GAN for Fashion: Fashion Image Generation Conditioned on Text Description

Chen Huang (https://youtu.be/nT0Qa-ioCi8)
Stanford University

Introduction

Motivation
- tremendous application: photo-editing, computer-aided design
- fashion gan: rapidly visualize and modify ideas, or share their ideas with others

long sleeve shirt black. spread collar. zip closure vertical zippered welt pockets

slim-fit jog jeans blue. fade effect distressing throughout. five-pocket styling. belt loop

singlet tank top white . ribbed collar cuff. logo print gray chest . black

short sleeve cotton piquecacete polo purple. ribbed spread collar sleeve cuffs .

Difficulties
- It is very difficult to capture details in original text description
- It is very difficult to train GAN to generate high-resolution photo-realistic images from text descriptions.

Datasets

The FashionGen Challenge

items	quantity
number of images	325,536
Resolution	1360 x 1360
categories	48
Pose	multiple
number of items	78,850

Methods

The whole architecture is follow AttnGAN[5]
- attentional generative network: draw different sub-regions of the image by focusing on words that are more relevant to sub-region being draw
- deep attentional multimodal similarity model (DAMSM): compute the similarity between the generated image and the sentence
  - text encoder: bi-directional LSTM that extracts semantic vectors from text description
  - image encoder: Convolutional Neural Network (CNN) that maps images into semantic vectors
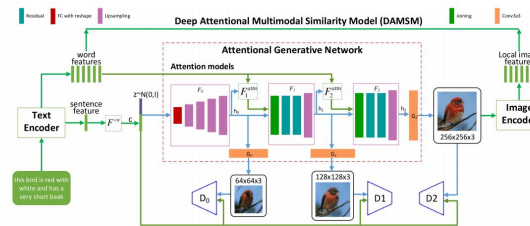
Loss Function: $L = L_G + \lambda L_{DAMSM}$, where $L_G = \sum_{i=0}^{m-1} L_{G_i}$

Results

DAMSM Training Results

AttnGAN Training Results

More Results

	Text				
Text	mid length wool skirt in black, off set vent at hem conceal closure, back fully lined tonal stitch	sleeve bomber black hardware coated ribbed way closure, front vertical pin at pocket, zipper button	slim fit distress cotton cargo multi cal camouflage pattern through mid rise, six pocket, stylish zip	relaxed red elastic waist pocket, styling embrioid in	mid length, 12 gauge knit dress in black, angular sleeve, up closure at sides boxy fit tonal
Stage 1					
Stage 2					

Discussion

Summary
- implement a conditional-gan, generate realistic fashion-photo based on text description
- use attention generative model to build relations between specific word and image region
- two-stage refinement product high-resolution outputs (256 x 256)

Discuss limitations or future work.
1. higher-resolution image (1280 × 1280)
2. multi-modal learning to transfer image to captions

References

1. Reed, S. (2016). Generative Adversarial Text to Image Synthesis. *ICML2016*.
2. Reed, S. (2016). Learning What and Where to Draw. *NIPS2016*.
3. Zhang, H. (2017). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, *ICCV2017*.
4. Zhang, H. (2017). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks, *T-PAMI*
5. Xu, T. (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks, *CVPR2018*

chuang4@stanford.edu