INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND APPLICATIONS
VOLUME XVI, ISSUE IV, APRIL 22, WWW.IJCEA.COM ISSN 2321-3469

REALISTIC FACE IMAGE GENERATION BASED ON GAN

Purna Nandiboina [1], Akshata Salian [1], Shweta Akhadmal [1], Megha V. Gupta [2]

[1]Research Scholar, [2]Vice Principal, Department of Computer Engineering, New Horizon Institute of Technology and Management, University of Mumbai, Thane, India.

ABSTRACT:

Text to face generation is a sub-domain of text to image synthesis. It has a huge impact on new research areas along with the wide range of applications in the public safety domain. Most of the work for text to face generation until now is based on the partially trained generative adversarial networks, in which the pre-trained text encoder has been used to extract the semantic features of the input sentence. Later, these semantic features have been utilized to train the image decoder. The proposed system will be a fully trained generative adversarial network to generate realistic and natural images. The system will train the text encoder as well as the image decoder at the same time to generate more accurate and efficient results. In addition to the proposed methodology, another contribution is to generate the dataset by the amalgamation of LFW and CelebA datasets. The dataset has also been labeled according to our defined classes. The system will create a model – a discriminator network and a generator network by eliminating the fully connected layer in the traditional network and applying batch normalization and deconvolution operations. The proposed work also presents the details of the similarity between the generated faces and the ground-truth input description sentences.

Keywords: GAN, CNN, text to face, image generation, face synthesis, legal identity for all.

[1] INTRODUCTION

In recent years, advances in generative machine learning techniques, in particular with image generation using generative adversarial networks (GANs), have shown impressive results. While exciting, many of these results have little application beyond art or novelty. Similarly, there have been great improvements in computer natural language understanding through the use of deep recurrent neural networks. We plan to take advantage of these breakthroughs in image generation and

natural language understanding by combining them to take as input a textual description of a face and output a set of photo-realistic image interpretations of the text.

The initial inspiration for our system was that it may be used to help identify suspects in a police investigation. Often a victim or eyewitness description of the suspect will be relayed to a sketch artist who will then synthesize that information to draw a sketch of what the suspect looks like. However, this is time consuming for the eyewitnesses and artists. A software system built around a model trained to generate an array of images based on unstructured text input would enable witnesses to more easily relay their description of the suspect and police to more easily begin their process of searching for or identifying the suspect. In addition to increased speed of suspect facial generation, other benefits of a software system include ameliorated costs to police departments and the possibility to generate an array of interpretations of the description rather than relying on a single interpretation from a single artist.

Even with multiple interpretations from multiple artists, a software system may more cheaply (with regard to time and other resources) generate more interpretations for police to analyze. Our model could potentially be more accurate than the sketch. It is also easier to recognize a person based on a digital image than a drawing. One may additionally imagine non-criminal scenarios in which one wishes to search for an individual only by their facial features. In addition to person identification and search, a range of scenarios (artistic and otherwise) photorealistic face generation based off text input may add value to people's lives. We evaluate the results of our project using various key performance indicators: photo-realistic accuracy and clarity of the image and accessibility of the system. Accuracy will be based entirely on the system we will build, and is the main motivation behind our project. Clarity will be a secondary objective that measures our improvements of previous iterations. Accessibility will be based on how we can demonstrate our project by making our research understandable to the public in addition to how easily one may use the system to generate image.

[2] DEVELOPMENT OF THE RESEARCH

[2.1] LITERATURE SURVEY

For deep learning algorithms, in order to understand the input data, they need to learn to create the data. The most promising approach is to use a generation model that learns to discover the rules of the data and find the best distribution to represent it. In addition, by learning to generate models, we can even draw samples that are not in the training set but follow the same distribution.

