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

Amulya Ambati, Bhuvanendra Putchala, Danna Gurari

University of Colorado, Boulder

Abstract. In this rapidly growing domain of computer vision, synthesizing photo-realistic images from textual descriptions presents a significant challenge with broad applications in the digital content creation to assistive technologies. We would like to use this technology and apply to the fashion industry. Current generative adversarial networks (GANs) have demonstrated a substantial impact on text-to-image synthesis; however, generating high-resolution, detailed images from text remains a challenging task. Samples generated by existing text-to-image approaches can roughly reflect the meaning of the given descriptions, but they fail to contain necessary details and vivid object parts StackGAN.This research aims at adapting one of the novel architecture's AttnGAN [1] which was originally trained on COCO dataset, for the problem of generating fashion images and tries to evaluate and compare the performance. We have obtained an inception score of 3.75 ± 0.2 on the fashion image dataset.

1 Introduction

The Fashion industry is demanding tools that can quickly turn ideas into image representations. This ability to generate fashion images from text descriptions can impact fashion personalized recommendations and e-commerce. This project aims to address the existing gap by introducing AI-driven methods to analyze fashion images from text descriptions, enabling efficient design and enhanced user experience in fashion platforms [2]. Furthermore, this technology holds promise to improve interactions between humans and computers, making it more understanding and productive by enabling machines to understand and visualize human descriptions.

2 Literature Review

Having GAN as a base model, StackGAN[2] 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, shapes, and the second stage refines this image to a high resolution. This method achieved a 28.47 % and 20.30 % increase in inception scores on the CUB and Oxford-102 datasets, respectively, compared to previous state-of-the-art methods, suggesting higher quality and diversity in generated images. While this model excels in controlled

scenarios (e.g., datasets with specific types of objects like birds and flowers), its performance on more general, diverse datasets like MS COCO, which contains images with multiple objects and varied backgrounds, can be less predictable.

Clip-gen[3]uses natural language supervision, where models learn from a massive dataset of image-text pairs, leveraging the descriptive power of text to improve visual understanding. It learns visual concepts from text by predicting the pairing of images and captions through a contrastive learning approach. It learns to map images and text into a shared embedding space, where semantically similar images and text are closer together and dissimilar ones are farther apart. CLIP achieves impressive zero-shot learning capabilities, where it can generalize to new tasks without any specific training on those tasks.

DALL-E[4]introduces an autoregressive transformer-based approach that has shown high efficacy in generating detailed, high-resolution images from textual descriptions, evaluated on diverse datasets such as MS-COCO. The model exhibits robust zero-shot learning capabilities, effectively generating images that align with textual descriptions without direct training on those specific tasks by using discrete variational autoencoder (dVAE) to compress image data, which significantly reduces the input dimensionality for the transformer model, allowing for efficient scaling. However, this failed to generate detailed images in higher resolution.

All the above models can be adapated for fashion image synthesis. However, AttnGAN tries to overcome the limitations of the models by adding an attention layer which thereby tries to capture intricate relations between wordings and it also proposes a 3 layer multi-stage GAN which generates images at three different resolutions. It also tries to incorporate the novelties proposed in all the above 3 models into a single architecture, like stacking GANS, proposing DAMSM loss, which looks for text-image similarity score. Hence, we have proceeded with AttnGAN for our experimentation.

3 Dataset

In this study, we have used DeepFashion MultiModal dataset[5], which is a famous dataset in the field of the fashion industry. Four benchmarks are developed by this database that include Attribute Prediction, Consumer-to-shop Clothes Retrieval, In-shop Clothes Retrieval, and Landmark Detection. The data and annotations of these benchmarks can be also employed as the training and test sets for the following computer vision tasks, such as Clothes Detection, Clothes Recognition, and Image Retrieval. Recently fashion synthesis benchmark has also become an integral part of this database. On a whole, deep fashion consists of 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos, constituting the largest visual fashion analysis database. However, we are interested in using fashion synthesis data which contains long textual descriptions, detailing the fashion image. For our analysis and, study we have considered a subset of 44k images from the entire database.

