Chapter 7
Reconnoitering Generative Deep Learning Through Image Generation From Text

Vishnu S. Pendyala
https://orcid.org/0000-0001-6494-7832
San Jose State University, USA

VigneshKumar Thangarajan
PayPal, USA

ABSTRACT

A picture is worth a thousand words goes the well-known adage. Generating images from text understandably has many uses. In this chapter, the authors explore a state-of-the-art generative deep learning method to produce synthetic images and a new better way for evaluating the same. The approach focuses on synthesizing high-resolution images with multiple objects present in an image, given the textual description of the images. The existing literature uses object pathway GAN (OP-GAN) to automatically generate images from text. The work described in this chapter attempts to improvise the discriminator network from the original implementation using OP-GAN. This eventually helps the generator network's learning rate adjustment based on the discriminator output. Finally, the trained model is evaluated using semantic object accuracy (SOA), the same metric that is used to evaluate the baseline implementation, which is better than the metrics used previously in the literature.

INTRODUCTION

The objective of this chapter is to explore generative deep learning through the specific example of generating images from text with some control over the placement of objects in the generated image. Expressing ideas in text is often much easier than doing the same in pictures or figures. Coming up with figures is an important skill that is often a formidable challenge even for human beings. In a sense, figures capture the latent semantic space of the corresponding text. Deep learning has been quite ef-

DOI: 10.4018/978-1-6684-6001-6.ch007

fective in processing natural language by capturing its latent space in the language models. With this background, we framed our research question: *to what extent can deep learning systems capture a pictorial representation of a given text and exercise control over the generated image?* We went on to search the literature to survey the existing work in this area and tried to leverage some of it as detailed in the following sections. In general, synthesizing high-quality photo-realistic images is a challenging computer vision problem and has a plethora of practical applications. Generating multiple objects in an image is an even more challenging problem, as there are high chances of missing out on the objects and overlapping of created objects in the image.

Generative Adversarial Networks (GANs) have shown major improvements and capabilities in generating photo-realistic images given an input of textual description. GANs have achieved superlative performance in synthesizing images containing a single object given the textual descriptions as input to the model. Figure 1 shows real-looking images of fictitious people that were generated using a GAN on the website, https://thispersondoesnotexist.com/. As can be seen, it is hard to tell the images are fake. The generator and the discriminator are neural networks that play a minmax game, acting as adversaries (Karras et al., 2020) to produce the astoundingly real-looking images. The generator starts with random Gaussian noise and improves in generating

Figure 1. Images of fictitious people generated on https://thispersondoesnotexist.com using a GAN

real-looking images over several iterations of feedback from the discriminator. The generator tries to make the discriminator believe that the image is that of a real person, while the discriminator acts as an adversary by denying the generator's claim as much as possible. Eventually, when the generator produces images like in Figure 1, the discriminator agrees with the generator's claim and the cycle stops. Given the outstanding performance of GANs combined with the advances in using pretrained deep learning models at text understanding, generating images conditioned on a given textual description seems feasible.

Generating images conditioned on textual description has many applications such as generating a quick visual summary for a text paragraph to enhance the learning experience for students and real-time image generation in sports for a commentary text. The images generated can depict anything from anywhere imagination can take one. For instance, images of flying lions, four-eyed tigers, and flowers that can see with eyes can all be generated by specifying the appropriate words in the text. There were attempts in the past to generate images from directed scene graphs, but they did not achieve significant results. In addition to that, the existing evaluation metrics like Inception Distance do not align with human eye