GAN for fashion: Fashion Image Generation Conditioned on Text Description

Chen Huang
Stanford University
chuang4@stanford.edu

Abstract

Automatic synthesis of realistic images from text would be interesting and useful while challenging problem. Samples generated by existing text-to-image approaches can roughly refect the meaning of given descriptions, but they fail to contain necessary details and vivid object parts. In this project, I explore the task of assisting fashion designers to share their ideas with others by translating verbal descriptions to images. Thus, given the description of a particular item, I generate images of clothes and accessories matching the description. I utilize the Attentional Generative Adversial Network (AttnGAN) [9] that allows attention-driven, multi-stage refinement for fine-grained text-to-image generation. Extensive experiments demostrates that the model achieves remarable results on generating photo-relatistic images conditioned on text descriptions.

1. Introduction

Generating realistic images from informal descriptions would have a wide range of applications, such as art generation and computer-aided design.Recently, Generative Adversarial Networks (GAN) [2] have shown promising results in synthesizing real-world images. Conditioned on given text descriptions, conditional-GANs are able to generate images that are highly related to the text meanings. It also drives research progress in multimodal learning and inference across vision and language, which is one of the most active research areas in recent years. [2, 9, 10, 11, 1, 6, 5].

In this solo project I explore the task of generating fashion images given text description about it. This would facility the fashion designers to rapidly visualize and modify ideas, or share their ideas with others. The difficulties remain in threefold: 1) text encoder to correctly capture the detailed information from the image captions. 2) high resolution, photo-realistic image generation. 3) image generation conditioned on the given text descriptions.

A commonly used approach is to encode the whole text description into a global sentence vector as the condition for GAN-based image generation [6, 5, 10, 11]. Although impressive results have been presented, conditioning GAN only on the global sentence vector lacks important fine-grained information at the word level, and prevents the generation of high quality images. So in this project I utilize the AttnGAN [9], which allows attention-driven, multi-stage refinement for fine-grained text-to-image generation.

2. Related Work

Generative image modeling is a fundamental problem in computer vision. There has been remarkable progress in this direction with the emergence of deep learning techniques. Variational Autoencoders (VAE) [4]formulated the problem with probabilistic graphical models whose goal was to maximize the lower bound of data likelihood. Recently, Generative Adversarial Networks (GAN) [2] have shown promising performance for generating sharper images, and have been used in a wide range of applications, including photo-realistic image super resolution, video generation, inpainting, image-to-image translation and text-to-image synthesis.

Build upon these generative model, conditional image generation has also been studied. Reed *et al.*[6] first showed that conditional GAN was capable of synthesizing plausible images from text descriptions. Their follow-up work [5] also demostrated that GAN was able to generate better samples by incorporating additional conditions. Zhang *et al.*[10, 11] stacked several GANs for text-to-image synthesis and use different GANs to generate images of different sizes.

3. Dataset

The training and evaluation is conducted on a large-scale dataset called Fashion-Gen [7]. This new dataset has 293,008 high definition (1360 x 1360 pixels) fashion images paired with item descriptions provided by professional stylists. 260,480 images for training, 32,528 for validation and 32,528 for test, which is larger than other available datasets for the task of text to image translation. Each full HD images are photographed under consistent studio con-

Figure 1: Samples of the dataset.

ditions, and all fashion items are photographed from 1 to 6 different angles depending on the category of the item. Each product belongs to a main category and a more fine-grained category (*i.e: subcategory*). There are 48 main categories, and 121 fine-grained categories in the dataset. Finally Each fashion item is paired with paragraph-length descriptive captions sourced from experts (*professional designers*)

4. Text-to-Image synthesis

I employed the AttnGAN [9] to generate images conditioned on their description. The AttnGAN decomposes conditional image generation into multiple stages. First, the initial GAN sketches a low resolution image (64 x 64) with the overall shape and colors of the image conditioned on the global sentence-level feature and a random noise vector. Subsequentially, the later GAN refines this low resolution image conditioned on the results of the previous stage and the word-level feature to generate higher resolution synthetic images, 128 x 128 and 256 x 256, respectively. The attention mechanism in the generative network enables the AttnGAN to automatically select word level condition for generating different sub-regions of the image. With an attention mechanism, the DAMSM is able to compute the fine-grained text-image matching loss L_{DAMSM}. For the algorithm details please refer to paper [9].

