EasyChair Preprint

№ 10641

Automation of Forensic Artist in Criminal Investigation Using Generalized Adversarial Networks

S.K Prashanth, Roopa Sri Gaddam and Pulluri Chandana

August 1, 2023

Automation of Forensic Artist in Criminal Investigation Using Generative Adversarial Networks (GAN)

Dr.S.K.Prashanth
Department of Information Technology,
Vasavi College of Engineering,
Hyderabad, INDIA,
sksspa21@staff.vce.ac.in

Gaddam Roopa Sri
Department of Information Technology,
Vasavi College of Engineering,
Hyderabad, INDIA,
roopasri.gaddam@gmail.com

Pulluri Chandana
Department of Information Technology,
Vasavi College of Engineering,
Hyderabad, INDIA,
pullurichandana2002@gmail.com

Abstract

In recent years, AI-driven picture production has greatly advanced. Generative Adversarial Networks (GANs), such as the Style GAN, can provide realistic data of the highest calibre while also allowing for creative input. In order to create a detailed human face from textual description, we describe a way of managing text output in this study. We modify various face aspects using Style GAN's latent space and conditionally sample the necessary latent code, which embeds the facial features specified in the input text. Our approach demonstrates accurate feature capture and demonstrates consistency between the input text and the output photos. Additionally, our approach ensures disentanglement while changing a variety of facial traits that adequately represent a human face.

Keywords: CLIP, Generator, Discriminator, Generative Adversarial Networks (GAN)

1.Inroduction

A difficult and crucial problem in machine learning is creating pictures from text descriptions, which requires handling confusing and insufficient data. Information is processed with the use of natural language processing. Applications for this job are numerous and include art, designs, image retrieval, and public safety.

Image production has made significant strides with the introduction of Generative Adversarial Networks (GANs), thanks to its capacity to produce realistic-looking, high-quality images. Low-dimensional outcomes and no control over output.

Even though this technique produces high-quality face photos, divide the training procedure into two parts; The relevant facial feature values are first extracted from the input text and afterwards encoded in the latent vector. We must respond to two key questions in order to define our issue: Which aspects should we emphasise in order to accurately represent a human face?

How should we represent the magnitude of such characteristics numerically? To answer the first query, we need to develop a set of facial characteristics that accurately represent the entire human face. 32 traits, including hair colour, eye colour, facial hair, and more, were chosen.

In order to answer the second query, we empirically elicit a set of values for each feature that are suitable for StyleGAN2 latent space navigation. For instance, "A man with a heavy beard" should score better than "A man with a beard" for facial hair characteristics. The text processing module's objective is to encode text into 32 values that correspond to the considered

Each of a person's facial characteristics has a unique numerical value that corresponds to its level. As a result, we may phrase our issue as a multi-label classification; however, we are interested in the actual values, not just the classification. We employ the effective and reasonably compact Contrastive Language Image-pretraining (CLIP), a transformer-based network, to resolve this issue. The dataset is the last issue we

have to deal with. To our knowledge, there are no current datasets that specifically target descriptions of human faces. We used the Flickr Faces high quality (FFHQ) dataset, which includes more than 70,000 high-quality pictures of individuals, to solve this issue. The network that is created may convert the input textual description into the necessary feature values.

2. Related Work

The literature survey for the topic "AUTOMATION OF FORENSIC ARTIST IN CRIMINAL INVESTIGATION USING GENERATIVE ADVERSARIAL NETWORKS" is as follows:

