Generative Adversarial Text to Image Synthesis

Amit Manchanda | Anshul Jain | Dr. Vinod Pankajakshan

Department of Electronics and Communication Engineering, IIT Roorkee

Abstract

We implemented a deep recurrent neural network architecture and Generative Adversarial Network(GAN) formulation to effectively bridge the advances in text and image modeling, translating visual concepts from characters to pixels. We show the capability of the model to generate images of flowers from detailed text descriptions.

Introduction

- Artificial synthesis of images using text descriptions could have profound applications in visual editing, animation, and digital design.
- The distribution of images conditioned on a text description is highly multimodal.
- In GANs, the discriminator D tries to distinguish real images from syntheticized images. The generator G tries to fool D.
- The discriminator views (text, image) pairs as joint observations and is trained to judge a pair as real or fake.

Subproblems

- Learn a text feature representation that captures the important visual details.
- Use these features to synthesize a compelling image.

Literature Survey

- [1] estimated generative models via adversarial process to generate image conditioned on text and input noise.
- In [2,4] authors describe architectural guidelines for stable GANs.
- In [3] authors gave unsupervised approach to train a generic sentence encoder.

Methodology

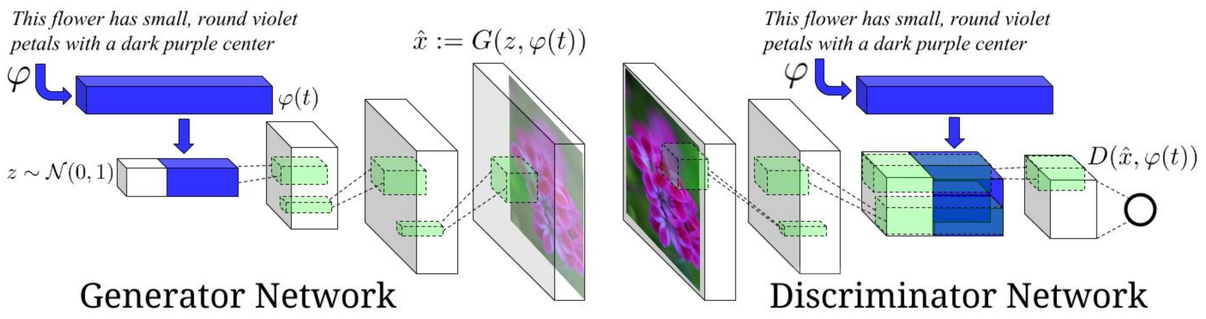

Figure 1: Text-conditional convolutional GAN architecture.

DCGAN

GAN training procedure is similar to a two-player min-max game with the following objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\ p_{data}}[logD(x)] + \mathbb{E}_{z\sim\ p_z}[log(1 - D(G(z)))]$$

where x is a real image from the true distribution, and z is a noise vector sampled from p_z, which might be a Gaussian or uniform distribution.

Skip Thought Vectors

An unsupervised approach to train a generic, distributed sentence encoder. We train an encoder-decoder model where encoder maps the input sentence to a vector and the decoder generates the surrounding sentences.

$$\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, h_i)$$

Objective is to reduce the sum of the log-probabilities for the forward and backward sentences conditioned on the encoder output.

Results

this flower has petals that are red and are bunched together

the flower has an abundance of yellow petals and brown anthers

flower is purple and pink in petal and feature a dark, dense core.

Figures generated from corresponding caption using the trained model.

Conclusion

- We developed a simple and effective model for generating images based on detailed visual descriptions.
- The images are able to capture shape and color of the flower but lacks other significant details to pass off as a realistic sample.
- The model could not generalize to images with multiple objects.

Future Works

- Improve Generator learning with manifold interpolation.
- Implementation of Stacked GANs to produce high quality images.
- Explore the possibility of using Wasserstein GANs and Cyclic GANs.
- Generalizing the model to generate images with multiple objects and variable backgrounds using MS-COCO dataset.

References

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016.

[2] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016.

[3] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In NIPS, 2015.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.