Proceedings of the Fourth International Conference on Inventive Systems and Control (ICISC 2020)
IEEE Xplore Part Number: CFP20J06-ART; ISBN: 978-1-7281-2813-9

Creation of Face for Aiding in Forensic Investigation based on Textual Description

Mukesh Jha
Information Technology
Sardar Patel Institute
Of Technology
Mumbai, India
jhamukesh998@gmail.com

Sahil Jobanputra
Information Technology
Sardar Patel Institute
Of Technology
Mumbai, India
sbkjobanputra@gmail.com

Divya Kamath
Information Technology
Sardar Patel Institute
Of Technology
Mumbai, India
divya.kamath@spit.ac.in

Prof. Varsha Hole
Information Technology
Sardar Patel Institute
Of Technology
Mumbai, India
varsha_hole@spit.ac.in

***Abstract*—In this work, we have developed a solution that will assist the police department in the criminal investigation process during drawing sketches of a suspect based on the textual description from a witness. We propose a deep learning implementation based on Generative Adversarial Network that consists of generator which generates an image based on text description and discriminator which validates the image generated by generator with the data set. We trained a subset of 400 imagescollectedfromtheLabeledFaceInWilddataset.The description of images was taken from Face2Text data provided by Albert Gatt. The images generated resemble certain features mentioned in the textual description which can directly aid the cops or can act as an intermediary for sketch artists to work upon.**

Keywords — generative adversarial networks; forensics; generators; discriminators; k-l divergence;

I. Introduction

In the sketching phase of the traditional system of investigation process, witnesses describe the facial features of the criminal or suspicious person and the forensic sketch artist draws the image based on the description so that the crime department can track down the person resembling the sketched face. This process is tedious and time consuming, takes hours to sketch the face. We propose a solution which can eventually reduce this time considerably. The limitation with traditional method is that we need a skilled forensic sketch artist who may or may not be available at a given point of time. In case of urgency, the traditional method fails, whereas our proposed solution will work under such conditions. Though the quality of image generated isn't quite appealing but it can be used as a basis for further investigation as it resembles almost all the features mentioned in description.

We have kept the description limit to 100 words and there are some input constraints like the input text must be in lower case characters and some numbers like approximate age of the person can be given. The accepted ASCII values of characters are (65-90) and (48-57) along with special characters like "," and ".". The input format is assumed to follow the standard English notation.

II. Objectives

A. Speeding the Process of Forensic Investigation

This system automatically generates images based upon best trained model and given textual description and thus the time taken to generate the image is very less.

B. Making the images of suspect which takes all features into consideration

The forensic sketch artist has to make versions of sketches and keep enhancing every feature of the face based on the user input given by the witness. In our case the model itself will consider all the relevant words in the text and generate images using the machine learning techniques and algorithms used.

C. No human intervention is required

The role of a forensic sketch artist is automated by our model. The human error can be minimized by our model if some features are not taken into account or the user description is interpreted incorrectly by forensic sketch artist.

III. Related Work

Concept Of GAN is widely used machine learning technique when working on generation of images. Applications of GAN in image synthesis, image to image translation, super resolution, classification and regression [6]. But in this field of text to image conversion by using GANs, some work has been done in the past using Caltech-USCD Birds generating the images of birds based on their textual description. In that model StackGan [8] is used which varies from the traditional GAN, as the traditional Generative Adversarial networks are difficult to train, unstable and highly dependant on choice of hyper-parameters. StackGAN is used because training in StackGan occurs in two stages that allows the new stage to learn from the previous one to provide better results with good resolution [2].
StackGAN version1 stage 1 generates low resolution 64x64

978-1-7281-2813-9/20/$31.00 ©2020 IEEE

image which acts as input for StackGAN stage 2 which will generate high resolution image taking into account some features omitted by stage 1. StackGAN version2 trains multiple generators and discriminators alternatively.

Another type of GAN is ProGan [3] which is used for improving the resolution of generted image. This method involves training GAN progressively and add layers to increase resolution as the training progresses. The generator and discriminator networks are opposite of each other and grow simultaneously. Multiple generator and discriminator work at varying resolution and every layer can be trained throughout the process. Thus, they speed up and stabilize the training process. The training on Celeb A data set generated images of 1024x1024 resolution. Training time was reduced by factor of 2.

