Text to Image Synthesis Using Stacked Generative Adversarial Networks

Ali Zaidi
Microsoft, Stanford University
alikazim@stanford.edu

STANFORD UNIVERSITY

Background and Introduction

Motivation. Human beings are quickly able to conjure and imagine images related to natural language descriptions. For example, when you read a story about a sunny field full of flowers, an image of a beautiful field with blossoming flowers might pop into your head. Artificial synthesis of images using text descriptions or human cues could have profound applications in visual editing, animation, and digital design.

Related Work. In order to tackle this problem, we utilize the stacked Generative Adversarial Network (S-GAN) architecture proposed by (ZXL+16). Stacked GANs improve upon an earlier DC-GAN architecture for text-to-image synthesis proposed by (RAY+16), which described how to utilize GANs to generate images by conditioning on text descriptions. However, the synthesized images from DC-GANs did not great fidelity or the ability to produce output at high-resolution. Stack-GANs attempt to improve the synthesis process by using a two-stage procedure, each of which is it's own manageable GAN implementation. The first stage, stage-I, learns to produce the rough shape and primary colors of the synthesized object conditioned on a given text description, and generates background regions from a random noise vector sampled from a prior distribution. The generated low resolution image is coarse and of lower fidelity than provided texts description. It may eve have entire objects or segments missing. However, to improve fidelity, a second GAN is stacked on top: a stage-II GAN, which aims to generate realistic high resolution images conditioned on both the low resolution image as well as the provided text description.

Dataset

The Story of the Birds and Flowers. I utilized the birds and flowers datasets from CUB and Oxford-102 respectively.

Methods and Model Architecture

For each stage, we utilize the GAN training procedure that is similar to a two-player min-max game with the following objective function:

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

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.

Moreover, in our architecture we will follow the conditional-GAN approach and additionally condition both the generator and the discriminator on additional variables, which will be the text embeddings of our descriptions, denoted by c, therefore giving us generator and discriminator $G(z,c)$ and $D(z,c)$.

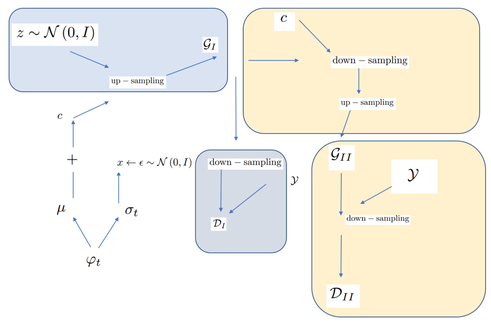
Our model architecture is shown in the figure below. In the figure $\mathcal{G}_I$ denote sthe generator from stage-I, which produces low-resolution images, and $\mathcal{G}_{II}$ is the generator from stage-II, which produces higher quality images by conditioning on the text c and $\mathcal{G}_{II}$.

Deep Structured Text Embeddings

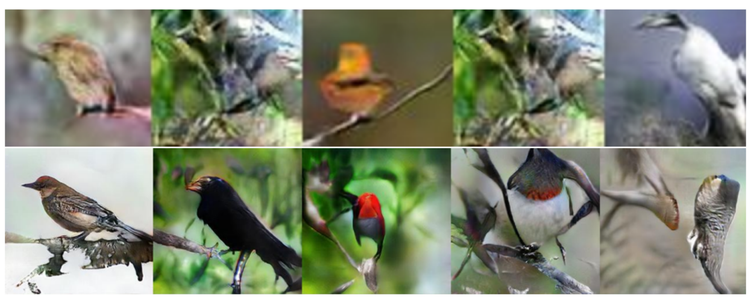
The text-embeddings we conditioned on were first pre-trained using a structured joint embedding approach. More precisely, we trained functions f_v and f_t that map image features $v \in \mathcal{V}$ and text descriptions $t \in \mathcal{T}$ to class labels $y \in \mathcal{Y}$, i.e., that minimize the empirical risk given by

$$\frac{1}{N}\sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)),$$

where $\Delta : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ is the zero-one loss derived from looking at one-hot encodings of our class labels. To make things differentiable, I used a convex surrogate rather than the discontinuous 0-1 loss.



Experimentation and Results



Experimentation and Results

The results in the bottom show the generated images from Stage I and Stage II respectively. As can hopefully be seen from the images here, the images from the first row contain the Stage I generations, which do indeed generate rough sketch of a bird, but are not very high-resolution or clean. The Stage II images, however, are much more realistic and have much higher fidelity.

Conclusion and Further Directions

In future work I'd like to try and scale to larger image-caption datasets like MSCOCO. I'd also like to try a sequential dual-training method, where we train do text-to-image synthesis in tandem with image-to-text synthesis. For multi-category datasets like MSCOCO these might perform better.

References

[RAY+16] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text-to-image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[ZXL+16] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv:1612.03242*, 2016.