College of Engineering & Applied Science

UNIVERSITY OF COLORADO **BOULDER**

Text to Texture: Adapting Text-Based Image Synthesis with AttnGAN for Fashion Industry

Amulya Ambati, Bhuvanendra Putchala, Danna Gurari

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Section 1: Introduction

- In the rapidly evolving world of fashion, the ability to quickly visualize and iterate on design concepts directly from textual descriptions can significantly enhance the creative process.
- Contemporary breakthroughs in deep learning, including Generative Adversarial Networks (GANs) and attention mechanisms, have significantly propelled the capabilities of text-to-image synthesis.
- This poster presents our findings and the innovative application of the AttnGAN model on fashion data, a novel method which uses concept of attention in the domain of image generation using GAN.

Section 2: Related Works

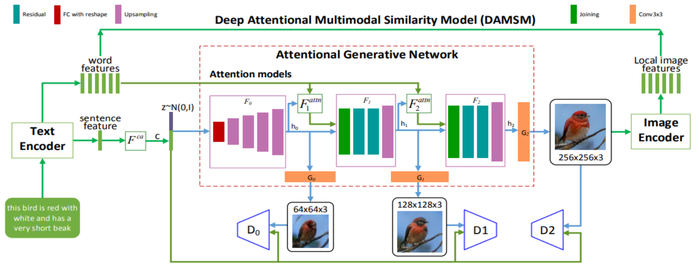
- Initial works have implemented Image synthesis using deterministic neural networks as function approximators. Autoregressive models (e.g., PixelRNN) that utilized neural networks to model the conditional distribution of the pixel space have also generated appealing synthetic images.
- Current generative adversarial networks (GAN) have done substantial impact in text-to- image synthesis, yet image generation in high resolution in detailed images from text remains a challenging task. Having GAN as a base model, Stack GAN(Zhang et al., 2017) proposed a novel idea of using a two-stage generative adversarial network (GAN) process, where the first stage generates a low-resolution image from text to include colors and shape, and the second stage refines this image to a high resolution.
- However, all the past works are conditioned on the global sentence vector, missing fine-grained word level information for image generation.

Section 3 : Dataset

The DeepFashion dataset is a comprehensive collection meticulously curated to advance the state-of-the-art in cloth analysis. It encompasses over 800,000 diverse fashion images ranging from daily wardrobe apparel to high-fashion extravagance, annotated with rich information such as clothing attributes and landmarks. This dataset is organized into several categories, providing a wide array of clothing styles and appearances that cater to various algorithmic tasks such as attribute prediction, consumer-to-shop clothes retrieval, and fashion landmark detection.

Section 4 : Model Implementation

- A novel attention mechanism within the AttnGAN allows the generative network to focus on different sub-regions of the image by keying into the relevant words from the natural language description, a first in layered attentional GANs.
- The multi-stage generative process begins with a base image that gradually refines to add detail. Each subsequent stage uses attention queries from the preceding stage's image features, ensuring each refinement adds relevant details.



Section 5: Model Results

An inception score of 3.75+/- 0.2 was obtained.

The attention map at each stage of generator provided a various attention map.

Limitations:

- Unable to consume multiple aspects of the text.
- Training for more epochs might have resulted in better performance.

Example:

Text description:

Man is wearing a white striped shirt.

Output (Attention Map of generated image):

References:

1) Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June.