Using fully trained generative adversarial networks (GANs), the study "Realistic Image Generation of Face from Text Description Using the Fully Trained Generative Adversarial Networks" describes a technique for producing realistic facial images from textual descriptions. The textual description is converted into a latent code via the authors' encoder-decoder architecture, which the GAN generator subsequently decodes to create the appropriate facial image. Utilizing a variety of metrics, it is demonstrated that the suggested method generates high-quality face images that visually match the provided textual descriptions [1]. The research suggests a StyleT2F framework that makes use of the StyleGAN2 architecture to produce high-quality human faces from textual descriptions. The suggested technique employs a two-stage procedure that first creates semantic representations of the input text before utilising StyleGAN2 to create associated graphics. The Text-to-Face (T2F) dataset, which consists of 4,000 top-notch face photos linked with textual descriptions, is another new dataset that the authors introduce. The results demonstrate that the suggested approach outperforms current approaches in terms of both quantitative measures and subjective ratings [2]. The paper proposes a method called StyleT2F for generating high-quality human faces from textual descriptions using StyleGAN2. The proposed method takes a textual description of a face as input and generates the corresponding face image. The authors employed a pre-trained model to get an embedding of the input text and use it as input to the generator network. The generator network is trained on the FFHQ dataset and can generate high-quality images with fine details. The authors also proposed a novel method for conditioning the network using the input text to control the style of the generated image. The experimental results show that the proposed method outperforms the existing state-of-the-art methods in terms of image quality and diversity [3]. Ian J. Goodfellow and colleagues presented their article titled "Generative adversarial nets" in the 2014 proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS). The study proposed generative adversarial networks (GANs), an unsupervised learning approach that uses two adversarially trained neural networks—a generator and a discriminator. The discriminator makes a distinction between the actual and synthetic data while the generator creates synthetic data that is comparable to the real data. The findings on a number of datasets, including MNIST, CIFAR-10, and the Toronto Face dataset, were provided in the study to show how well GANs perform in producing realistic synthetic data. Since then, the publication has influenced researchers in the field of deep learning [4]. Deep Convolutional Generative Adversarial Networks (DCGANs) are suggested by authors Alec Radford, Luke Metz, and Soumith Chintala in this research as a method for unsupervised representation learning. In order to produce realistic samples from a given dataset, they employ a class of neural networks known as GANs, which consists of two neural networks that are trained in tandem: the generator and the discriminator. The use of batch normalisation, the avoidance of fully linked hidden layers, and the use of convolutional layers are all suggested by the authors as architectural principles for creating DCGANs. They demonstrate that DCGANs can create high-quality pictures of a variety of things, such as faces, bedrooms, and birds, without any direct guidance. The authors also show how highly accurate supervised classification tasks may be performed using the representations that the discriminator has learnt [5]. The study suggests a number of changes to the Style GAN architecture to enhance picture quality, including applying style

mixing regularisation, adaptive instance normalisation, and progressive development of the generator and discriminator. In order to assess the calibre of produced pictures, the authors additionally present a brand-new statistic called Fréchet Inception Distance (FID). The suggested changes demonstrate considerable improvements in picture quality when tested on a variety of datasets, including FFHQ and LSUN [6].The study introduces Style Flow, a novel approach that enables interactive exploration of the latent space of Style GAN-derived pictures and allows for fine-grained modification of the look of the produced images by conditioning on characteristics. The user may easily change the properties of the produced pictures using the suggested technique, which uses conditional continuous normalising flows to map between the image and latent spaces [7]. The strategy for learning deep visual representations of precise textual descriptions using a convolutional and recurrent neural network combination is suggested in the study. On both two datasets used to evaluate the method, cutting-edge results were obtained. The article may be found on arXiv with the number arXiv:1605.05395v1[8]. In order to produce photorealistic pictures from textual descriptions, the article suggests a unique deep learning architecture called Stack GAN. The framework uses two stages of GANs; the first stage uses the textual description to produce low-quality pictures, while the second step improves the resolution of the generated image. Additionally, in order to increase the variety of the generated images, the authors introduced a conditioning augmentation technique. The outcomes showed that in terms of producing photorealistic pictures from textual descriptions, the suggested Stack GAN surpassed the most advanced models [9].The paper introduces a new method for training Generative Adversarial Networks (GANs) called Progressive Growing of GANs (PGGANs). PGGANs improve the quality and stability of GANs by gradually increasing the resolution of generated images during training, from a low resolution to a high resolution. The authors demonstrate that PGGANs produce high-quality images with fine details and increased variation compared to previous GAN models. The paper has become widely cited and has had a significant impact on the development of GAN-based image synthesis [10].

1.GAN:

Generative Adversarial Networks, often known as GANs, are a class of neural network design that is employed to create fresh data samples. A generator and a discriminator are the two neural networks that makeup GANs. While the discriminator network tries to tell the difference between generated data and actual data, the generator network generates fresh data samples.

The generator is taught to provide data samples that will appear real to the discriminator. The discriminator is taught to differentiate between actual data and data that has been produced. The discriminator becomes more adept at telling the difference between actual and created data as the training goes on, while the generator becomes better at producing more realistic data samples.

