DS 5500 - Capstone Project: Text to Image Translation

Aishwarya Vantipuli, Aveek Choudhury, Harshita Ved

1 Motivation

Converting natural language text descriptions into images is an amazing demonstration of Deep Learning. Text classification tasks such as sentiment analysis have been successful that are able to learn vector representations from the text. In another domain, Deep Convolutional GANs are able to synthesize images. In this work, we develop a deep GAN to effectively bridge these advances in text and image modeling by translating visual concepts from characters to pixels. Generating photo-realistic images from text has tremendous applications in photo-editing, computer-aided design, art generation etc.

1.1 Related Work

The latest progress in the study of text-to-image synthesis provides various persuasive techniques and algorithms. At first, the primary goal of text-to-image synthesis was to generate images from simple texts, and that goal later adjusted to natural languages using powerful neural network architectures like GANs (Generative Adversarial Networks). Through this project, we want to explore architectures that could help us achieve our task of generating images from given text descriptions.

1.2 Dataset

Depending on the risk levels, we are considering three different datasets: Oxford-102, CUB-200, MS-COCO. Oxford-102 is an open-source dataset consists of 8k images of flowers with 10 captions comprising of 102 categories. Caltech-UCSD Birds 200 (CUB-200) is an image dataset with photos of 200 bird species (mostly North American). Common Objects in Context (COCO) is a large-scale dataset with 330K images each containing 5 captions used for object detection and Image captioning.

2 Methodology

The main idea behind generative adversarial networks is to learn two networks- a Generator network ***G*** which tries to generate images, and a Discriminator network ***D***, which tries to distinguish between 'real' and 'fake'

generated images. We train these networks against each other in a min-max game where the ***G*** seeks to maximally fool the ***D*** while simultaneously the ***D*** seeks to detect fake examples. We aim to make use of 2 such GANs in a stacked fashion to accomplish decomposed tasks of the overall problem.

2.1 Low Risk ***Date for completion:*** *March 2nd, 2020*

The goal during this phase is to build an end-to-end pipeline to generate low resolution images from textual input. This text is first converted into an embedding vector using Word2Vec and Embeddings. These embeddings with an added random noise are given to the ***Stage-I G*** which generates a low-resolution *(64 x 64)* RGB image. This image represents low-level features of input like shape, color, segments of areas, etc. ***Stage-I D*** takes real and generated images and discriminate between them and gives an output of 0 *(fake)* and 1*(real).*

2.2 Medium risk ***Date for completion:*** *March 22nd, 2020*

In this phase, we aim to generate high resolution images from the output of the previous level GAN. ***Stage-II G*** takes the input of low-resolution image *(64 x 64)* which is generated in ***Stage-1*** and embedding noise, computes the series of convolution operations and generates the high-resolution *(256 x 256)* RGB image. ***Stage-II D*** takes generated images by ***Stage-II G*** and real images from data set and discriminate them. This could be done through a series of down sampling steps. Additionally, we would like to wrap our model into a consumable REST service.

2.3 High risk ***Date of completion:*** *April 4th, 2020*

In the final phase of work, we would like to extend our system to a much wider dataset like the MS-COCO, which would involve elevated risk due to the nature of the constituting data. In addition to that, we also look forward to building a robust platform that can consume our API service and act as an interface for end-users.

3 Impact

Despite the significant success of GANs in the field of vision and natural language processing, real-world problems in 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 is aimed at exploring architectures and pipelines that could be used to perform these tasks with comprehensive understanding of text for image generation.