Text To Image Using Stack-Gan

*

Lakshmi G Pillai
dept. Computer Science
Artificial Intelligence
lakshmigpillai@am.students.amrita.edu

Nithin Sylesh
dept. Computer Science
Artificial Intelligence
nithinsylesh@am.students.amrita.edu

Ritika R Prasad
dept. Computer Science
Artificial Intelligence
nithinsylesh@am.students.amrita.edu

Vysakh S Nair
dept. Computer Science
Artificial Intelligence
vysakhsnair@am.students.amrita.edu

***Abstract*—Computing high-quality images from text descriptions is a challenging problem in computer vision, but it has many applications. Currently available text-to-picture algorithms produce images that largely match the meaning of given descriptions, but lack the necessary details and vibrant object elements. We introduce a StackGan for creating stunning graphics. GAN can generate delivered high pictures of 256x256 pixel based on the text specifications in the paper. Through a sketch-refinement process, we are able to break up a dif icult problem into smaller parts. Based on the given text description, Stage-I GAN draws the primitive shape and colours of the object, creating Stage-I low-resolution pictures. Relying on stage-I results and description, GAN Stage-II produces significant images with photo-realistic features. A Level II GAN takes the results of a Stage-I GAN as input and creates 256x256 high-resolution pictures with set of pictures features.. The refinement process can correct faults in Stage-I results and address com pelling details. In this paper, we provide an unique ConditioningAugmentation approach that uses roughness in the hidden conditioning manifold to promote picture variety and sustain the conditional-training GAN's**

***Index Terms*—Stackgan,Generator,Discriminator**

I. Introduction

There are many applications for creating photo-realistic images from text, including photo editing, computer-aided design, and so on. Recently, Generative Adversarial Networks (GANs) have shown promise for creating real-world images. Conditional GANs are able to create high-resolution photorealistic images from text descriptions that are highly connected to those meanings. However, it is challenging to teach GANs how to generate such images based on text descriptions. GAN models that generate high-resolution images (e.g., 256256) are unstable during training, causing nonsensical outputs when additional upsampling layers are added. The fundamental challenge in using GANs to generate high-resolution images is that in high-dimensional pixel space, the supports of natural image distribution and suggested model distribution may not overlap. This problem becomes more pronounced as the image resolution increases. In comparison to how human artists draw paintings, we are here decomposing the problem of converting text to a photo realistic image into two tractable sub-problems using StackGAN. Our Stage-I GAN generates low-resolution images first. We stack Stage-II GAN on top of our Stage-I GAN to create realistic high-resolution (e.g., 256256) images based on Stage-I findings and text descriptions. Stage-II GAN learns to capture the text information that Stage-I GAN omits and draws extra details for the object by conditioning on the Stage-I result and the text again. The model distribution support created from a roughly aligned low-resolution image has a higher chance of intersecting with the image distribution support. Furthermore,the restricted number of training text-image pairs for the text-to-image generation task frequently results in sparsity in the text conditioning manifold, making it difficult to train GAN. To encourage smoothness in the latent conditioning manifold, we suggest a unique Conditioning Augmentation approach. It increases the diversity of synthesised images by allowing modest random disturbances in the conditioning manifold. There are three main attributes for the proposed model: (1) For synthesising photo-realistic images from text descriptions, we offer a unique Stacked Generative Adversarial Networks. It breaks down the challenging task of creating high-resolution photographs into smaller, more manageable chunks, considerably improving the state of the art. For the first time, StackGAN creates 256x256 quality images with photorealistic details from text descriptions. (2) A new Conditioning Augmentation strategy is proposed to boost the diversity of the generated samples while stabilising conditional GAN training. (3) Extensive qualitative and quantitative trials indicate the overall model's performance as well as the influence of individual components, providing useful knowledge for future conditional GAN model creation

II. Related Works

1) StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks To generate photo-realistic images conditioned on text descriptions, this research presents stacked Generative

Adversarial Networks (StackGAN). Stage-I GAN creates low-resolution Stage-I pictures by sketching the object's rudimentary shape and basic colours based on the given text description. Stage-II GANs create high-resolution images with lifelike detail using Stage-I results and text descriptions as inputs