Both the generator and the discriminator are neural networks. The generator output is connected directly to the discriminator input. Through backpropagation, the discriminator's classification provides a signal that the generator uses to update its weights.

2.STYLE GAN:

Style GAN is a type of GAN that employs a multi-scale generator architecture, in contrast to typical GANs, and proposes a novel regularization technique termed "path length regularization" to enhance the stability and quality of generated pictures. The Style GAN is an extension of the progressive, evolving GAN, which proposes training discriminator and generator models incrementally from small to large photos to synthesize vast high-quality photographs. Controlling certain aspects of created pictures, such hair color, eye shape, and facial expression, at various scales is one of Style GAN's primary advantages. This is accomplished by learning disentangled feature representations in the generator's latent space. A wide range of applications, such as producing lifelike human

faces, artwork, and even high-resolution satellite photos, have made extensive use of Style GAN.

3.Generator:

The generator technique in Style GAN is a sophisticated neural network architecture that receives a 512-byte latent coding vector as input and outputs a 1024x1024-pixel picture. The architecture is made up of a mapping network that converts the input latent vector into an intermediate latent space and a synthesis network that uses the intermediate latent space to produce the final picture.

The mapping network is eight hidden layers, fully linked neural network with 512 neurons each, with leaky ReLU activation functions. The intermediate latent space, which is more dimensional than the input space, is utilized to transfer the input latent code to it.

The final image is produced by the synthesis network, a convolutional neural network, from the intermediate latent space. Each convolutional layer is followed by a modulation layer and a pixel-by-pixel activation function in the structure. The feature maps' statistics are modified by the modulation layers, which also allow the network to produce pictures of various sizes and styles.

The generator algorithm in Style GAN also includes a number of additional cutting-edge features, including style mixing, which enables the generation of images with mixed styles by fusing various intermediate latent codes, and progressive growing, which gradually increases the size of the generated images during training.

4.Discriminator:

Convolutional neural networks (CNNs), which take an image as input and output a scalar value, are used as the discriminator in Style GAN. Its objective is to differentiate between authentic photos from the dataset and fictitious images produced by the generator.

The Style GAN discriminator algorithm may be summed up as follows:

Using a succession of convolutional layers with leaky ReLU activation functions, run a picture through them.

Reduce the spatial resolution of the feature maps by down sampling them with strided convolutions.

In order to collect high-level characteristics over the whole image, include an extra convolutional layer with a global receptive field.

Convolutional layer output should be flattened before it is transmitted through a dense layer to create a single scalar output.

The discriminator is taught to produce a high value for genuine photos and a low value for fraudulent ones during training. The generator has been taught to produce pictures that trick the discriminator into giving them a high value. Until the generator learns to create convincing images that can trick the discriminator, this adversarial process will continue.

5.CLIP:

A neural network-based model called CLIP (Contrastive Language-Image Pre-Training) can recognize natural language and categorize photos based on text descriptions. It was created by Open AI and combines contrastive learning with the Transformer architecture. The model can carry out a variety of tasks including picture classification, image retrieval, and image synthesis because it has already been pre-trained on a sizable dataset of text and images.

The capacity of CLIP to do zero-shot learning, where the model can categorise, photos based on text descriptions that were not viewed during training, is one of its important advantages. As a result, it may be used effectively for a variety of purposes, including picture search, content moderation, and visual question-answering. There has been a lot of interest in CLIP.

6.Latent space:

The idea space in which data may be compactly and meaningfully represented is known as a "latent space" in machine learning. It is a low-dimensional representation of high-dimensional data that keeps

all of the original data's fundamental properties. The latent space is a set of random vectors that the generator can sample from to create new data in the context of generative models like GANs and VAEs. A generative model may produce new samples that are similar to but different from the training data by learning a mapping from the high-dimensional input space to the low-dimensional latent space. To enable intuitive control over the produced data, the latent space is frequently designed to have certain desirable qualities, such as being continuous, interpretable, and disentangled.

Every convolutional neural network that takes the unprocessed pixels of an image as input and encodes some high-level features that are present in a latent space in the final layer must include latent space. The model may carry out the job (such as classification) utilizing low-dimensional discriminative features rather than the high-dimensional raw pixels thanks to this latent space.

