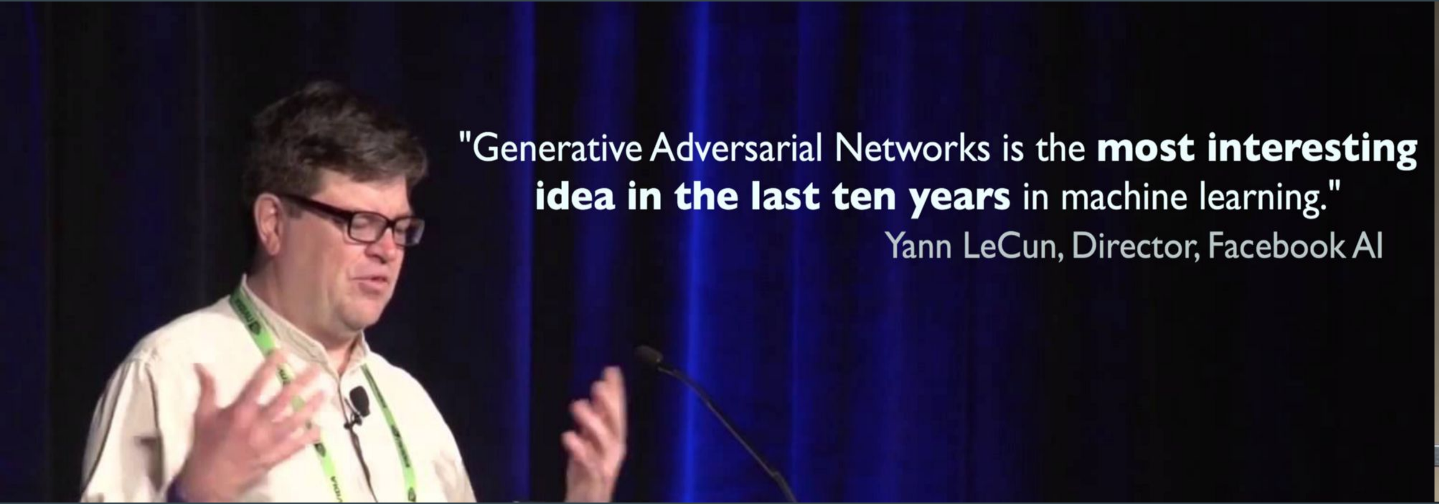
Synthesizing Novel Pairs of Image and Text: Repurposing Generative Adversarial Networks as Semantic Autoencoders

Jason Xie, Tingwen Bao

Overview

In this project, we present strategies for generating coherent image and caption pairs based on an existing captioning dataset. Our model takes advantage of recent advances in generative adversarial networks and sequence-to-sequence modeling. We were able to generate novel paired samples from multiple domains and it can further be repurposed as semantic autoencoders.



Background

- Generative Adversarial Networks (GAN)
- Text to Image Synthesis
- Image to Text Caption