Below are a few instances from the data:

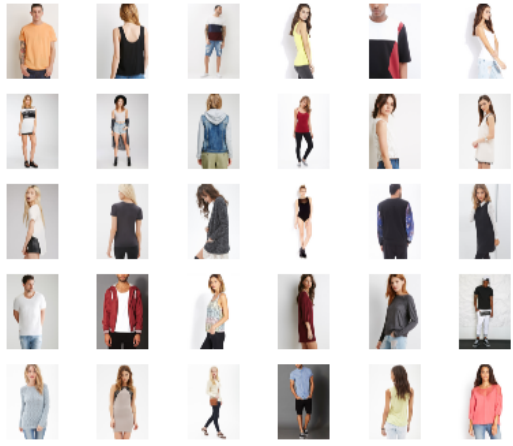

Fig. 1. Grid of images present in the dataset.

Below is the word cloud which provides an idea about the frequent words or the kind of words present in the textual descriptions.

Fig. 2. Word cloud of the text descriptions for fashion images

4 Experimentation

4.1 Architecture

Attentional GAN's have revolutionarized image synthesis process by proposing a novel architecture that uses contempoary booming concepts of Attention combined with conditional GAN. The proposed architecture also enables the gen-

erative network to draw different subregions of the image conditioned on words that are most relevant to those sub-regions.

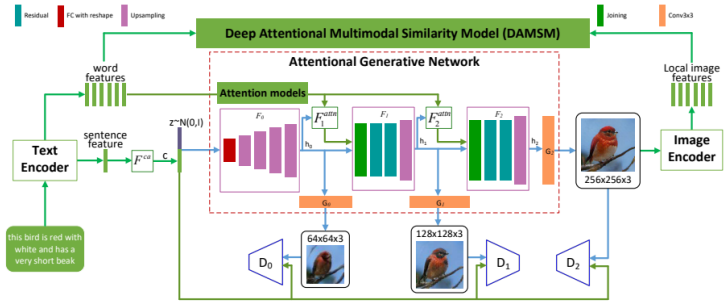

Fig. 3. Attentional GAN architecture

The attention model, which is the heart of this model $F^{attn}(e, h)$ has two inputs: the word features $e \in R^{D*T}$ and the image features from the previous hidden layer$h \in R^{D*N}$. The word features are first converted into the common semantic space of the image features by adding a new perceptron layer, i.e., $e^1 = Ue$, where $U \in R^{D*D}$. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h (query)[6]. Each column of h is a feature vector of a sub-region of the image. For the j^{th} sub-region, its word context vector is a dynamic representation of word vectors relevant to h^j, which is calculated by

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e_i^1$$

$$\beta_{j,i} = \frac{exp(s_{j,i}^1)}{\sum_{k=0}^{T-1} exp(s_{j,k}^1)}$$

and $\beta_{j,i}$indicates the weight the model attends to the i^{th} word when generating the j^{th} sub-region of the image.

It also proposes DAMSM loss, such that it learns two neural networks that map subregions of the image and words of the sentence to a common semantic space, and thereby it measures the image-text similarity at the word level to compute a fine-grained loss for image synthesis[7]. The two networks it learns are Text encoder and image encoder.

For the text encoder, we will be using a bi-directional LSTM, which provides us with 2 hidden states for a single word, which can be further concatenated into a single embedding for the given word and the final last state embedding would be global embedding which will be passed to gaussian noise which will be further passed to GAN.

The next network, image-encoder is a Convolutional Neural Network that maps images to semantic vectors. We have used pre-trained Inception-v3 model for our image encoder. The embedding of size 768x289 is reshaped from and is extracted from “mixed 6e” layer of shape 768x17x17, and also global embedding

of size 2089x1 is extracted from the last average pooling layer. In the end, image features are converted to a common semantic space of text features by passing it through a perceptron layer.

The DAMSM loss is designed to learn the attention model in a semi-supervised manner, in which the only supervision is the matching between entire images and whole sentences (a sequence of words). It consists of 4 parts, sentence level loss and the world level loss such that we maximize the posterior probability of obtaining the given image for a text description and vice versa, which is maximizing the posterior probability of obtaining the given text description for an image, using both word level embeddings and sentence level embeddings.

