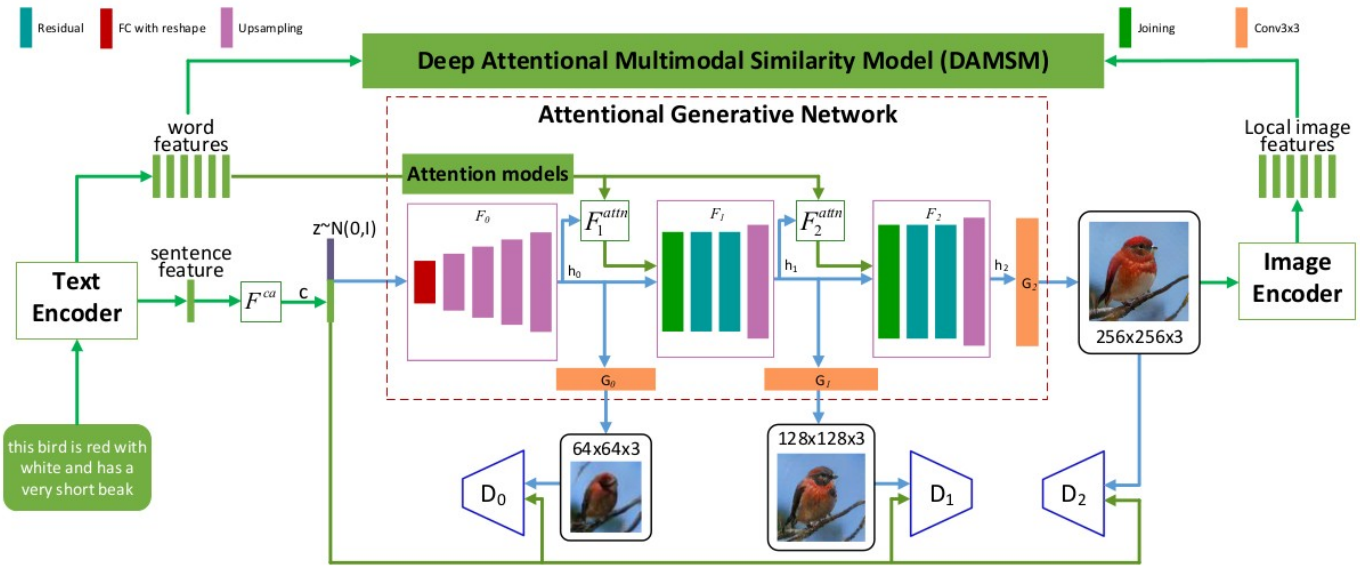


Fig. 1: The internal structure of the AttnGAN

A. Installation

The proposed networks are successfully installed in two different computers with $4GB$ 10^{st} series GPU and $6GB$ 15^{st} series GPU. By running the AttnGAN on these computers, I got the error of "CUDA out of memory". It seems that running the AttnGAN network needs at least the $8GB$ memory for the GPU. It seems that the AttnGAN is a bigger network compared to the AttnDCGAN as it can generate three different resolution images. Thus, in this report, the performance of the AttnDCGAN is just assessed and extended due to hardware limitations.

B. Generating image with AttnDCGAN

The AttnDCGAN is used to generate the image from the text description with the following config. This network is trained over the CUB dataset. For assessing the performance of AtnDCGAN, the same text descriptions in Figure 1 and Figure 4 of [1] are applied to compare the generated images (Figure 3, 4 and 5) to those images in the paper.

```
Using config:
{'B_VALIDATION': False,
 'CONFIG_NAME': 'attn2-dcgan',
 'CUDA': True,
 'DATASET_NAME': 'birds',
 'DATA_DIR': '../data/birds',
 'GAN': {'B_ATTENTION': True,
         'B_DCGAN': True,
         'CONDITION_DIM': 100,
         'DF_DIM': 64,
         'GF_DIM': 32,
         'R_NUM': 0,
         'Z_DIM': 100},
 'GPU_ID': 0,
 'RNN_TYPE': 'LSTM',
 'TEXT': {'CAPTIONS_PER_IMAGE': 10, 'EMBEDDING_DIM': 256, 'WORDS_NUM': 25},
 'TRAIN': {'BATCH_SIZE': 100,
           'B_NET_D': False,
           'DISCRIMINATOR_LR': 0.0002,
           'ENCODER_LR': 0.0002,
           'FLAG': False,
           'GENERATOR_LR': 0.0002,
           'MAX_EPOCH': 600,
           'NET_E': '../DAMSMencoders/bird/text_encoder200.pth',
           'NET_G': '../models/bird_AttnDCGAN2.pth',
           'RNN_GRAD_CLIP': 0.25,
           'SMOOTH': {'GAMMA1': 5.0,
                      'GAMMA2': 5.0,
                      'GAMMA3': 10.0,
                      'LAMBDA': 1.0},
           'SNAPSHOT_INTERVAL': 2000},
 'TREE': {'BASE_SIZE': 64, 'BRANCH_NUM': 3},
 'WORKERS': 1}
```

Fig. 2: The config values to apply the trained AttnDCGAN on the CUB dataset for generating images from text descriptions.

