Text to Image Generation

Abhishek Rajgaria 2017276
Preyansh Rastogi 2017176
Tanish Jain 2017115

1 Introduction

Generating images from text description is an important problem for computer aided design and art/image generation, it also drives research in multi modal learning direction and inference in both language and vision. As a human we try to draw the image unconsciously for various words which helps in understanding and learning, though a very challenging problem for computers. There has been researches [1] [2] which have tried to produce analogy between the generation of image to that drawing the image by a real painter. This has produced great insights for the architecture used for this problem.

Problem Statement: Generating realistic images from their text descriptions. The image generated should be consistent with the text that was input. The problem can be divided into two main sub problems: finding a meaningful visual representation for the text description and using this representation to generate the overall image.

2 Literature Review

In the past decade, great progress has been achieved in this field with the emergence and advancement of Deep Generative Models. In DC GAN [3], a more stable GAN with Convolutional network is proposed which emphasise on using strided convolution so that the model learn its own spatial down sampling. Eliminating Fully connected layer with global pooling which increase stability at the cost of convergence speed and finally batch normalization which stabilizes learning by normalizing the input. In [4] GAN-CLS was proposed, which was among the first architectures which provided promising results for text to image synthesis. They computed the textual representaions from the char-CNN-RNN encoder. It uses a deep convolutional architecture for both the generator and the discriminator, similar to DC-GAN.

The previous researches did not address the network architectures which could be analogous to that human drawing behaviour StackGAN [1] uses two GANs and build their architecture on this notion. The Stage I GAN generates images from captions at a lower resolution. The Stage II GAN has a generator which takes as input the image generated by the Stage I generator and produces a higher resolution image with more fine-grained details. Later [2], provided an improvised version StackGAN++, it uses multiple generators and discriminators in a tree-like structure. At each branch of the tree, "the generator captures the image distribution at that scale and the discriminator estimates the probability that a sample came from training images of that scale rather than the generator". Thus, generating high-resolution images from low-resolution images.

Many GANs focus on global sentence vector, missing the important word level information for image generation. In AttnGAN [5], an attention mechanism is developed which performs multilevel conditioning(word level and sentence level). For each sub-region of image, attention weights are learned to

emphasis the importance of the word w.r.t image sub-region, thus obtaining more fine-grained image.

DFGAN [6] proposes a novel simplified text-image backbone that directly generates realistic images using one pair of generator and discriminator, a novel Deep text-image Fusion Block that fuses the text and image features more effectively and Matching-Aware zero-centred Gradient Penalty that significantly improves text-image consistency. It also employs a one-way discriminator that helps the generator converge faster.

3 Baselines

Dataset: We have used the CUBDataset [7]. It has 11,788 images of birds belonging to 200 different categories. It also includes 15 locations of birds' body parts, 312 binary attributes and a bounding box for the bird in each image. For each Images 10 caption is also available.

3.1 Baseline 1 : DC-GAN with Sentence condition

DC GAN [3] focuses on learning image representation and generation from normal distribution commonly called "noise". We concatenated sentence level embedding to noise in order to add text/sentence condition on DC GAN. Generator uses three Up sampling while contracting feature and enlarging image, with 2 Linear transformation layer for transforming the noise and sentence embedding. Discriminator uses three down sampling layer with a global pooling in the end. Preprocessing : We have used pretrained sent2vec Bert model for text to vector embedding, each image has been resized to 32x32 for computation reasons. One major limitation of DC GAN with sentence condition is that the sentence level information is not propagated efficiently through the generator. Model is trained on 5994 images from CUB dataset.

Figure 1: Generated Image with $DFGAN(left)$ and $DCGAN(right)$

3.2 Baseline 2 : DF-GAN

In [6], the text data is encoded into a sentence vector and with a noise vector sampled from the Gaussian distribution, they are input to the generator. The noise vector is first fed into a fully connected layer and the output is reshaped to (-1,4,4). A series of UPBlocks that comprises of upsample layers, a residual block and DFBlocks, are then applied to upsample the image features that are converted into images using a convolution layer. The discriminator, composed of DownBlocks and convolution layers, converts images into feature maps and downsamples the output by a series of DownBlocks. Then the sentence vector is replicated and concatenated onto the image feature. An adversarial loss is predicted to evaluate the visual realism and semantic consistency of inputs. By distinguishing generated images from real samples, the discriminator promotes the generator to synthesize images with higher quality and text-image semantic consistency.

References

[1] Han Zhang et al. *StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks*. 2017. arXiv: 1612.03242 [cs.CV].

[2] Han Zhang et al. *StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks*. 2018. arXiv: 1710.10916 [cs.CV].

[3] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG].

[4] Scott Reed et al. *Generative Adversarial Text to Image Synthesis*. 2016. arXiv: 1605.05396 [cs.NE].

[5] Tao Xu et al. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018). DOI: 10.1109/cvpr.2018.00143.

[6] Ming Tao et al. *DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis*. 2020. arXiv: 2008.05865 [cs.CV].

[7] C. Wah et al. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011. URL: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html.