4.1. Text and Image Encoding

According to [7], the method by which I encode the textual descriptions can indeed have a big impact on the quality of the generated images. Here I discuss the text embedding that I applied.

The text encoder is a bi-directional Long Short-Term Memory (LSTM) that extracts semantic vectors from the text description. In the bi-directional LSTM, each word corresponds to two hidden states, one for each direction. Thus I concatenate its two hidden states to represent the semantic meaning of a word. Meanwhile, the last hidden states of the bi-directional LSTM are concatenated to be the global sentence vector.

The image encoder is a convolutional Neural Network (CNN) that maps images to semantic vector. The intermediate layers of the CNN learn local features of different subregions of the image, while the later layers learn global features of the image. More specifically, our image encoder is built upon the Inception-v3 model [8] pretrained on ImageNet.

4.2. Impmentation Details

Throughout all the experiments, the descriptions were lowercased, tokenized and cleared of stop words. I used the first 18 token of the descriptions as input sequence to the encode model.

5. Experiment

This is the interesting part. Extensive experimentation is carried out to evaluate the AttnGAN. I first experiment each important component of the AttnGAN, including the DAMSM and the attentional generative network. Then, I compare this AttnGAN with previous state-of-the-art GAN models on the Fashion-gen dataset.

5.1. Deep Attentional Multimodal Similarity Model

The DAMSM learns two neural networks that map subregions of the image and words of the sentence to a common

PANTS

Relaxed-fit trousers in mottled navy blue. Five-pocket styling. Tonal Zip fly.

(a)

SUITS & BLAZERS

Long sleeve blazer in deep navy. Notched lapel collar. Padded shoulders. closure at front. Welt pocket at breast. Flap pockets at waist. Four-button cuffs. Two vents at back. Partial lining. Tonal stitching.

(b)

Figure 2: Pictures a, b present samples of the dataset. Each description is associated with all the images below it. And each item ie. a, b is photographed from different angles. I also provide each images attributes, and its relationship to other objects in the dataset

semantic space, thus measures the image-text similarity at the word level to compute a fine-grained loss for image generations.

After 10 epochs training, I learned a text encoder (LSTM) network and an image encoder (CNN) network. For a word w_i and an image sub-region r_j, if their encoded vectors are relatively close in the common semantic space, then we assume they are somehow related. This kind of corresponsence is visualized in Figure 3. The row is image from Fashion-gen, and column is the key word from their description. What interesting here is the relationship can not only learn some highly-localized feature (like text), but also can learn abstract feature like color and style. We have more than 20,000 words in dictionary and most of them find consistent target sub-regions in the training images.

Figure 3: Semantic correspondence between word and image sub-region

5.2. Attentional Generative Network

Using the previous image encoder and text encoder as feature extractor, we perform text conditional image generation. In this section we want to evaluate the results in threefold: 1) is the generated image realistic enough, 2)is the generated image matching the input text description, 3) is the semantic corresponding between word and image sub-region accurate?

Here is some generated example images in Figure 4. On the top are text descriptions, they are the inputs to the network. In the middle are generated outputs from different stages. 64 x 64 size is the output from stage I conditioned on sentence-level feature, while the 128 x 128 and 256 x 256 are outputs from later stage, refined by word-level features. The bottoms are corresponding relationships between word and image sub-regions. From here we can tell that the generated image is realistic enough, and well matching the text description.

This is more examples 5 on the text conditional generating output.

Also here I use Inception scores on the validation set to evaluate the overall performance of this model. The result is shown in Table 1

6. Conclusion

Recent progress in generative modeling techniques has great potential to give designers tools for rapidly visualizing and modifying ideas. While recent advances in generative