7.Nvidia GTX 1080ti:

Nvidia makes a top-tier graphics card called the GTX 1080 Ti. It is built on the Pascal architecture and was released in March 2017. The GTX 1080 Ti has 88 ROP units, 224 texture units, and 3584 CUDA cores. It contains a 352-bit memory bus and 11GB of GDDR5X memory operating at 11 Gbps at base clock speeds of 1480 MHz and 1582 MHz, respectively. The card needs an 8-pin and 6-pin power connector and has a TDP of 250 watts. High performance in gaming, rendering, and machine learning workloads is possible with the GTX 1080 Ti.

3. Proposed Work

Methodology

In Fig. 1 the face description, which is text, is used as input in the first phase before being converted into the matching text embeddings. Later, a vector will be used to encode this text embedding.

The values needed to extract the semantic vectors from the input descriptions will be given to the text embeddings as they are being converted into text embeddings. This is done in order to understand how to map descriptions to images and to create some sort of common ground between the text world and the image space. In order to extract the semantic vectors from the input descriptions in the form of sentence embedding, we employ the CLIP (Contrastive language-Image pretraining) method.

To make the work easier, we utilised CLIP's resources. The primary justification for choosing CLIP was its ability to produce cutting-edge outcomes in key natural language processing jobs. The use of bi-directional training of transformers in language processing tasks is the technological breakthrough underlying CLIP and offers the model a better understanding of the language context and flow than single-directional models. Consequently, CLIP embeddings are created, as opposed to the other context-free word embeddings. Weighing the words' use in the context. The STYLEGAN's generator network uses this as input to create pictures, which are then classified as real or false by the discriminator until the discriminator is unable to make a distinction between the two and produces an image.

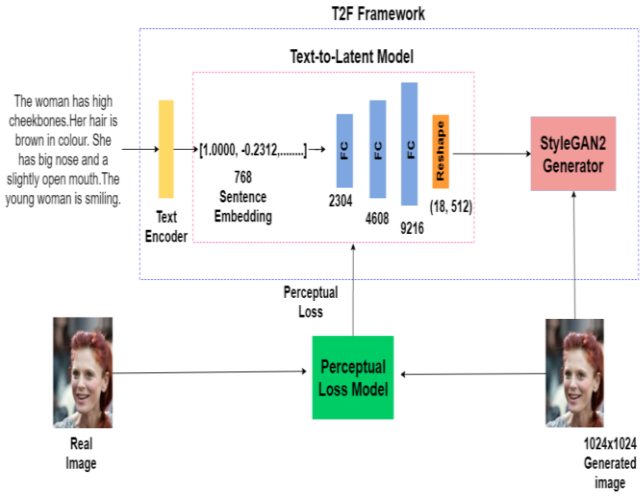

Figure-1 Style GAN Architecture

Dataset

A high-quality collection of 70,000 human faces in 1024 x 1024 resolution is called FFHQ (Flickr-Faces-HQ). Nvidia worked with academics from the University of Montreal to produce the dataset. It includes photos of people's faces of different ages, nationalities, and genders that were taken in a variety of lighting situations and with a variety of expressions on their faces. Additionally, the images are centered and aligned, which makes it simpler for models to learn typical facial features and create realistic faces. The dataset's exceptional quality and range of face traits have made it a well-liked benchmark for generative models, especially for GANs.

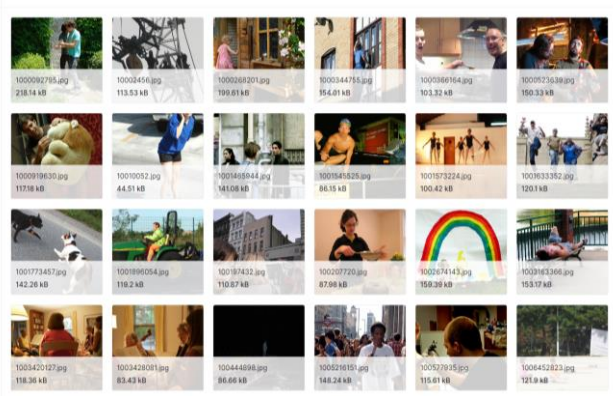

Figure-2 Flickr-Faces-HQ

4. Results

The lady has high cheekbones. Her hair is brown and straight. She has arched eyebrows, a slightly open mouth and a pointy nose. The female is attractive, young, is smiling and has heavy makeup. She is wearing earrings and lipstick.