Hence $L_{DAMSM} = L_w^1 + L_w^2 + L_s^1 + L_s^2$

4.2 Training

Originally, this architecture was developed and tested out on CUB 200 bird dataset which contained 200 unique bird classes and, COCO dataset. It has given an overall best inception score of $25.89 \pm .47$ on COCO dataset and an overall best inception score of $4.36 \pm .03$ on CUB dataset. Hence, we have adapted this architecture for our problem statement of image synthesis for the fashion dataset. Our train data consisted of 25,582 images and was tested on 4,515 images.

The Entire process of training was divided into 2 steps, the first part trains the text encoding module and image encoding module as a separate entity. The next model uses these encoders along with conditional Attentional GAN that facilitates the generation of synthetic images.

DAMSM or Encoders Training

architecture	Bi-Directional LSTM
batch size	16
learning rate	0.002
drop out probability	0.5
dimension of output text embedding	256

Image encoder settings were as follows:

architecture	Inception Net-V3
batch size	16
learning rate	0.002
dimension of output image embedding	256

Below are the loss graphs obtained during the training of encoder modules.

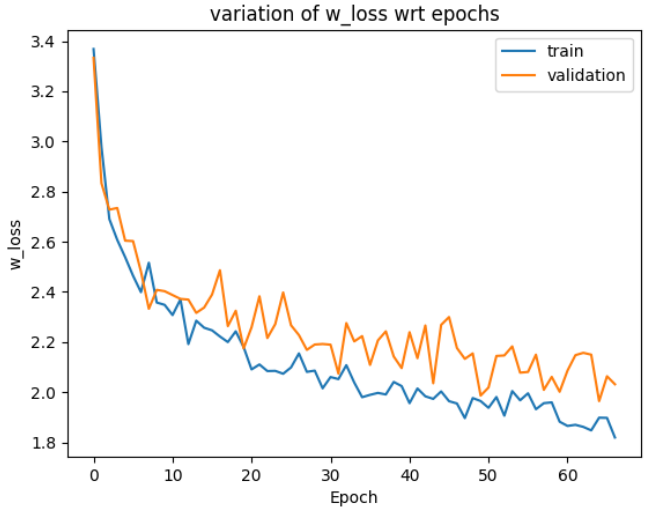

Fig. 4. Word level loss

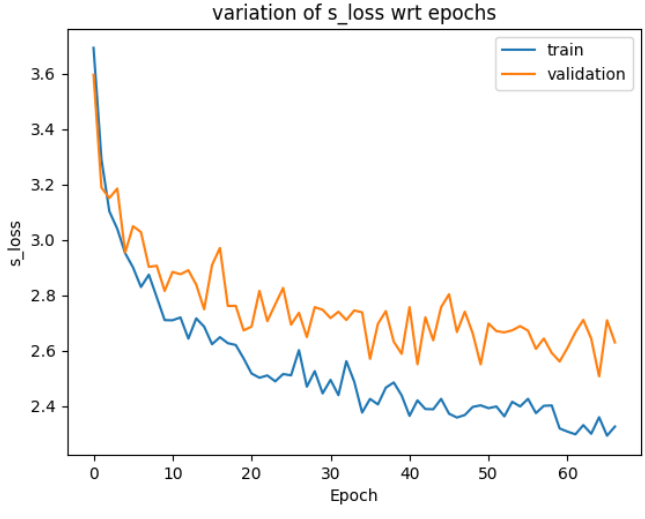

Fig. 5. Sentence level loss

We can observe the reduction in loss as epochs are progressing. However, for validation set sentence level loss has become more or less constant.

GAN After training robust text and image encoders, we train our multi-stage stacked GAN.

batch size	32
Generator learning rate	0.0002
Discriminator learning rate	0.0002
dimension of output text embedding	256
gamma1	4.0
gamma2	5.0
gamma3	10.0
lambda	5.0