Conditional Augmentation [7] has been used to effectively achieve text encoding in the text description. The problem of high dimensional latent space for text embedding is eliminated by using Conditional Augmentation. It produces additional conditional variables on top of the conditional latent variables generated by transforming the text embedding.

In text to image generation, most methods don't differentiate between image foreground and background causing the main object in the image to get distorted. This problem has been eliminated by using a method that uses combination of context aware variational auto-encoder and conditional GAN [5]. Context aware variational auto-encoder captures image layout, color and separates image foreground from background. This approach of context-aware mechanism is advantageous and generates more clear images which caters to better images.

IV. Challenges

The main challenge in implementing this system was to find a relevant dataset consisting of annotations of given images. So, we made use of dataset consisting of 400 images collected by Albert Gatt [4] of University of Malta for his research paper Face2Text [4] corpus for generating descriptions of faces. This RIVAL group of researchers belonging to University of Malta created annotated image descriptions by making a crowd sourcing website which took descriptions from various users on images in LFW dataset. The description is cleaned and converted to json format.

V. Methodology

The model makes use of the concept of Generative Adversarial Network abbreviated as GAN. The Generator generates images from the Gaussian noise and text embedding vector while discriminator validates it. These are the following processes while generating the face.

A. Input and Data Preprocessing:

The model takes the input from the user in the form of a sentence. We have kept the description limit to 100 words which considers all the alphanumeric numbers as valid input. The input format is assumed to follow the standard English notation. The network removes all the stop words from the sentence.

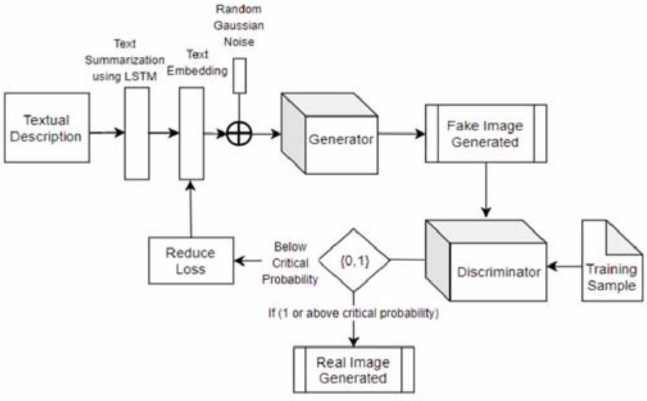

Fig. 1: System Architecture Diagram

The sentence is split into words which is called Tokenization. We have maintained a counter in the sentence for all the useful words. The counter is used to assign a unique value at each unique word just like HashMap in the form of key, value pair. The key consists of words and value consists of the unique value assigned.

B. Text Encoding, LSTM and Conditioning Augmentation:

The model performs text encoding which uses LSTM (Long Short Term Memory), which is a type of Recurrent Neural Network. LSTM network is used to maintain the context of the sentence so that we can check whether similar kind of words have occurred in the previous part of the sentence. If it has occurred in the past then the value is updated to similar one. For instance, if a sentence is " a bald man with black eyes is smiling . he seems furious ", now in this sentence "man" and "he" as well as "black" and "eyes" are assigned same values respectively. So our context of the sentence with it's literals is maintained. This process is called Text Embedding.

The text embedding vector is then passed to conditioning augmentation [7] which is responsible to maintain the embedding vector size same as latent size. If the size is less then it adds up the conditioning Variables which acts as a padding to input vector. Conditioning augmentation makes a latent vector which is then fed as an input to generator.

