FASHION-GEN

ABSTRACT:

New dataset of 293.008 high-definition fashion images paired with item descriptions provided by professional stylists.

INTRODUCTION:

It is explored the task of assisting fashion designers to share their ideas with others by translating verbal descriptions to images. Thus, given a description of a particular item, images of clothes and accessories are generated matching the description.

THE FASHION DATASET:

- It consists of 293.008 images:
  - 260.480 images for training
  - 32.528 for validation
  - 32.528 for test
- Full HD images photographed under consistent studio conditions.
- 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.
- Each fashion item is paired with paragraph-length descriptive captions sourced from experts.
- Metadata is provided for each item. Also, the color distribution extracted from the text description presented.

EXPERIMENTS WITH THE DATASET:

- **Generating high-resolution images using P-GANs**

It follows the same experimental setup and architectural details of the original P-GAN paper.

In order to quantitatively evaluate the quality of the generated images, it is computed the *Inception Score* for the down-sampled version (256x256) of them.

- ## Text-to-Image Synthesis:

StackGAN-v1 decomposes conditional image generation into two stages. First, the *stage1 GAN* sketches a low-resolution image *(64x64)* with the overall shape and colors of the image conditioned on the text and random noise vector. Subsequently, the *stage2 GAN* refines this low-resolution mage conditioned on the results of the first stage and the same text embeddings, and generates a 256x256 image.

StackGAN-v2 follows a similar architecture consisting of multiple chained generators and discriminators. The input of each stage of the chain is the output of the previous stage. One of the major differences between *StackGAN-v2* and *StackGAN-v1* is that these stages are trained jointly, whereas in *StackGAN-v1*, they are trained independently.

- ## Text Embeddings:

It is important for the embedding of the description to correctly relate to the visual contents of the product image.

The experiments are conduced using different encoders from a wide range of complexity, namely averaging word vectors, concatenating word vectors, a slightly modified encoder from the *Transformer* and a *bidirectional LSTM*.

- The *bi-LSTM* model is the one that achieves the highest category classification accuracy on the validation dataset in the pre-training process.
- Irrespective of the encoder architecture, pre-training the encoder model results in better correspondence between the descriptions and the generated images.
- Using the pre-trained *bi-LSTM* with fixed weights as the encoder leads to better results both visually and qualitatively.

- ## Implementation Details:

The descriptions were lowercased, tokenized and cleared of stop words. It is used the first 15 tokens of the descriptions as the input sequence to the encoder model.

StackGAN-v1: Same architecture *Zang et al. 2017a*. The first stage trained for 80 epochs, and the second stage for 185 epochs.

StackGAN-v2: Same architecture and hyper-parameters as *Zang et al. 2017b*.

Score of the StackGAN-v1 is better than Stack-GAN-v2, while the quality of the images in the StackGAN-v2 is better and the reason is due to a significant mode-collapse that was faced in StackGAN-v2.

Most of the faces in StackGAN-v1 and StackGAN-v2 are blurry. It suggests that since the images are conditioned on the text, the model is focusing more on the clothing material than face information.

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 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.