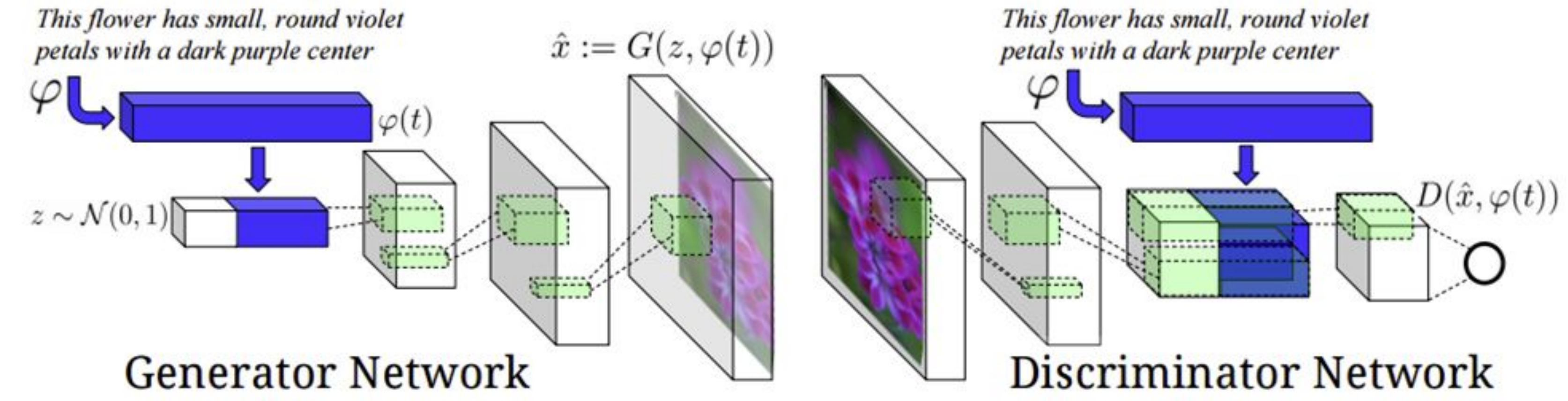


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

https://arxiv.org/pdf/1605.05396.pdf

Our Model

$$\mathbf{Image_{new}}, \mathbf{Text_{new}} \sim \mathbf{S\ (Image, Text)}\ , \text{ where } \{Image_{new}, Text_{new}\} \notin \{Image_i, Text_i\}_N$$

1. Source Domain Generation:

$$Image_{new} \sim I(Embedding\{Image_i\}) \ \ \& \ \ Text_{new} \sim T(Embedding\{Text_i\})$$

   a. Prototype Based for Text: λ* Embedding$\{i_1\}$ + (1-λ)*Embedding$\{i_2\}$
   b. Density Based for Image: Fit GMM with Image Embedding and sample from Gaussian Distribution

2. Target Domain Generation:

$$Text_{new} \sim F(Image_{new}) \ \ \& \ \ Image_{new} \sim G(Text_{new})$$

   a. Text to Image: GAN-CLS, loss function:

$$L_D \leftarrow -\log(D(x,h))+(\log(1-D(x,h_{fake}))+\log(1-D(x_{fake},h)))/2$$

   b. Image to Text: LSTM on the last convolution layer of DCGAN, maximize by the likelihood of correct caption:

$$\theta^* = \mathrm{argmax}_\theta \sum_{(Imagei,\ Texti)} \log p(Text_i \mid Image_i; \theta)$$

Our Model--Continued

- **Formulating a Circle: Image -> Text -> Image -> Text ->.........**

  $\text{Image}_{new} \sim G(\text{ Embedding}\{ F(\text{ Embedding}\{ \text{Image}_{new} \}) \})$

  $\text{Text}_{new} \sim F(\text{ Embedding}\{ G(\text{ Embedding}\{ \text{Text}_{new} \}) \})$

- **Semantic Autoencoder**

  Text to Image: G is learning the true distribution of the data $p_{data}(image|text)$

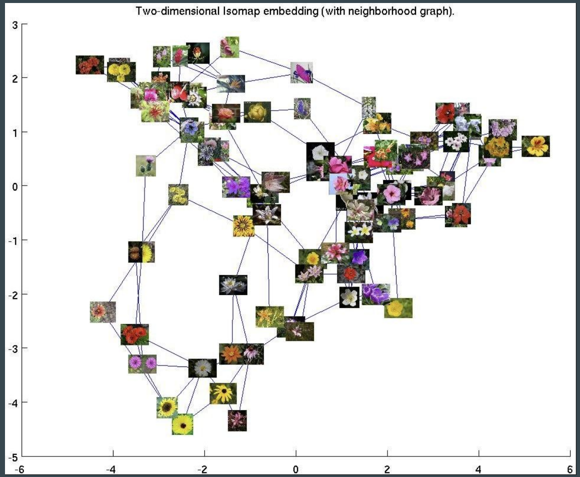
  Image to Text: we are maximizing $p_{data}(text|image)$ directly

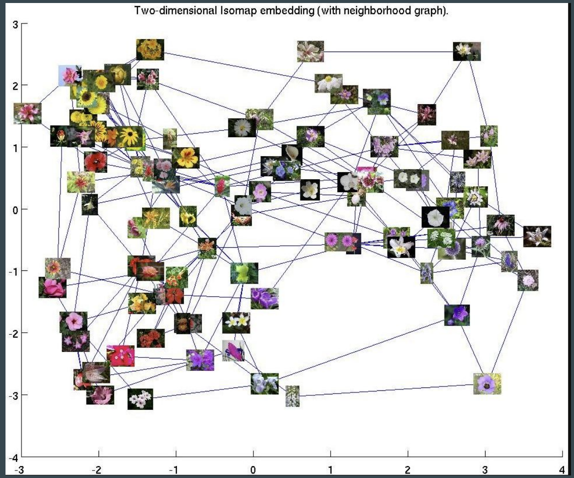
  Note that $p_{data}(image|text)$ and $p_{data}(text|image)$ only preserving information that are shared between image and text. Information not shared are lossed.