As a new generation model framework, the Generative Adversarial Network [1] proposed in 2014 can generate a composite image that is better than the previous generation model, and has since become one of the most popular research fields. The Generative Adversarial Network includes two neural networks, a generator and a discriminator, wherein the generator attempts to generate a real sample to spoof the discriminator, and the discriminator attempts to distinguish between the real sample and the generated sample from the generator. Generating images is by far the most widely studied area of GAN. The main methods of GAN in image generation are hierarchical methods, iterative methods and direct methods. The algorithm under the hierarchical approach uses two generators and two discriminators in its model, where different generators have different purposes, and the relationship between the two generators can be parallel or in series. In short, two neural networks play a minimax game in which one model, the discriminative model, learns to discern whether a sample is from the training data distribution or generated by the second mode, the

generative model that learns to generate samples of a distribution [1]. The "competition" between the models should lead towards images generated by the generative model that are indistinguishable from samples of the training data.

Goodfellow et. al. shows many current methods of generating samples of a distribution have limitations in either computational methods or variability in types of inputs [1]. The authors argue that the ability to use deep neural networks enables taking advantage of advancements made with deep neural network training in general. For example, generative adversarial networks are very effective at generating images. One of the main motivating factors is that, while deep learning had seen great progress in classification problems, generative problems had seen relatively limited progress. We are choosing to use generative adversarial networks because they are so powerful and there has been much improvement on them since the original paper in 2014.

There are two large scale datasets which are publicly available for face synthesis task. These datasets are the CelebA [2] and LFW [3]. Most of the state of the artwork has tested their model capabilities and abilities for face synthesis using the GAN and conditional GAN. DCGAN [4], CycleGAN [5], Pro-GAN [6], BigGAN [7], StyleGAN [8], StarGAN [9] are the examples of this problem. The quality of the generated face images is improving day by day with the development in the generative adversarial networks. Some of the networks can generate good quality face images with a size of 1024 × 1024. These face images are much larger than the original images present in the face dataset. These described models first learned through the noise vector with the help of mapping and followed the normal distribution to generate the natural images of the face. However, they are not able to generate an accurate and precise face based on the input description.

To overcome and tackle this problem, many researchers have worked on different directions of face synthesis. These directions include converting the face edges into the natural face images [10], swapping the facial attributes of two different face images [11], generating the face with the help of the side face [12], generating the face with the help of the human eye's region [13], draw sketches from the human face [14], face make-up [15] and many more. But as per our best knowledge, no one combined the different face-related information in a single methodology to generate the natural and realistic face images.

Some of the researchers have also worked on the face generation through the attribute's description. Li et. al [16] proposed the work, in which they generated the face with the help of the attribute description by making sure that they preserve the identity of the face. The drawback of their proposed methodology is that it is only applicable to those faces which can be generated using the simple attributes. Another work named TP-GAN [12] has been proposed by the researchers.

In this work, they have proposed the generative adversarial network based on the two pathways. They synthesized the frontal face images using the proposed network. Although they succeeded to generate the good results but required a large amount of labeled data of frontal faces. Some of the researchers have also explored the disentangled representation learning for face synthesis using the defined attributes of the face. DC-IGN [17] has proposed the variational auto-encoder using the patterns and techniques of disentangled representation learning. However, the major drawback of this work is that it only tackles one attribute in particularly one batch. It makes it computationally weak as well as it also requires the large explicitly annotated data for training. Luan et. al [18] proposed the algorithm, which they named as the DR-GAN. It is used for the learning purpose of generative and discriminative representation of face synthesis.

Their proposed work was based on the poses of the face and did not focus on specified face attributes. However, our proposed framework makes sure to preserve the identity of the generated

image by incorporating all the attributes information related to the face. As per our best knowledge and based on the literature survey, the work on the face generation through the attribute description using the generative adversarial network is very less. Most of the work on this problem is done on the limited scope and failed to generate impressive results by not preserving the face identity. Moreover, most of the relevant proposed networks have trai-