C. Generator and Discriminator:

In our proposed system the function V has two terms which calculates entropy value based on real data and another based on random data. The real distribution is used by discriminator because it has the access of training sample and it will try to set it's value to 1. The value 1 indicates the image taken is

$$\min_{G} \max_{D} V(D, G)$$

$$V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z))]$$

Fig. 2: Min Max function

resembling the given description and since we pass the training sample to Discriminator, then it must compulsorily have a value of 1. The second and random data is passed to generator. The generator generates fake images in initial iteration which is fed to discriminator to evaluate whether image generated is real or not. Initially the value of random data is 0 since image is assumed to be not real. This is kind of min-max function where Generator will try to minimise the loss while discriminator will try to maximize the loss which gives rise to good images.

The Generator also takes Random Gaussian Noise in earlier stage. It consists of a Convolution Neural Network which generates the images. The discriminator competes with generator because it contains all the real images as a training sample. The Generator generates pictures which it assumes to be real while discriminator assumes every picture to be not real unless it meets stopping condition. At the end point when discriminator is not able to label the generated image by generator as fake furthermore, then the actual image is generated. At each point the difference between the images of Generator and Discriminator is calculated using K L divergence.

D. Kullback Leibler divergence:

KL Divergence is the loss metric which states the similarity between two probability distributions by calculating and analysing the difference among entropy and cross-entropy. It forces the latent variables distribution to convert into a normal distribution. In the further iteration the input fed up to the Generator makes up for the loss in the previous training. Thus at the end, the image is generated which somewhat resembles the textual description.

VI. RESULTS

We trained the model till sixth depth and found that on increasing the depth, resolution of the images also increases as can be seen in below mentioned figures. However, we could not increase the resolution of images beyond 128*128 as the GPU was exhausted by this GAN implementation. We are generating 16 images of 128*128 resolution. The reason behind that is that some features are taken into consideration in a particular image and some features in other image but all the features are not depicted together in a single generated image. Our study revealed that depth is proportional to the resolution of images. Depth increases the complexity of the model. In the initial depths, the model training time was faster compared to the later stages as the resolution increases on increasing depth. These are best results that we were able to obtain based on our

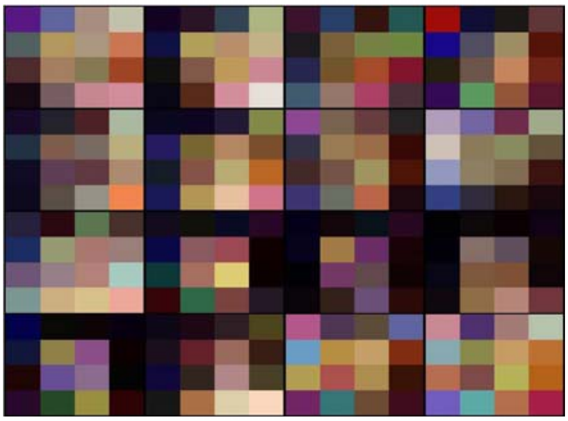

Fig. 3: Depth 1

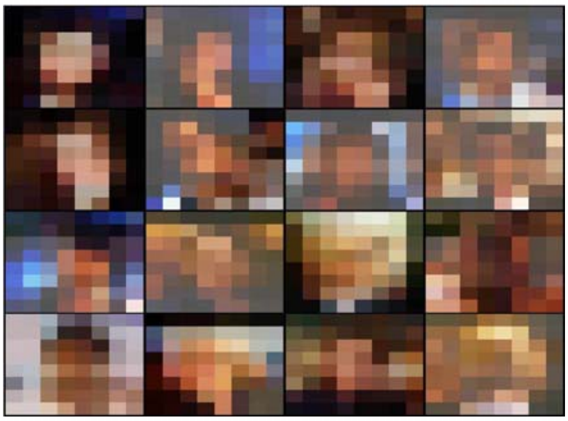

Fig. 4: Depth 2

hyper parameter tuning. The figures below show the progress that we got in each depth in the best training that we achieved and the images generated by using the best model on giving description as input. We have trained the model on Tesla T40 GPU of 12GB and the training process took more than 8 hours for minimum 5 depths i.e resolution of 64*64.