Below is the loss curve obtained during the training

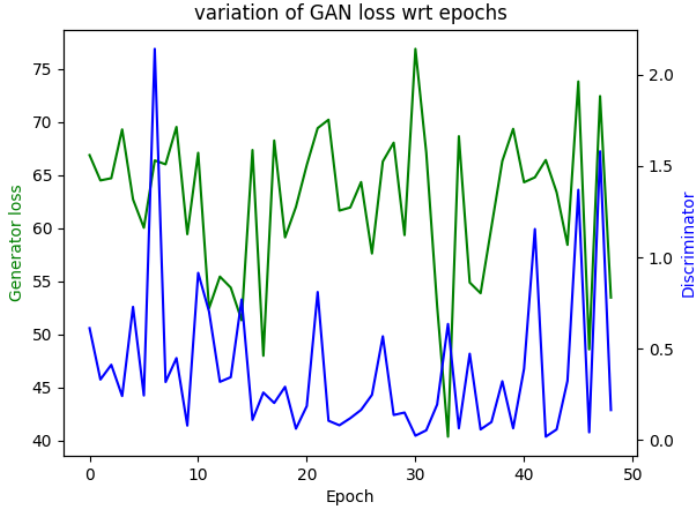

Fig. 6. GAN loss

The loss seems to fluctuate for both generator as well as discriminator, which implies the network hasn't been trained properly. A plausible reason for such fluctuation could be a high learning rate.

The obtained metrics based on the images generated by GAN are as follows:

w_loss	6.33
s_loss	10.79
kl_loss	0.98

The obtained w_loss and s_loss for true image vs descriptions were in the range of 1-2 and 3-2.5 respectively. Whereas for the generated images the w_loss and s_loss are as above, which is high when compared to the true text-image pair, which suggests there is scope for improvement in the model and a higher level of tuning is required.

Below are the gan generated images for different text descriptions present in the dataset.

Fig. 7. Fashion images generated by the model

Below is the attention map obtained on a given test example with the description "The lady wears a long-sleeve shirt with lattice patterns and a long trousers. The shirt is with cotton fabric. The trousers are with denim fabric and solid color patterns."

Fig. 8. Attention map for generated image

The above attention map clearly shows, the model has learned meaningful word embeddings and has a semantic understanding between words vs image regions.

Along with the proposed w_loss and s_loss, the dataset was also evaluated on inception score. An inception score of 3.75 ± 0.2 was obtained.

The loss curve of the GAN indicates that the model has not been adequately trained, as evidenced by fluctuating losses, likely due to a high learning rate. Furthermore, heuristic analysis of the images reveals significant distortions in the human faces, suggesting that this aspect of the model requires further training. As a next step, we propose re-running the model with a lower learning rate, simplifying the text dimensions, and extending the training over more epochs to enhance performance and image quality.

5 Conclusion

In conclusion, the AttnGAN demonstrated effective generalization capabilities on the fashion dataset. Upon reviewing the images and attention maps, it is evident that the model successfully captured semantic relationships between words and image sub-regions. However, the loss curves indicate that while the encoders were adequately trained, the loss exhibited fluctuations across epochs and did not converge for the GAN model. A lower learning rate and extended training might enhance performance. Due to computational constraints, further retraining of the model was not feasible. Another observed limitation was the model's reduced ability to recognize multiple components in a given description, particularly as the length of the text input increased.

The foremost concern about this problem statement is fairness, ensuring that our model does not perpetuate or exacerbate biases present in the training data. This includes biases related to body size, race, gender, and age, which could lead to discriminatory practices or reinforce negative stereotypes. Privacy must also be safeguarded, particularly in scenarios where user-generated descriptions are used for training or when personal data could inadvertently be included in the dataset[8]. Lastly, the social impact of automating such creative processes should be considered, including the potential displacement of jobs and the cultural implications of globalized fashion trends being dictated by AI systems.

For future prospects, we aim to elevate our project by creating a website or an API that would generate fashion images based on the text descriptions provided.

References

1. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. CoRR **abs/1711.10485** (2017)
2. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
3. Wang, Z., Liu, W., He, Q., Wu, X., Yi, Z.: Clip-gen: Language-free training of a text-to-image generator with clip (2022)
4. Seneviratne, S., Senanayake, D., Rasnayaka, S., Vidanaarachchi, R., Thompson, J.: Dalle-urban: Capturing the urban design expertise of large text to image transformers. In: 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA). (2022) 1–9
5. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 1096–1104
6. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022) 1552–1565

7. Yang, B., Xiang, X., Kong, W., Zhang, J., Peng, Y.: Dmf-gan: Deep multimodal fusion generative adversarial networks for text-to-image synthesis. IEEE Transactions on Multimedia **26** (2024) 6956–6967
8. Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: The generative fashion dataset and challenge (2018)