2) Text to Image Generation with Semantic-Spatial Aware GAN They suggested a unique framework for Semantic-Spatial Aware GANs for T2I generation in this paper. The system is trained from start to finish and only has one set of generators and discriminators. As part of the SSA-GAN framework, a Semantic Spatial Aware Convolution Network (SSACN) block operates Semantic-Spatial Condition Batch Normalization by predicting mask maps based on the current image features, and by learning the affine parameters from the encoded text vector. A block called SSACN deepens the text-image fusion through the image-generation process and ensures that the text-image consistency is maintained
3) Generative Adversarial Text to Image Synthesis They demonstrated a straightforward and efficient way for creating graphics from precise visual descriptions. The model was able to synthesize a large number of plausible interpretations based on a given caption. On CUB, the conditional GAN manifold interpolation regularizer dramatically enhanced image to text conversion. From query images, they showed the disentanglement of style and content, as well as the transfer of bird poses and backgrounds to text descriptions. Their results on the MS-COCO dataset showed that their strategy for creating images with many objects and changing backgrounds is generalizable.
4) Text to Image Generation with Segmentation Attention For text-to-image synthesis, they propose SegAttnGAN, which uses segmentation attention to constrain the training process of the GAN and is capable of producing images that are of higher quality compared to other methods. The SegAttnGAN achieved the highest Inception Scores on both the CUB and Oxford-102 datasets that we used segmentation masks as inputs. In comparison to other state-of-the-art methods, our self-attention SegAttnGAN also generates more realistic results when masks are generated by our self-attention generator.

III. Methodology

1) Embedding and Conditional Augmentation The generator's goal is to minimise the discriminator classification error.Similarly, the discriminator's goal is to minimise classification error.Initially, the textual description is converted in the embedding and conditional augmentation section.Text embedding to vector representation of fixed length.Later on, concatenate embedding with the noise vector and feeding the final result to the generator The image is then generated.According to the text description, generates an image based on the bias of the text embedding text.An FC layer with LeakyReLU nonlinearity was used to pass the text embedding output. The text conditioning variable after processing.
2) Stage-I GAN Rather than directly creating a high-resolution image conditioned on the text description, we first generate a low-resolution image with our Stage-I GAN, which focuses on drawing only the rough outline and correct colours for the object. In the Stage-I GAN , to generate values for the Gaussian distribution, the text embedding is first fed into a fully connected generator. In order to obtain the low-resolution image, a series of blocks of up-sampling are employed to compute the noise vector. Similar to the discriminator in Stage-1 GAN, the embeddings are first compressed using a fully connected layer, then spatially copied to generate a three-dimensional tensor. The image produced by the generator is sent through a series of down-sampling blocks until it is reduced to a two-dimensional spatial dimension. Concatenating the image filter map with the text tensor along the channel dimension achieves this. To concurrently learn picture and text features, the tensor is put into a 1x1 convolutional layer. A fully linked layer with one node is used to generate the decision score
3) Stage-II GAN Low-resolution Stage-I GAN pictures may lack bright object elements and have shape distortions. Some text features may be removed in the initial step, which is necessary for making photo-realistic graphics. Based on the Stage-I GAN results, our Stage-II GAN generates high-resolution images. It is conditioned on low-resolution photographs and text embedding to correct flaws in Stage-I outcomes. The Stage-II GAN fills in missing text information, resulting in more photo-realistic details. Our Stage-II generator is designed as an encoder-decoder network with residual blocks that are designed to learn multi-modal representations across image and text features. Lastly, the high resolution image is generated with a series of up-sampling layers. In this instance, the generator may fix the flaws in the input image while also adding extra details for a more realistic high quality image. There are only a few extra blocks in the discriminator structure compared to Stage-I discriminator due to the larger image size at this stage.To directly encourage GAN to develop good orientation between the image and the conditioning text, we utilise trying to match differentiation instead of the plain discriminator at all steps.The discriminator employs genuine photos and their text descriptions as positive sample pairs during training, whereas negative sample pairs are made up of two groups. In the first case, the text embeddings on real images are mismatched, but in the second case, the text embeddings on synthetic images are matched.

IV. Dataset

CUB includes 200 bird species and 11,788 pictures. Because 80 percent of the species in this sample possess image scale proportions of less than 0.5, we resize all photos as a

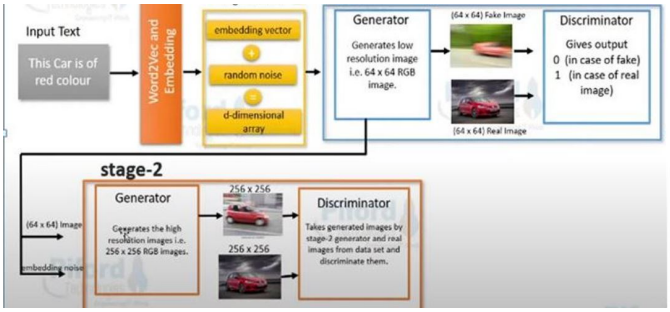

Fig. 1. StackGAN

pre-processing phase to guarantee that bird bounding boxes possess element size ratios larger than 0.75. Oxford-102 includes 8,189 flower photos from 102 separate types. MS COCO, a more difficult dataset, is also used for assessment to demonstrate the generalisation capabilities of our technique.