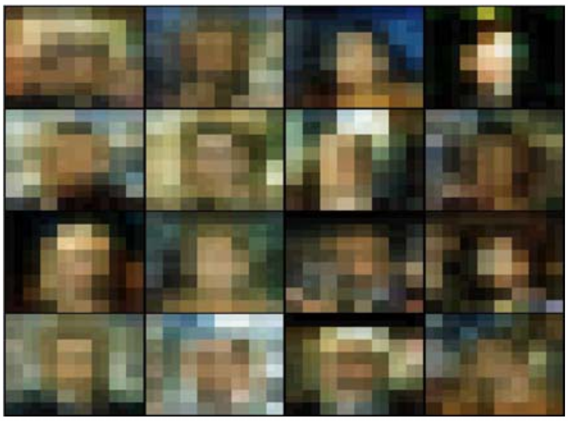

Fig. 5: Depth 3

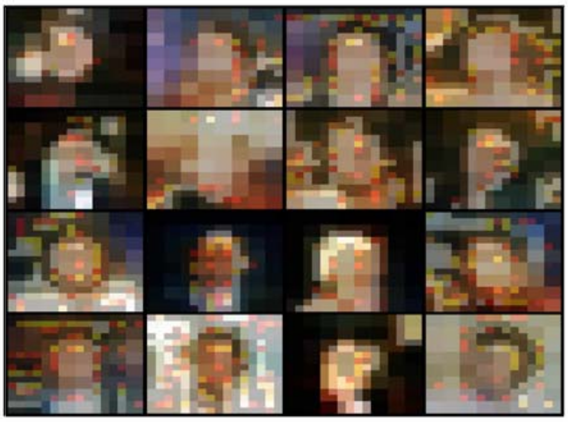

Fig. 6: Depth 4

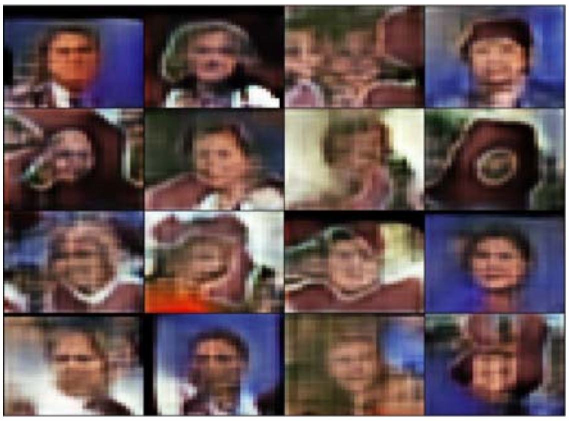

Fig. 7: Depth 5

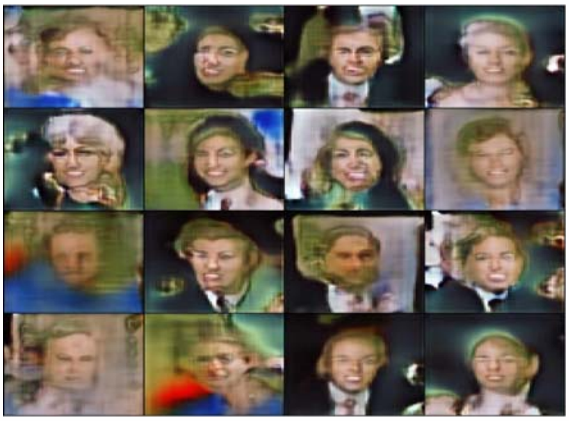

Fig. 8: Depth 6

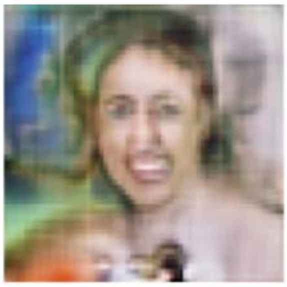

Fig. 9: thin woman in her late 2 0 s with gentle facial features and without heavy makeup . her long brown hair is pulled up at to top only , giving her a more casual and relaxed look . she seems to be at ease with what she is doing .