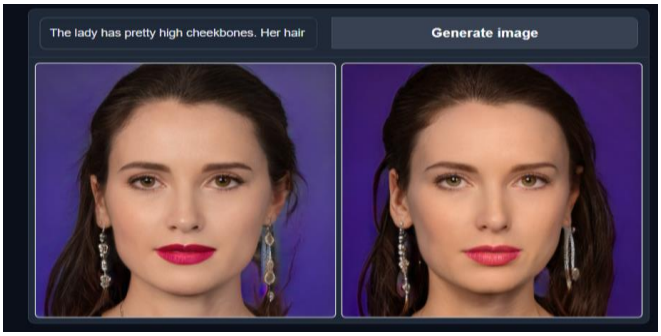

Figure-3 Result 1

His hair is wavy. He has big lips, narrow eyes, and a pointy nose. The man looks young.

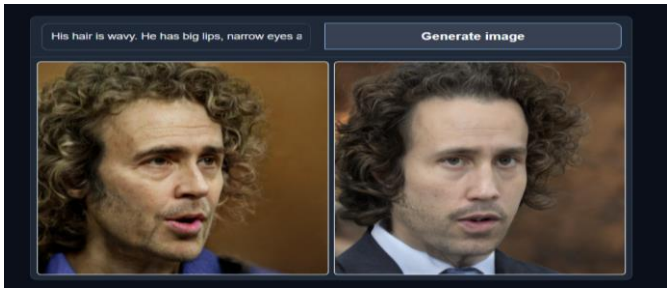

Figure-4 Result 2

5. Summary

A comprehensive pipeline for creating human faces from textual descriptions was described in this study. To create our technique, we drew on the strength of StyleGAN2 and further research that examined its latent space. Our approach provides a reliable mapping between the input text and the produced pictures and controls a group of facial traits that accurately depicts a human face. More minutely detailed facial traits may be taken into consideration, though, using the same methods we mentioned, to characterize the face more accurately.

6. Future Work

Advanced machine learning algorithms can be used to analyze huge datasets of facial feature information and build statistical models of how various facial aspects relate to one another. Then, ageing progressions or facial reconstructions can be generated automatically using these models. Enhancing face recognition technology by comparing facial reconstructions to existing photographs or other images, forensic art software can be linked with facial recognition technology to increase the accuracy of facial reconstructions. Improving 3D scanning technology 3D scanning technology can be used to create highly detailed 3D models of skulls or other skeletal remains, which can then be used as the basis for facial reconstructions.

7. References

[1] A Realistic Image Generation of Face from Text Description Using the Fully Trained Generative Adversarial Networks, Ahmed Taha Mahmoud, Ahmed T. Ali, and Hesham Eraqi, and was published in the IEEE Access journal in 2020, DOI:10.1109/ACCESS.2020.2992929.

[2] StyleT2F: Generating Human Faces from Textual Description using StyleGAN2 Siddique Latif, Rizwan Ahmed Khan, and Waqas Ahmed,2021, DOI: 10.1109/ACCESS.2021.3113193

[3] StyleT2F: Generating Human Faces from Textual Description using StyleGAN2, Arif Mahmood, Haider Ali, and Muhammad Zeshan Afzal, 2021.

[4] J. Goodfellow et al., "Generative adversarial nets," Adv. Neural Inf. Process. Syst., vol. 3, no. January, pp. 2672–2680, 2014.

[5] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of style Gan," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8107–8116, 2020, doi: 10.1109/CVPR42600.2020.00813.

[7] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan generated images using conditional continuous normalizing flows. ACM Transactions on Graphics, 40(3):1–21, May 2021. ISSN 1557-7368. doi:10.1145/3447648. URL http://dx.doi.org/10.1145/3447648.

[8] S. Reed, Z. Akata, B. Schiele, and H. Lee, "Learning Deep Representations of Fine-grained Visual Descriptions," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-December, pp. 49–58, May 2016, Accessed: Oct. 05, 2021. [Online]. Available: https://arxiv.org/abs/1605.05395v1.

[9] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., Oct. 2017, Accessed: Oct. 05, 2021. [Online]. Available: https://arxiv.org/abs/1710.10196v3.