

# DS 5500: Text to Image Translation

Aishwarya Vantipuli, Aveek Choudhury, Harshita Ved

## 1. PROBLEM STATEMENT

Generating images in high quality from textual descriptions is a challenging problem in the field of natural language and computer vision. The domain has a lot of room for improvement considering the current state-of-the-art results. Developments in the field of deep learning have shown significant promise in vision and language processing tasks. Recently, Generative Adversarial Networks (GANs) have shown promising results in synthesizing real-world images. However, existing methods could generate low level images with rough relations to the description but fail to capture explicit features. Through the course of this project, we explore architectural variations of GANs to generate high quality images. Generating photo-realistic images from text has tremendous applications in space of photo-editing, computer-aided design, art generation etc.

## 2. DATASET & PHASE I RECAP

The Caltech Birds dataset (CUB) containing 11,788 bird images across 200 classes was used for this phase of work. Along with the images, the dataset provides 10 captions for each image in text format.

The dataset provides annotation of bounding boxes and a rough bird segmentation in order to identify the outline of



Figure 1: Annotated sample bird image

the bird. A set of attribute labels depicting various features of the bird is also given in the dataset. The images are cropped to ensure bounding boxes of birds have greater-than- 3/4th object-image size ratio during pre-processing.

Phase 1 dealt majorly with the initially described low-risk goals of the end-to-end pipeline for generating low-resolution images from text descriptions. The initial data processing pipeline was created which involved standardizing the images by taking the center crops using the bounding boxes around the birds and the feature attributes, ensuring the images were centered around the birds. Captions were encoded using the character CNN-RNN encoding model into fixed-length vectors and the stage-1 Generative Adversarial Network was trained successfully using the text encodings and images for 1000 epochs. The Stage 1 GAN generated low-resolution color images in 64 x 64 dimensions, showing rough boundaries of birds with some basic features like beaks and wings.

### 3. METHODOLOGY (Phase 2)

#### 3.1 Medium Risk (Success)

The medium risk goal of our project involved the generation of high-resolution images from the low-resolution images generated by the stage 1 generator. Towards this objective, we implement a second GAN (as shown in Fig. 2). The text embeddings from the caption are concatenated with the noise vector and then provided to the conditioning augmentation. Additionally, the results from the Stage 1 generator are used. The conditioning augmentation block adds randomness to the network and also makes the generator network robust by capturing various objects with different poses and appearances. It also produces more image-text pairs which makes the network more robust towards handling perturbations.

The Generator consists a series of down-sampling and upsampling layers to first generate the image features and then scale up to a high-resolution output. Then, the image features and the text conditioning variables are concatenated along the channel dimensions which is fed into the residual blocks that learn multimodal representations across image and text features. Finally, the

output of the last operation is fed into a set of upsampling layers, which generate a high-resolution image with dimensions of 256x256x3.

The Stage 2 discriminator contains extra downsampling layers as the image is of a larger size than the discriminator network in Stage-I. It is a matching-aware discriminator which allows us to achieve better alignment between the image and the conditioning text. During training, the discriminator takes real images and their corresponding text descriptions as positive sample pairs, whereas negative sample pairs consist of two groups. The first group is real images with mismatched text embeddings, while the second is synthetic.