Dataset

Oxford 102 flowers with caption:

- 102 Classes of flower
- 8000 images
- 10 captions per image

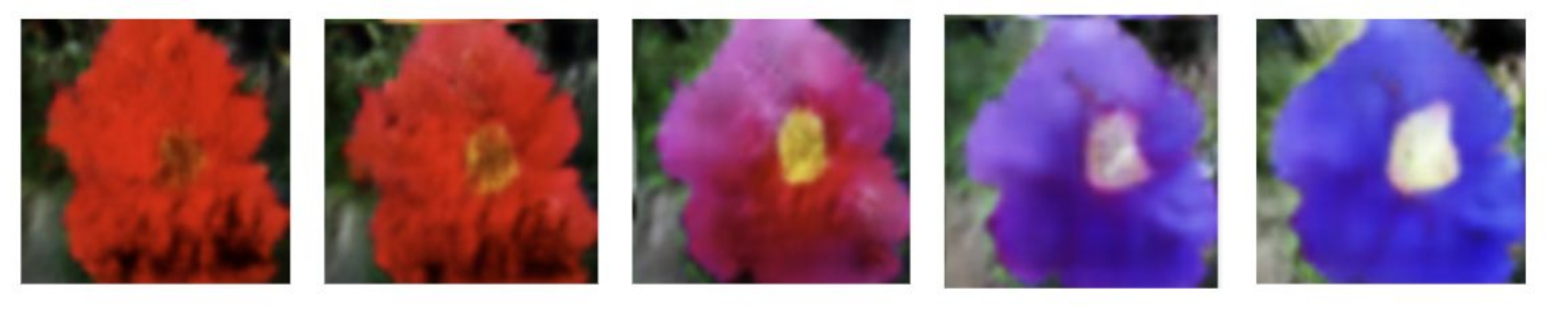




-- *Image from: http://www.robots.ox.ac.uk/~vgg/data/flowers/102/*

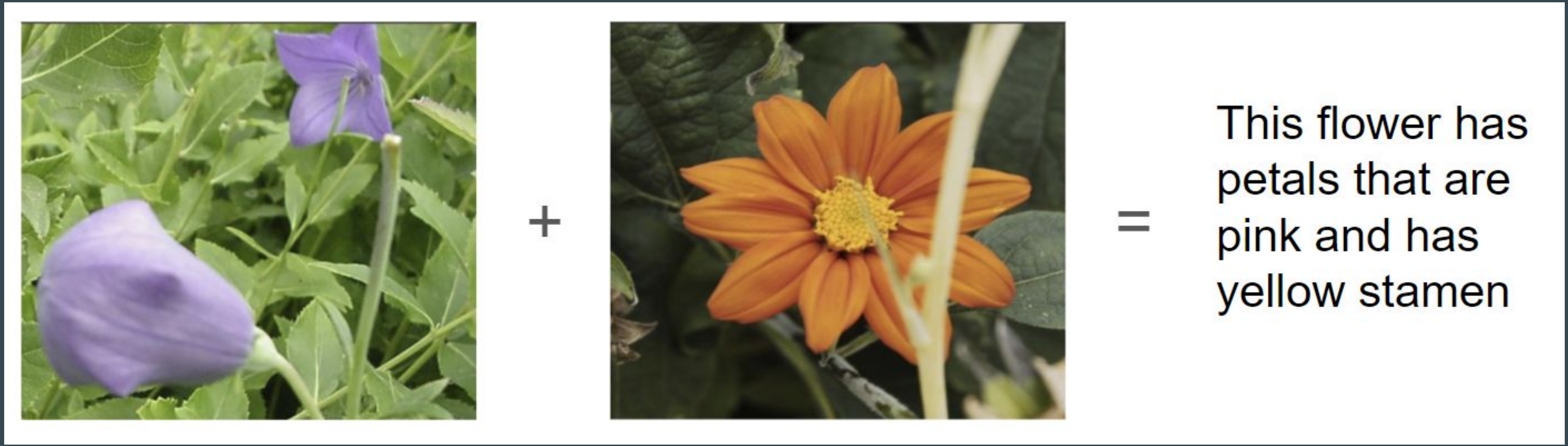
Results

$$Image_{new} \sim F(Text_{new})$$

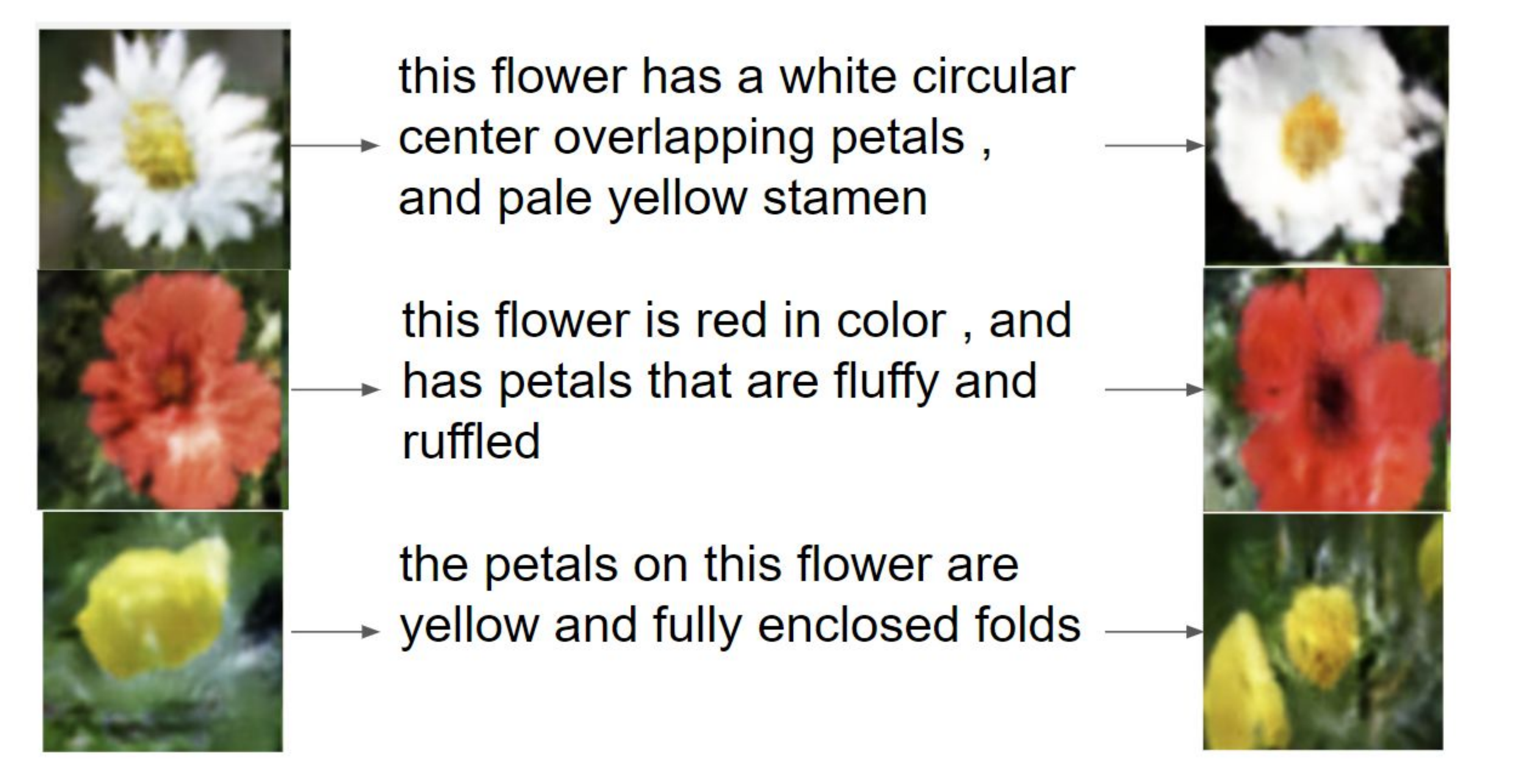


ƛ* Embedding{“The flower is red.”} + (1-ƛ)*Embedding{“The flower is blue.”}

$\text{Text}_{new} \sim F(\text{Image}_{new})$



Cycle-- Image-> Text -> Image



Cycle-- Text -> Image -> Text

this flower has wing like petals with sharp long orange leaves →

→ this flower is orange, with petals that are skinny and pointed.