Text2Image Generation:
ECE 542 Final Project Proposal

Zachary Johnston
North Carolina State University
ztjohnst@ncsu.edu

Michael Patel
North Carolina State University
mrpatel5@ncsu.edu

John Ravi
North Carolina State University
jjravi@ncsu.edu

***Abstract*—Deep learning has unlocked advances in Natural Language Processing (NLP) and Generative Adversarial Networks (GANs). GANs are still mainly used to just generate photo-realistic images. Instead, GANs can be further leveraged to perform image synthesis from text [3]. The goal of this project is to use a GAN approach to generate photo-realistic images from text-based descriptions (i.e. captions). The overall challenge is to use character inputs to produce pixel-based outputs. This end-to-end technique is called image synthesis. In this work, we propose to develop a model that builds upon GANs and Recurrent Neural Networks (RNNs) to generate plausible images of flowers from short and simple text descriptions of flowers. This work is motivated by a desire to extend GAN applications beyond just photo-realism, but also to text-to-image matching.**

I. Introduction

While image classification has been widely regarded as a success because of deep learning, image generation is still fairly new and unstudied. Imagine if by simply typing out a phrase, you could instantly generate an image that perfectly embodied that description. The impacts of such a technology would be profound. Album covers and movie posters could be quickly and cheaply made. Journalists and bloggers would no longer need to license photographs for newsworthy captions. Image synthesis can become a transformative tool in the daily lives of people.

A. Motivation

GANs are composed of two competing neural networks, the generator and the discriminator. Traditionally, the generator only takes in a vector of noise as input [1]. Instead, we look to add a second input component that is text-based encoded to account for the text descriptions. This modification allows the GAN to learn the text-to-image matching.

The purpose of this project is to perform image synthesis from input-provided text descriptions. This extends the original scope of GANs as first detailed by Goodfellow [1] by incorporating aspects of other deep learning models such as RNNs. The overall end-to-end application of this work would map character inputs to pixel-based outputs and focuses on two problems. First, information about the visuals must be learned from text data. Then, that learned information must be used to create photo-realistic images.

II. Outline of Methodology

A. Architecture

The network architecture will be the same as described by Reed [3]. The network will be a DCGAN. A high-level overview can be seen in Figure 1. The generator has two input components. An RNN will be added to the front of the generator to pick out the key words needed to learn visual information. The generator will be constructed using a series of convolutional and upsampling layers. The discriminator will be a CNN classifier using a stride of 2 in lieu of pooling layers.

B. Project Environment

Language: Python 3.6
Libraries: tensorflow-gpu, numpy, matplotlib, keras
Dataset: Oxford-102

C. Evaluation

Since GANs have competing neural networks, judging performance is relative. Some measures of performance include image quality and image distribution diversity. Image quality can be ascertained through an image classifier or through a survey approach with actual individuals (since the generated image should be easily identifiable). If we encounter GAN collapse, we will consider our implementation to have failed the image diversity criterion.

The dataset we will be using is Oxford-102, which consists of flower images and descriptions [2]. The desired outcome is to create images of flowers that are similar to Oxford-102, but by using text descriptions as input. Thereby, we will have synthesized images of flowers from matching short descriptions.

We propose two testing approaches that are qualitatively based. We will use human assessments to 'score' whether an image looks authentic or fake. This checks for image quality and overall photo-realism. Additionally, before any training, we will separate a portion of the flower dataset that will be left unused. Crucially, this includes image and text description pairs. We will then generate images using that separate subset's text descriptions. We will compare (qualitatively and also using RGB pixel values) the unused, original flower data subset images and our generated images. Ideally, both map from the same text description, so they should look similar.

To gain a more quantitative understanding of our GAN's performance, we may make use of the Frechet Inception Distance (FID) score to compute image quality and image diversity distances between the real and generated images. Perhaps there is also room in this project to determine a more useful measuring tool to judge GAN performance.

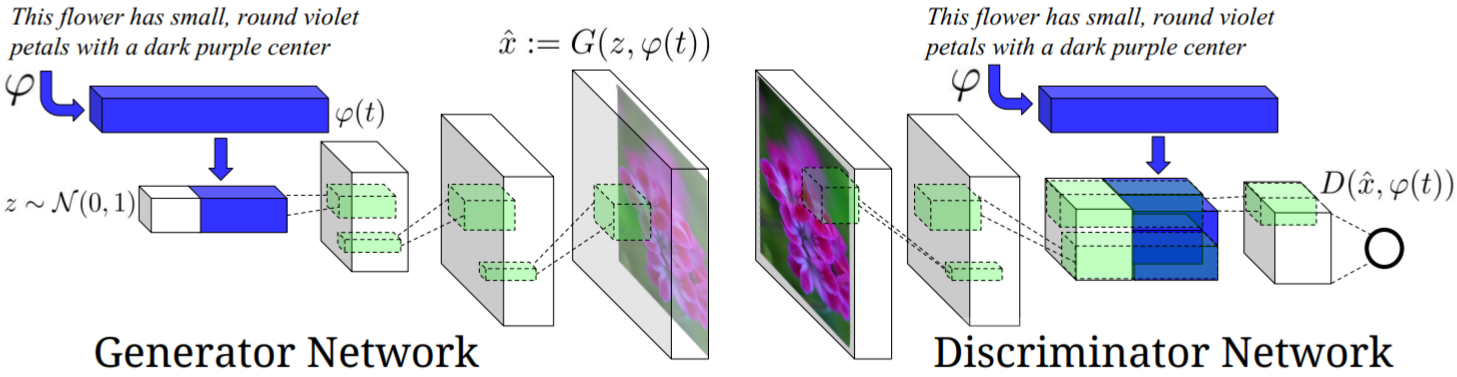
This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
Generator Network
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Discriminator Network

Fig. 1. GAN architecture from Reed [3]

References

[1] Goodfellow, Ian, et al. Generative adversarial nets. Advances in neural information processing systems. 2014.

[2] Nilsback, M-E. and Zisserman, A. Dec 2008. Automated Flower Classification over a Large Number of Classes

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48