International Conference - Innovation-2021-Innovation-2021
International Journal of Scientific Research in Computer Science, Engineering and Information TechnologyISSN : 2456-3307 (www.ijsrcseit.com)
Volume 8, Issue 3, May-June-2021

Text to Image Synthesis

Chaitanya Ghadling[1], Firosh Vasudevan[1], Ruchin Dhama[1], Shreya Lad[1], Sunil Rathod[2]
[1]Student, Department Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon, Pune, Maharashtra, India
[2]Professor, Department Computer Engineering, Dr. D. Y. Patil School of Engineering, Lohegaon, Pune, Maharashtra, India

ABSTRACT

One of the most difficult things for current Artificial Intelligence and Machine Learning systems to replicate is human creativity and imagination. Humans have the ability to create mental images of objects by just visualizing and having a general look at the description of that particular object. In recent years with the evolution of GANs (Generative Adversarial Network) and its gaining popularity for being able to somewhat replicate human creativity and imagination, research on generating high quality images from text description is boosted tremendously.

Through this research paper, we are trying to explore a newly developed GAN architecture known as Attentional Generative Adversarial Network (AttnGAN) that generates plausible images of birds from detailed text descriptions with visual realism and semantic accuracy.

Keywords : GAN, AI, ML, Deep Learning, AttnGAN

I. INTRODUCTION

GAN (Generative Adversarial network):

GANs consists of two components- Generator and Discriminator which are constantly in touch with each other working in tandem. The generator generates images and the discriminator then assesses those images and provides feedback to generator about the correctness of the generated image in comparison with real images of the same object. The two neural networks constantly compete with each other to become more accurate in their predictions. The generator creates new images based on the feedback provided by the discriminator and the discriminator is trained by providing real images. The generator improves to fool the discriminator and the discriminator trains itself not to get fooled by the generator. The basic structure of GAN is shown in Fig-1.

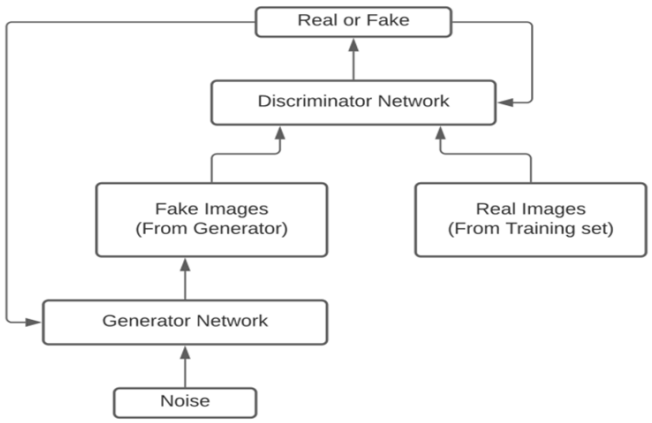

Fig-1. Basic Structure of GAN

II. LITERATURE SURVEY

In 2014, Ian Goodfellow and his colleagues designed Generative Adversarial Network with the idea of broadening scope of neural networks from just prediction and classification to allowing them to generate their own images.

Though originally proposed as a form of generative model for unsupervised learning, GANs have also proven useful for semi-supervised learning, fully supervised learning and reinforcement learning. After various architectures developed to generate images by providing text description the quality of images along with semantic accuracy can be discussed from the Table-1.

Table-1 : Literature Survey

Sr. No.	Paper Name	Advantages	Limitations
1.	Generative Adversarial Text to Image synthesis	1st major model for text to image synthesis	Lacks image quality.Does not work properly with different variety of datasets
2.	StackGAN++: Realistic Image synthesis with Stacked Generative Adversarial Networks	Improves the quality of image substantially	Difficult to train.Highly unstable and sensitive to hyper parameters.
3.	MirrorGAN: Learning Text to Image Generation by Redescription	Semantic consistency of image is highly improved.	Modules are not jointly optimized with complete end-to-end training.
4.	Learn, Imagine and Create: Text to Image Generation from prior knowledge.	Both visual realism and semantic accuracy is highly improved over baseline models.	Modules are not jointly optimized with complete end-to-end training.

III. TAXONOMY CHART

The two main attributes that the performance of text to image converting GANs are

i. Image Quality- How real the image drawn looks.

ii. Semantic Accuracy- How accurate the image is with respect to the given text description.