Fig. 3: Text description: this bird is red with white and has a very short beak

Fig. 4: Text description: the bird has a yellow crown and a black eyering that is round

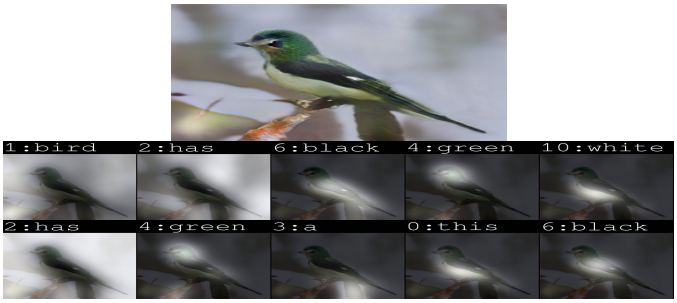


Fig. 5: Text description: this bird has a green crown black primaries and a white belly

This network provides the image under the $G0$ label and two groups of images. Each of them shows the top 5 most attended words by F_1^{attn} and F_2^{attn} of the AttnDCGAN, respectively. As shown, the generated images differ from the images in the paper but all of these images are tried to consider the main words in the sentence and are illustrated by the text description.

III. Extention

In this section, it is tried to figure out that the AttnDCGAN can detect which kind of words. I tried to insert different kinds of words, such as different adjectives into the network.

A. The effect of with vs without

Two different sentences are inserted into the network to see the effect of "with" and "without" in generating the images from the text by AttnDCGAN. As shown in Figure 7, it seems this network cannot understand the "without" word correctly. It detects this word as a "withou" instead of " without" as illustrated in Figure 7 after running different sentences with this specific word. It seems, this network has not trained for this word.

B. The effect of determining the size

Here, it is tried to determine the size of the beak to figure out if this network can distinguish among them. It seems that the network has a good performance in distinguishing between "short" and "long".

C. The effect of different shades of color

In this part, I try to determine the different shades of color to figure out if this network can correctly distinguish among them. As shown in Figure 10, the AttnDCGAN can detect the difference between the dark and bright shades of color.

Fig. 6: Text description: This is a bird with a beak

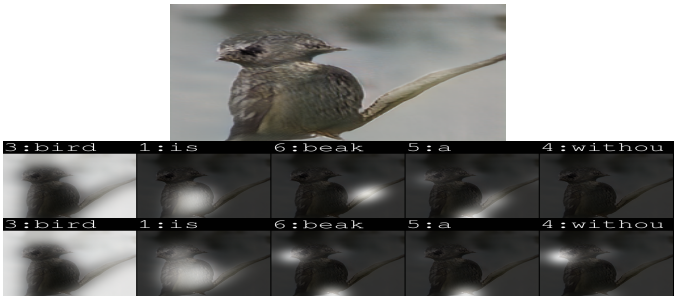


Fig. 7: Text description: This is a bird without a beak

Fig. 8: Text description: This bird is yellow and has a short beak

Fig. 9: Text description: This bird is yellow and has a long beak

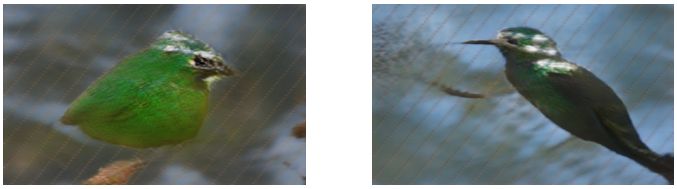


Fig. 10: Text description: a) This is a bright green bird with a white head, b) This is a dark green bird with a white head

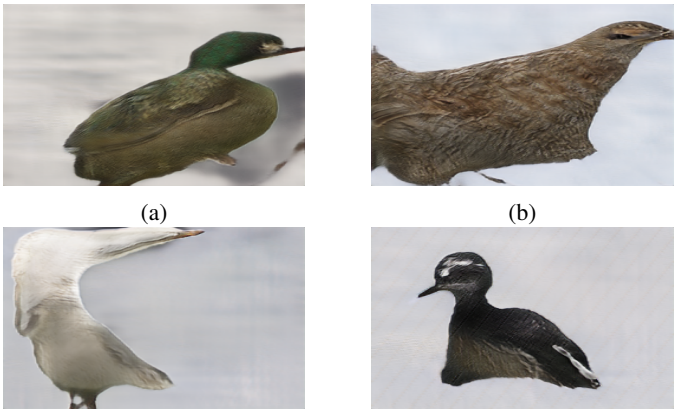
D. Determining the type of the bird

Based on my investigation on the annotation of the CUB dataset, this network should be trained to distinguish the different types of birds. It means that the network can illustrate some main features of the specific type of bird. As shown, some types of birds, such as duck and pigeon (Figure 11 (a) and (d)) are vividly illustrated in the generated images but the generated images for the gull and the owl (Figure 11 (b) and (c)) just illustrate some features of these types of birds.

E. Analysis the network can count

I changed the text description to include the number to figure out if these kinds of words are defined in this network. As shown in Figure 12, this network cannot count and this network has not trained to illustrate these words.

This part shows that the proposed network (AttnDCGAN) has good performance in the sentences and words that are already trained. It means the network can understand the words and illustrate them in the images when it has already been seen in the training phase.



(a) (b)

(c) (d)

Fig. 11: Text description: a) A duck-like bird with a green primary color, b) An owl-like bird with a brown primary color, c) A gull-like bird with a white primary color, d) A big pigeon-like bird with black and white primary colors

(a) (b)

Fig. 12: Text description: a) Two birds are yellow with red wings, b) A yellow bird with two red wings

IV. Conclusion

This paper is about showing the performance of AttnDCGAN in generating images from text descriptions. The performance of this method is evaluated by inserting some text descriptions in the original paper. The comparison shows that the generated images are different from the images shown in the paper but all of them are correct because they show what is mentioned in the text description. In the extension part, it tried to text the different kinds of words to see that the AttnDCGAN can understand these words and illustrate them in the generated images.

References

[1] 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.

[2] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Trans. Signal Processing, 45(11):2673–2681, 1997.

[3] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.