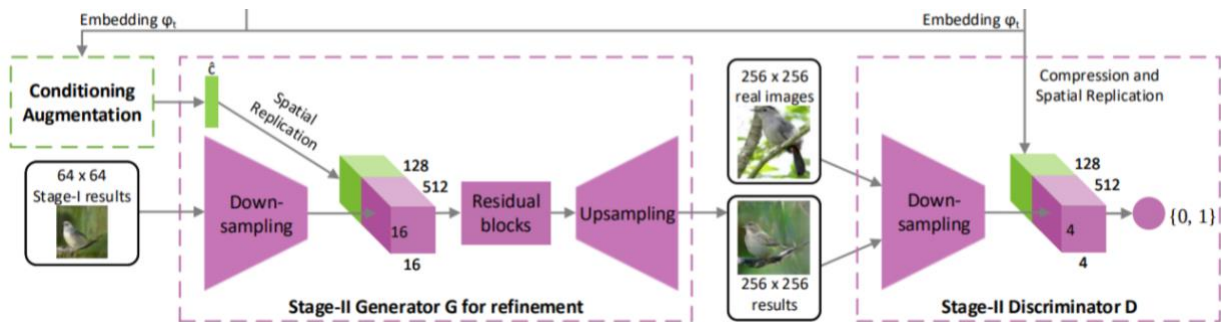


Figure 2: Network Architecture for Stage 2 GAN

### 3.2 High-Risk (Partial Success)

Our first high risk goal involved wrapping the entire end-to-end flow in an easy-to-use executable or REST-service. We were able to wrap our final generator models into a Flask API that could be requested to generate images in runtime. One challenge that we envision with models that are heavier is to store the model in-memory. This could however be achieved by having a single copy of the model weights in memory and parallelizing or balancing the requests through a load balancer setup when exposing the service to public domain.

The other high-risk task was to extend this project to other datasets comprising of images from a wide variety of classes. We tried training our network on the MS-COCO dataset, but the network

was not able to learn such varied representations. This was hence a failure for us (sample representations in Appendix).

## 4. RESULTS




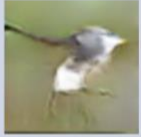





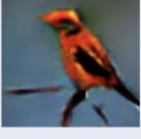






Text Description	Sample 1	Sample 2	Sample 3	Sample 4
This bird sits close to the ground with his short yellow tarsus and feet; his bill is long and is also yellow and his color is mostly white with a black crown and primary feathers				
				
The small bird has a red head with feathers that faded from red to gray from head to tail.				
				

Figure 3: Sample images generated by Stage-I and Stage-II GAN for two text descriptions

## 5. CONCLUSION

Despite the significant success of GANs in the field of vision and natural language processing, real-world problems in the text-to-image generation still pose significant challenges in - (1) the generation of high-quality images, (2) diversity of image generation, and (3) stable training. Our work explored architectures and pipelines that could be used to perform these tasks with a comprehensive understanding of the text for image generation. Through this work, we successfully established an end-to-end pipeline to generate low-resolution images from textual input and then rectify them and generate high-resolution images as final output.

## REFERENCES

- [1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. arXiv preprint arXiv:1710.10916, 2017. 4325, 4326, 4327
- [2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. 2016.4321, 4325, 4326
- [3] 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 (NIPS), 2014. 4321
- [4] Implementing GANs in Python - <https://realpython.com/generative-adversarial-networks/>