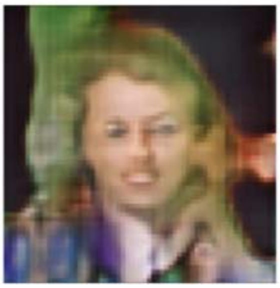

Fig. 10: a young woman with long light brown hair cut in layers and parted in the middle , a small nose and a nice open smile

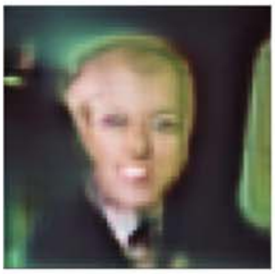

Fig. 11: a serious man , most probably politically involved . whiting hair with intense eyes . smart and looks important

VII. Conclusion

Despite the low resolution images generated, this method can be used for criminal identification with substantial amount of data to work upon. From the 16 images generated for one description the suspect might resemble to one of the faces. This can also act as a base layer for sketch artist to draw images from description. The witness can in real time say within minutes whether the suspect resembled the generated picture or not. On matching the image with the ones in a criminal database can lead to quick identification of criminals if they are present in the database.

VIII. Limitations and Future Scope

The limitation of this model is that the resolution of images generated is quite low due to relatively small number of images and of low resolution in dataset. Also, higher resolution images require more processing, so high GPU is required. There is a good amount of scope in this method. The creation of a dataset consisting of high resolution images with their annotations and using it for our model can give better results. Also if we could manage to have more annotated faces in a data set it could have given better results. As new type of GAN models are increasingly developed by making changes in the basic traditional GAN architecture, there is a possibility that we get better results by using some new GAN model. After getting better quality and resolution in the generated images, we can superimpose the generated 16 images to generate an individual image taking into account the features from these 16 images.

IX. Acknowledgment

The work done in this implementation wouldn't have been possible without Prof Albert Gatt of University of Malta who provided us with the dataset of 400 images from Labelled Face in the wild with it's annotation. We sincerely thank them for the same.

References

[1] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris N. Metaxas, 'StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks' *rXiv:1710.10916v2 [cs.CV] 25 Dec 2017.*

[2] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, Dimitris Metaxas, 'StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks' *2017 IEEE International Conference on Computer Vision (ICCV).*

[3] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, 'Progressive Growing of GANS for Improved Quality, Stability, and Variation' *ICLR 2018.*

[4] Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth P Camilleri, Michael Rosner, Lonneke van der Plas, 'Face2Text: Collecting Annotated Image Descrip-tion corpus for generating,' *arXiv:1803.03827v1 [cs.CL]*

[5] Chenrui Zhang, Yuxin Peng, 'Stacking VAE and GAN for Context-aware Text-to-Image Generation,' *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*

[6] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, Anil A Bharath, 'Generative Adversarial Networks: An Overview' *arXiv:1710.07035 19 Oct 2017*

[7] Sosuke Kobayashi, 'Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations' *arXiv:1805.06201 16 Oct 2018*

[8] Zhang, Zhaoxiang, David Suter, Yingli Tian, Alexandra Branzan Albu, Nicolas Sidère, and Hugo Jair Escalante, 'Pattern Recognition and Information Forensics' *ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers. Vol. 11188. Springer, 2019*

[9] Behmer, Ernst-Josef, Krishna Chandramouli, Victor Garrido, Dirk Mühlenberg, Dennis Müller, Wilmuth Müller, Dirk Pallmer, Francisco J. Pérez, Tomas Piatrik, and Camilo Vargas,'Ontology Population Framework of MAGNETO for Instantiating Heterogeneous Forensic Data Modalities" *IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 520-531. Springer, Cham, 2019*

[10] Xing Di · Vishal M. Patel,'Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs' *arXiv:1801.00077v1 [cs.CV] 30 Dec 2017*

[11] J. Yang, A. Kannan, D. Batra, and D. Parikh,'LR-GAN: layered recursive generative adversarial networks for image generation' *arXiv:1703.01560v3 [cs.CV] 2 Aug 2017*

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, 'Generative adversarial nets' *arXiv:1406.2661v1 [stat.ML] 10 Jun 2014*

[13] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. 'Generative adversarial text-to-image synthesis' *arXiv:1605.05396v2 [cs.NE] 5 Jun 2016*