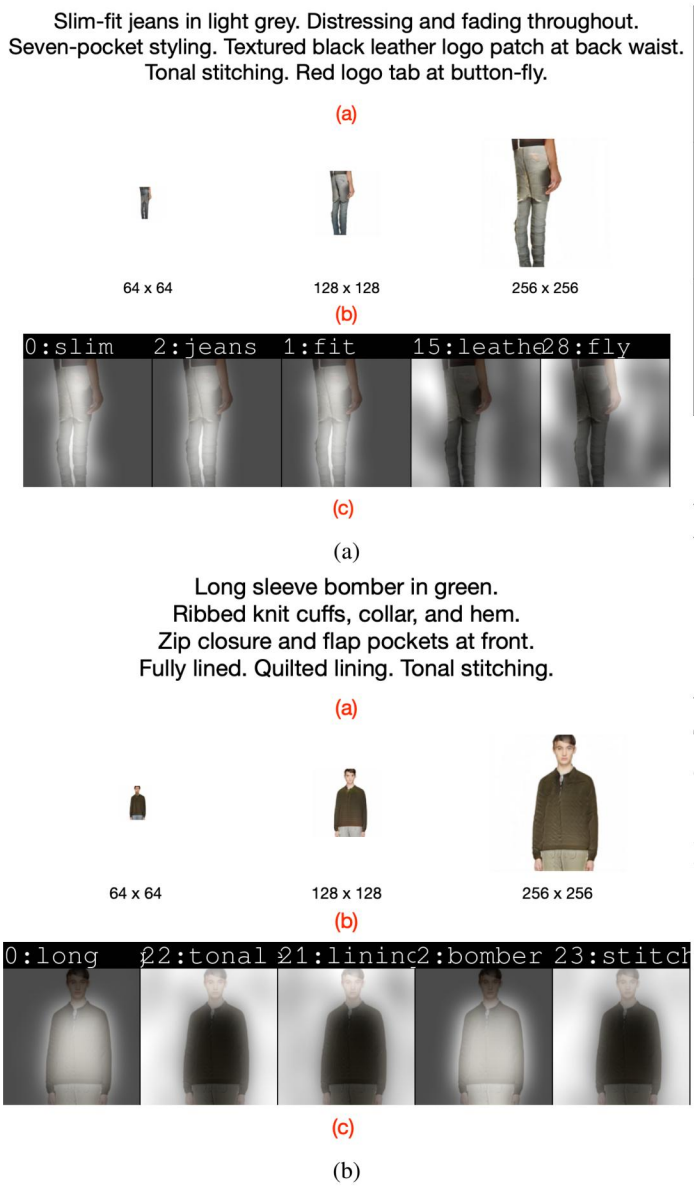

Figure 4: Generated image results. (a) is the text description, (b) is the multi-resolution outputs from different generator, (c) is the visual corresponding between word and image region.

models can be used to generate images of unprecedented realism, the quality of images generated from textual descriptions has so-far remained far from realistic. In this project I explored to AttnGAN models to generate photo-realistic based on fashion product descriptions. Compared to existing text-to-image generative models, this method generates high resolution images (256 x 256) with more photo-realistic details and diversity.

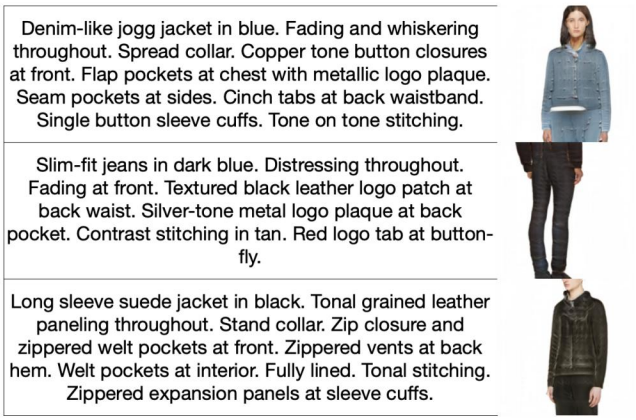

Figure 5: More generating examples

	Inception Score
Fashion Real data 256 x 256	9.71 ± 2.14
StackGAN-v1 [10]	6.50 ± 0.05
StackGAN-v2 [11]	5.54 ± 0.07
P-GAN [3]	7.91 ± 0.15
AttnGAN (This Work)	6.84 ± 0.04

Table 1: Inception Scores on the validation set, i.e: trained on the Fashion train set

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. VQA: visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[2] 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 Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[4] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[5] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 217–225, 2016.

[6] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In M. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning,*

ICML 2016, New York City, NY, USA, June 19-24, 2016, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1060–1069. JMLR.org, 2016.

[7] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal. Fashion-gen: The generative fashion dataset and challenge. *CoRR*, abs/1806.08317, 2018.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

[9] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. IEEE Computer Society, 2018.

[10] H. Zhang, T. Xu, and H. Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5908–5916. IEEE Computer Society, 2017.

[11] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017.