[2.2] PROPOSED SYSTEM

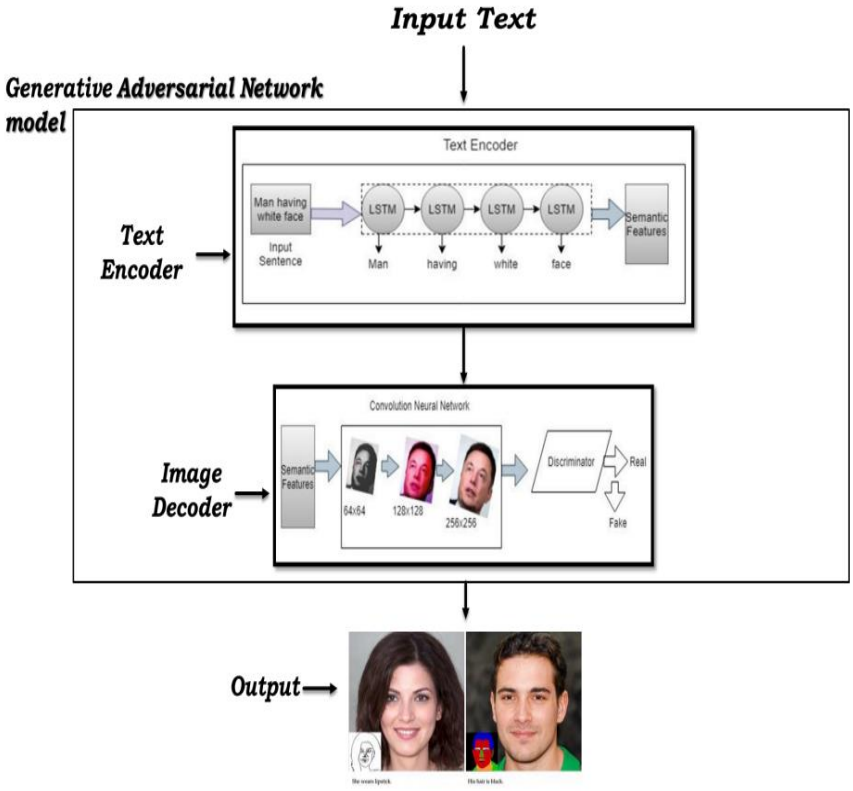


Figure: 1. Proposed System Architecture.

-ned the image decoder and used the [19]. So, in this work we have proposed the fully trainable generative adversarial network.

The proposed system will be able to generate realistic face image using the text description given by the user. The system will take the text description from the user. This is description is entered into the system using the text encoder. Each word present in the sentence is connected with the two hidden states. Each hidden state corresponds to the one direction. The outputs of these two hidden states have been concatenated to get the semantic meaning against each word of the sentence. These extracted semantic features from the description is passed on to the database. The database will use the semantic features and the image matched with those semantic features are fed as input to the image decoder network. The image decoder will generate an image using the image sent by the database. So, we have total 3 blocks and 9 deconvolution layers, which up-sample feature map twice to its original size. The layers present into the blocks take the input from the encoded features of text as semantic vectors and generates realistic images.

In the first stage, the semantic vectors are extracted from the text along with the noise concatenation and are passed as an input, and then this input vector is reduced to the 4 x 4 feature map. The up-sampling on the feature map increases the size twice the feature maps in all three layers

of each block. So, the size of the feature map is up-sampled to 8x8, 16x16 and 32x32. The up-sampling block helped in fine-tuning the training parameters. The input to the second block is the feature map with a size of 64x64. The second block contains the

same de-

Table 1. Quantitative Comparison of Text-to-Image Generation. We use FID, LPIPS, accuracy (Acc.), and realism (Real.) to compare the state of the art and our method on the proposed Multimodal CelebA-HQ dataset. ↓ means the lower the better while ↑ means the opposite.

n-th	Attribute	n-th	Attribute
1	eye glasses	7	hair color
2	head pose	8	face color
3	face shape	9	age
4	Hair length, nose, lip	10	gender
5	cheekbones	11	micro features
6	chin	12	micro features

Table 2. The Empirical Layer Wise Analysis of a 14-layer Style-GAN Generator. The 13-th and 14-th layers are omitted since there is basically no visible difference.

-

Method	FID ↓	LPIPS ↓	Acc. (%) ↑	Real. (%) ↑
AttnGAN	45.56	0.512	14.2	20.3
FTGAN	44.49	0.522	18.2	22.5
Our model	42.50	0.456	25.3	31.7

convolution layers the same as the first block and outputs the 128x128 feature map. Whereas, in the 3rd block, we get the 256 x 256 feature map using the same layer architecture which was previously used in the first two blocks. After the three blocks of up-sampling layers, we have generated the 256 x 256 image. This output is further passed to the discriminator network, to find the effectiveness of the generated face features.