## APPENDIX

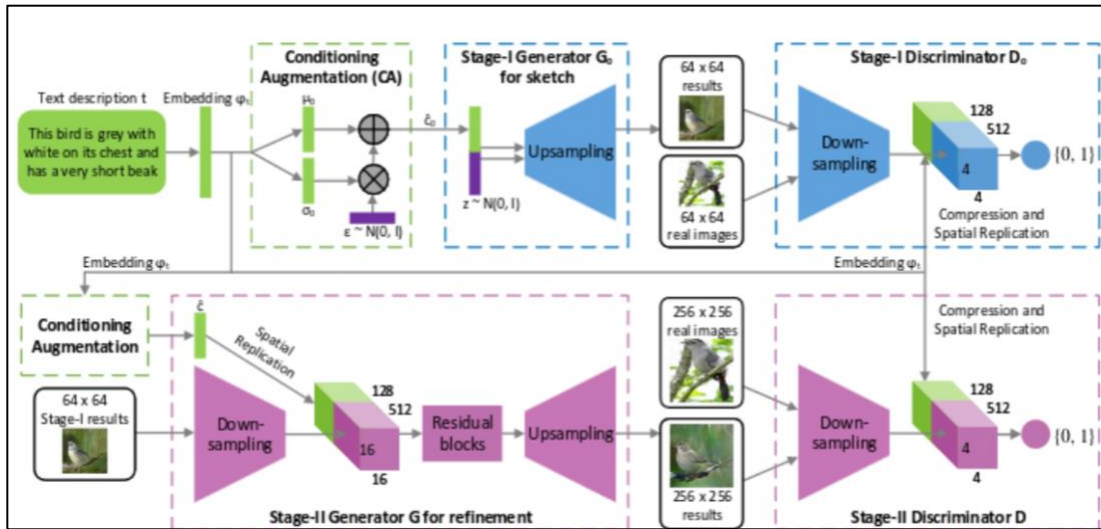


Figure 4: Overall architecture for Stage-I & II GANs

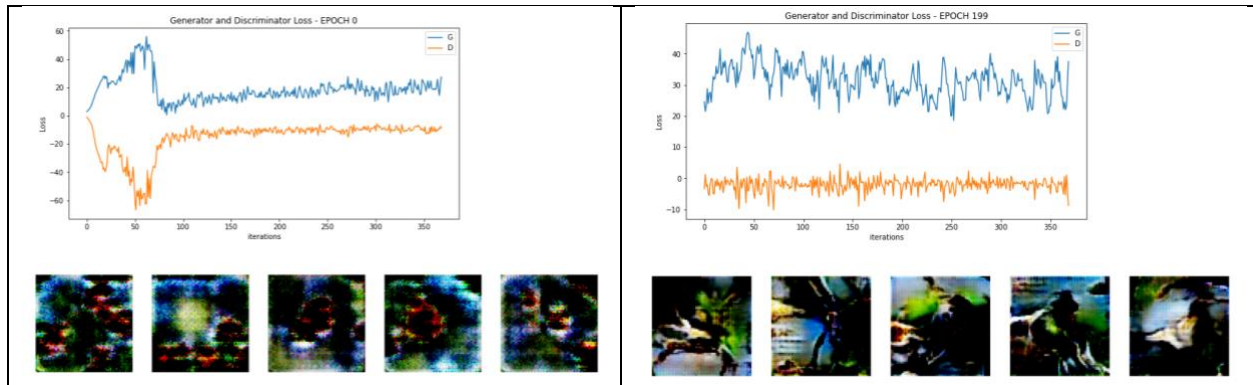


Figure 5: Comparing intermediate Generator-Discriminator loss

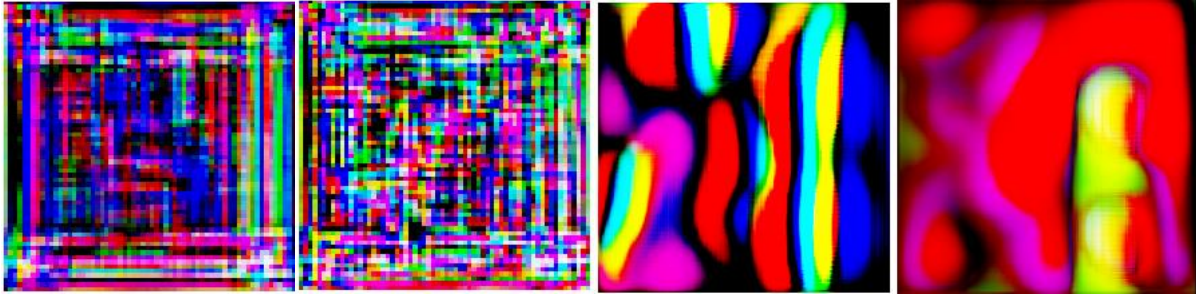


Figure 6: Sample representations generated by Stage 1 & Stage 2 Generators for MS-COCO (*FAILURE*)