V. Performance Evaluation

The accuracy we obtained was comparatively less due to the lack of resources.We could only perform the loop for 30 epochs.We successfully implemented the project text to image using a generative adversarial network in this project. Depending on the written explanation supplied, we were able to produce photos. We then calculated computational accuracy as well as discriminator and generator inefficiencies. In this project, we implemented our project using two stack gan, with the first gan producing a low resolution image and the second gan producing a high resolution image.

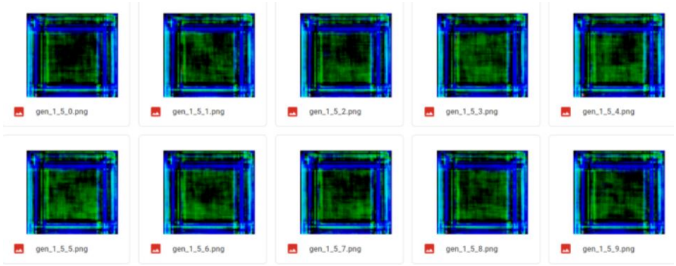

Fig. 2. Result obtained during Epoch-5

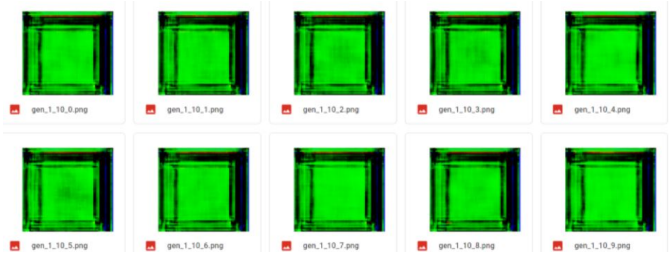

Fig. 3. Result obtained during Epoch-10

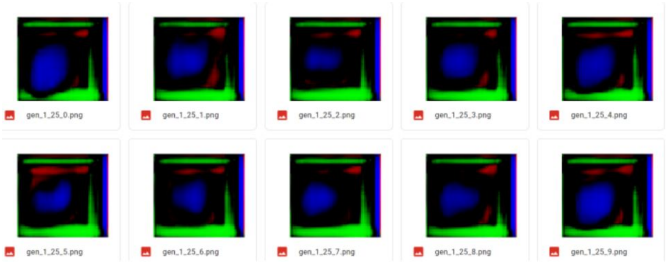

Fig. 4. Result obtained during Epoch-15

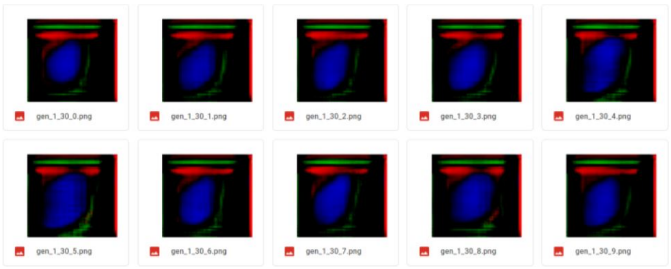

Fig. 5. Result obtained during Epoch-25

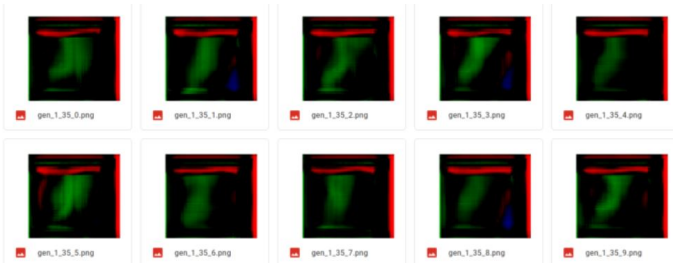

Fig. 6. Result obtained during Epoch-30

References

[1] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv:1411.1784, 2014.

[2] https://arxiv.org/pdf/1612.03242v2.pdf.

[3] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov. Generating images from captions with attention. In ICLR, 2016.

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[5] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In ICML, 2016

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017