

Text to Image Synthesis

Andrew Drozdov

NYU Department of Computer Science

August 8, 2017

Overview

Paper Summary: Reed et al., 2016

Generative Adversarial Text to Image Synthesis

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran
Bernt Schiele, Honglak Lee

REEDSCOT¹, AKATA², XCYAN¹, LLAJAN¹
SCHIELE², HONGLAK¹

¹ University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)

² Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)

this small bird has a pink breast and crown, and black primaries and secondaries.



this magnificent fellow is almost all black with a red crest, and white cheek patch.



the flower has petals that are bright pinkish purple with white stigma



this white and yellow flower have thin white petals and a round yellow stamen



Figure 1. Examples of generated images from text descriptions.
Left: captions are from zero-shot (held out) categories, unseen text. Right: captions are from the training set.

What is the problem?

However, one difficult remaining issue not solved by deep learning alone is that the distribution of images conditioned on a text description is highly multimodal, in the sense that there are very many plausible configurations of pixels that correctly illustrate the description. The reverse direction (image to text) also suffers from this problem but learning is made practical by the fact that the word or character sequence can be decomposed sequentially according to the chain rule; i.e. one trains the model to predict the next token conditioned on the image and all previous tokens, which is a more well-defined prediction problem.

What is the solution?

Just GAN it!

Thank you Goodfellow et al., 2014.

Generative Adversarial Networks (GANs)

Play this minimax game until reaching a Nash equilibrium:

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

Label image with a class and text description

$$\frac{1}{N} \sum_{n=1}^N \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \quad (2)$$

- ▶ v_n - image
- ▶ y_n - image class
- ▶ t_n - image description

$$f_v(v) = \arg \max_{y \in \mathcal{Y}} \mathbb{E}_{t \sim \mathcal{T}(y)} [\phi(v)^T \varphi(t)] \quad (3)$$

$$f_t(t) = \arg \max_{y \in \mathcal{Y}} \mathbb{E}_{v \sim \mathcal{V}(y)} [\phi(v)^T \varphi(t)] \quad (4)$$

- ▶ ϕ - image encoder
- ▶ ψ - text encoder
- ▶ f_v - image classifier
- ▶ f_t - text classifier

Deep Convolutional GAN (DC-GAN)

Architecture

We will cover the architecture later.

Training

Joint Conditioning of Image and Description

There might be a problem. (Gauthier, 2015)

What we have:

- ▶ Real Image + Matching Description
- ▶ Fake Image + Arbitrary Description (Matching or Not Matching)

What we may need:

- ▶ Real Image + Not Matching Description

Data Augmentation?

Problem: There are few text descriptions.

Solution: Create additional *text embeddings* (using interpolation between existing embeddings).

Assumption: “interpolations between embedding pairs tend to be near the data manifold” (Bengio et al., 2013; Reed et al., 2014)

$$\mathbb{E}_{t_1, t_2 \sim p_{data}} [\log(1 - D(G(z, \beta t_1 + (1 - \beta)t_2)))] \quad (5)$$

Style Transfer

The text encodes the content; the image encodes the style.
The text is easy to manipulate; the image not so.

Convolutional Style Network:

$$S : R^D \rightarrow R^Z$$

Style Loss:

$$\mathcal{L}_{style} = \mathbb{E}_{t,z \sim \mathcal{N}(0,1)} \|z - S(G(z, \varphi(t)))\|_2^2 \quad (6)$$

Style Transfer:

$$s \leftarrow S(x), \hat{x} \leftarrow G(s, \varphi(t))$$

Modely Summary

- ▶ GAN
- ▶ GAN-CLS
- ▶ GAN-INT
- ▶ GAN-INT-CLS
- ▶ Style Transfer

Results

Datasets

- ▶ Birds - CUB dataset
 - ▶ 11,788 images
 - ▶ 200 categories
 - ▶ 150 train+val classes
 - ▶ 50 test classes
- ▶ Flowers - Oxford-102 dataset
 - ▶ 8,189 images
 - ▶ 102 categories
 - ▶ 82 train+val classes
 - ▶ 20 test classes
- ▶ 5 captions per image
- ▶ Augment images with random transformation: scale, crop, etc.

Zero Shot: Birds



Figure 3. Zero-shot (i.e. conditioned on text from unseen test set categories) generated bird images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. We found that interpolation regularizer was needed to reliably achieve visually-plausible results.

Zero Shot: Flowers



Figure 4. Zero-shot generated flower images using GAN, GAN-CLS, GAN-INT and GAN-INT-CLS. All variants generated plausible images. Although some shapes of test categories were not seen during training (e.g. columns 3 and 4), the color information is preserved.

Discussion

Question for the Audience

What is the closest representation to an image, in pure 100% text?

Followup

- ▶ Can we do text to image synthesis without z ?
- ▶ Can we do text to image synthesis with longer descriptions?
- ▶ Can we do text to text synthesis (inverse summarization)?