To have a quantitative evaluating metric to measure the performance, we have used Inception Score for two datasets namely COCO and CUB.

Table 2 contains the comparison on how different GAN architecture performed on given parameters.

Inception Scores of all the models is taken from their respective papers.[2],[3], [4], [5],[6].

Table-2: Taxonomy Chart

Attributes Model	Image Quality	Semantic Accuracy	Inception Score (COCO dataset)	Inception Score (CUB dataset)
DC GAN	LOW	LOW	8.20	3.6
STACK GAN	MEDIUM	LOW	8.45	3.7
STACK GAN++	HIGH	MEDIUM	8.30	3.82
MIRROR GAN	MEDIUM	HIGH	26.47	4.56
LEICA GAN	MEDIUM	MEDIUM	20.42	4.62

IV. PROPOSED METHODOLOGY

The GAN model used in the proposed system is called as Attentional Generative Adversarial Network(AttnGAN). The architecture of the model is shown in Figure- 2. This model has two major components:

i. Attentional Generative Network.

ii. Deep Attentional Multimodal Similarity Model.

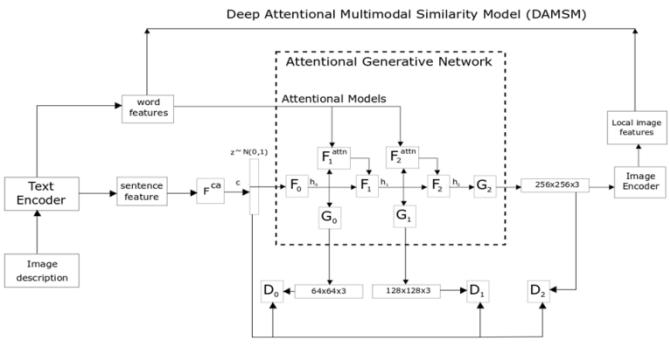

Figure -2. Proposed System Architecture

i. Attentional Generative Network.

Earlier Models for Text to Image Synthesis, typically encoded the entire text description into a single vector as condition for image creation. This enables us to generate various sub-regions of image conditioned on text that are relevant to those sub-regions. The proposed attentional generative network has m generators G0, G1….,Gm-1which take the hidden states h0, h1….,hm-1as input and generate images of small-to-large scales x0, x1….,xm-1.

Specifically,

$$h_0 = F_0(z, F^{ca}(\bar{e}));$$
$$h_i = F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \text{ for } i = 1, 2, ..., m-1;$$
$$\hat{x}_i = G_i(h_i).$$

(Equation 1)

Here, z is a noise vector usually sampled from a standard normal distribution. ‘ē’ is a global sentence vector, and ‘e’ is the matrix of word vectors. Fca represents the Conditioning Augmentation that converts the sentence vector e to the conditioning vector. Fi attn is the proposed attention model at the I th stage of the AttnGAN. Fca, Fi attn, Fi, and Gi are modeled as neural networks. The attention model Fattn(e,h) has two inputs: the word features e ϵ RDxTand the image features from the previous hidden layer h ϵ RDxT. The word features are first converted into the common semantic space of the image features by adding a new perceptron layer, i.e., e′= Ue, where U ϵ RDxD. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h (query). Each column of h is a feature vector of a sub-region of the image. For the jth sub-region, its word context vector is a dynamic representation of word vectors relevant to hj , which is calculated by vector is a dynamic representation of word vectors relevant to hj , which is calculated by

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e'_i, \quad \text{where } \beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})},$$

(Equation 2)

S'I,j = hTje'I and $\beta_{j,i}$ indicates the weight the model attendsto the ith word when generating the jth sub-region of theimage. We then donate the word-context matrix for imagefeature set h by
$F^{attn}(e,h) = (c_0, c_1, ..., c_{N-1}) \in \mathbb{R}^{\hat{D}\times N}$.

Attentional Model:

To generate realistic images with multiple levels (i.e., sentence level and word level) of conditions, the final objective function of the attentional generative network is defined as

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \text{ where } \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}.$$

(Equation 3)

Here, λ is a hyperparameter to balance the two terms of the above equation. The first term is the GAN loss that jointly approximates conditional and unconditional distributions.

Generator Model:

At the i th stage of the AttnGAN, the generator Gi has a corresponding discriminator Di . The adversarial loss for Gi is defined as

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log(D_i(\hat{x}_i)]}_{\text{unconditional loss}} \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log(D_i(\hat{x}_i,\bar{e})]}_{\text{conditional loss}},$$

(Equation 4)

where the unconditional loss determines whether the image is real or fake while the conditional loss determines whether the image and the sentence match or not.

Discriminator Model:

Alternately to the training of Gi, each discriminator Di is trained to classify the input into the class of real or fake by minimizing the cross-entropy loss defined by

$$\mathcal{L}_{D_i} = \underbrace{-\frac{1}{2}\mathbb{E}_{x_i\sim p_{data_i}}[\log D_i(x_i)] - \frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log(1-D_i(\hat{x}_i)]+}_{\text{unconditional loss}}$$
$$\underbrace{-\frac{1}{2}\mathbb{E}_{x_i\sim p_{data_i}}[\log D_i(x_i,\bar{e})] - \frac{1}{2}\mathbb{E}_{\hat{x}_i\sim p_{G_i}}[\log(1-D_i(\hat{x}_i,\bar{e})]}_{\text{conditional loss}},$$

(Equation 5)

where xi is from the true image distribution pdatai at the i th scale, and xˆi is from the model distribution pGi at the same scale. Discriminators of the AttnGAN are structurally disjoint, so they can be trained in parallel and each of them focuses on a single image scale.

ii. Deep Attentional multimodal similarity model

The attention-driven image-text matching score is designed to measure the matching of an image-sentence pair based on an attention model between the image and the text. We first calculate the similarity matrix for all possible pairs of words in the sentence and sub-regions in the image by:

$$s = e^T v,$$

where s 2 RT_289 and si;j is the dot-product similarity between the ith word of the sentence and the jth sub-region of the image. We find that it is beneficial to normalize the similarity matrix as follows

$$\bar{s}_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=0}^{T-1}\exp(s_{k,j})}.$$

(Equation 6)

Then, we build an attention model to compute a region context vector for each word (query). The region-context vector ci is a dynamic representation of the image's sub-regions related to the ith word of the sentence. It is computed as the weighted sum over all regional visual vectors, i.e.

$$c_i = \sum_{j=0}^{288}\alpha_j v_j, \text{ where } \alpha_j = \frac{\exp(\gamma_1\bar{s}_{i,j})}{\sum_{k=0}^{288}\exp(\gamma_1\bar{s}_{i,k})}.$$

(Equation 7)

Here, γ1 is a factor that determines how much attention is paid to features of its relevant sub-regions when computing the region-context vector for a word. Finally, we define the relevance between the ith word and the image using the cosine similarity between ci and ei,

$$i.e., R(c_i, e_i) = (c_i^T e_i)/(||c_i||||e_i||).$$

Inspired by the minimum classification error formulation in speech recognition the attention-driven image-text matching score between the entire image (Q) and the whole text description (D) is defined as

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

(Equation 8)

where γ2 is a factor that determines how much to magnify the importance of the most relevant word-to-region context pair. When $\gamma_2 \to \infty$, $R(Q, D)$ approximates to $\max_{i=1}^{T-1} R(c_i, e_i)$.

The DAMSM loss is designed to learn the attention model in a semi-supervised manner, in which the only supervision is the matching between entire images and whole sentences (a sequence of words). For a batch of image-sentence pairs $\{(Q_i, D_i)\}_{i=1}^{M}$, the posterior probability of sentence Di being matching with image Qi is computed as

$$P(D_i|Q_i) = \frac{\exp(\gamma_3 R(Q_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma_3 R(Q_i, D_j))},$$

(Equation 9)

V. RESULT AND DISCUSSION

Following the Zhang et al. [6], we have used Inception Score [7] as the quantitative evaluation measure. Also we have used R-precision, a common evaluation metric as a complementary evaluation measure for the text to image synthesis. The main feature that distinguishes our model from the pre-existing models is the presence of DAMSM which improves the performance of the model.

To test the proposed ŁDAMSM

We adjust the value of λ (See equation 3). From Table we can see as the value of λ increases both inception score and R-precision increases substantially.

Method (AttnGAN)	Inception Score	R-Precision (%)
No DAMSM	3.92 + 0.03	11.26 + 6.20
λ= 0.1	4.21 + 0.05	18.62 + 4.05
λ= 1	4.28 + 0.04	30.28 + 3.28
λ= 5	4.36 + 0.04	55.76 + 5.69
λ= 10	4.30 + 0.05	60.81 +3.44

Figure 3 and 4 shows us immediate results of the CUB dataset as images of 64 x 64 generated by G0, 128 x 128 generated by G1, 256 x 256 generated by G2 of the AttnGAN.

Figure 3: Text: A yellow bird with long beak

Figure 4: Text: A red bird with long beak

To elaborate the results the first stage of AttnGAN (G0) generates the skeleton of the object in low resolution. Since only single vector input is utilized here, word level detail is generally missing. These mistakes are later rectified during next stages of high resolution image generation by G1 and G2. As seen in Figure 3 and Figure 4, words like the, this, bird are generally handled by the Fattn model for locating the object. The initial image of 64 x 64 resolution does not

give attention to colour or shape of the bird's attributes. However in 128 x 128 image we can see that the image starts to give more attention to words like 'short beak', 'black crown', 'red wings' and'blue wings' along with better quality of image. In the end the image generated contains all the word level features we described in the text with a resolution of 256 x 256.

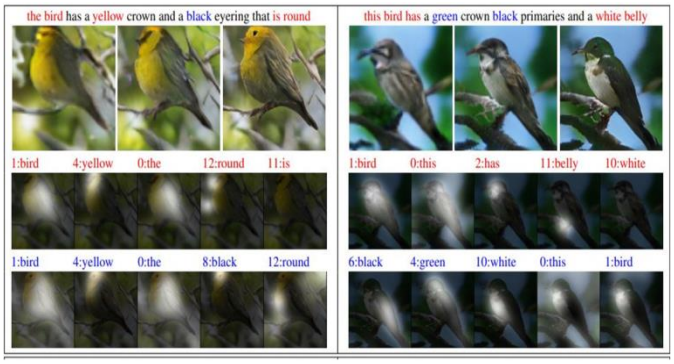

Figure 5. Intermediate results of our AttnGAN on CUB test sets. In each block, the first row gives 64×64 images by G0, 128×128 images by G1 and 256×256 images by G2 of the AttnGAN; the second and third row shows the top-5 most attended words by F attn 1 and F attn 2 of the AttnGAN, respectively.

VI.CONCLUSION

The performance of Attentional Generative Attentional Model is greatly improved over the existing models that deal with generation of image from text description. The quality of image can be improved even more by adding more feature generator. But due to the limitation on memory and computing capabilities we limit the resolution quality to 256 x 256. But in future if the hardware performance improves we can go ahead for even better resolution. Increase in inception. score by over 12% from other models on the CUB dataset shows the effectiveness of the model. Exhaustive experimentation can greatly demonstrate the ability of the proposed model in handling complex scenes having various word level detail that needs to be drawn on the image.

VII.

ACKNOWLEDGEMENT

It gives us a great pleasure in presenting the paper on "TEXT TO IMAGE SYNTHESIS". We are really grateful to Dr Sunil Rathod for giving an opportunity to work with R&D cell of our department and providing us with necessary guidance with our project. We would like to take this opportunity to thank Dr. Pankaj Agarkar, Head of Computer Engineering Department, DYPSOE, Pune for providing us with an opportunity to present this paper. Our special thanks to Dr. Ashok Kasnale, Principal DYPSOE who motivated us and created a healthy environment for us to learn in the best possible way. We also thank all the staff members of our college for their support and guidance.

VII. REFERENCES

[1]. AttnGAN: Fine - grained Text to Image Generation with Attentional Generative Adversarial Networks.

[2]. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks.

[3]. Generative Adversarial Text to Image Synthesis.

[4]. MirrorGAN: Learning Text to Image Generation by Redescription.

[5]. Learn, Imagine, and Create: Text to Image Generation from Prior Knowledge.

[6]. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017. 1, 2, 3, 5, 7

[7]. A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. VQA: visual question answering. IJCV, 123(1):4–31, 2017.

[8]. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014.

[9]. E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In NIPS, 2015.

[10]. T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In NIPS, 2016. 2, 5.