This generated image along with its sentence encoding is passed to the discriminator CNN network that extracts the low-level region features using attention mechanism to be compared with ground truth image. The attention layer is added that allows convolution layers of discriminator to attend the region-based features of eyes, nose, lips as well as entire facial features. Here the discriminator will measure the realness of human face region as well as the face features. The resultant realistic face image is then sent to the user which is derived from the text description given by the user.

[2.3] RESULT ANALYSIS

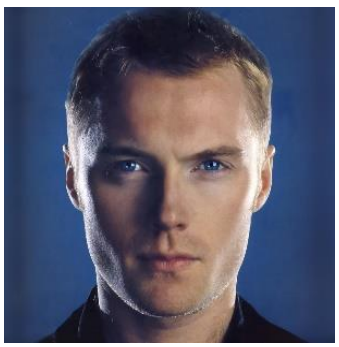
Quantitative Comparison: - In our experiments, we evaluate the FID and LPIPS on a large number of samples gen generated from randomly selected text descriptions. To evaluate accuracy and realism, we generate images from 50 randomly sampled texts using different methods. In a user study, users are asked to judge which one is the most photorealistic and most coherent with the given texts. The results are demonstrated in Table 2. Compared with the state-of-the-arts, our method achieves better FID, LPIPS, accuracy, and realism values, which proves that our methods can generate images with the highest quality, diversity, photorealism, and text-relevance.

Evaluation Metric: - For evaluation, there are four important aspects: image quality, image diversity, accuracy, and realism. The quality of generated or manipulated images is evaluated

through Frechet Inception Distance (FID). The diversity is measured by the Learne-

This woman wears earrings. She has oval face and high bones. She is smiling.

He has no beard.



He is old. He has beard. He has big nose.

Fig 2. Qualitative Comparison of Image Manipulation using Natural Language Descriptions.

-d Perceptual Image Patch Similarity (LPIPS). For image generation, the accuracy is evaluated by the similarity between the text and the corresponding generated image. For manipulation, the accuracy is evaluated by whether the modified visual attributes of the synthetic image are aligned with the given description and text-irrelevant contents are preserved. The accuracy and realism are evaluated through a user study, where the users are asked to judge which, one is more photo-realistic, and more coherent with the given texts. We test accuracy and realism by randomly sampling 50 images with the same conditions.

Layer wise Analysis: - The pre-trained StyleGAN we used in most experiments is to generate images of 256 × 256 (i.e., size 256), whose has 14 layers of the intermediate vector. For a synthesis network trained to generate images of 512 × 512, the intermediate vector would be of shape (16,

512) (and (18, 512) for 1024 × 1024), where the number of the layers L is determined by 2 log2 R−2 and R is the image size. In general, layers in the generator at lower resolutions (e.g., 4 × 4 and 8 × 8) control high-level styles such as eyeglasses and head pose, layers in the middle (e.g., as 16 × 16 and 32 × 32) control hairstyle and facial expression, while the final layers (e.g., 64 × 64 to 1024 × 1024) control color schemes and fine-grained details. Based on empirical observations, we list the attributes represented by different layers of a 14-layer StyleGAN in Table 2. The layers from 11- 14 represent micro features or fine structures, such as stubble, freckles, or skin pores, which can be regarded as the stochastic variation. High-resolution images contain lots of facial details and cannot be obtained by simply up sampling from the lower-resolutions, making the stochastic variations especially important as they improve the visual perception without affecting the main structures and attributes of the synthesized image.

Qualitative Comparison: - Most existing text-to-image generation methods, as shown in Fig 2, can generate photo-realistic and text-relevant results. However, some attributes contained in the text do not appear in the generated image, and the generated image looks like featureless paint and lacks details. This “featureless painterly” look would be significantly obvious and irredeemable when generating higher resolution images using the multi-stage training methods. Furthermore, most existing solutions have limited diversity of the outputs, even if the provided conditions contain different meanings. For example, “has a beard” might mean a goatee, short or long beard, and could have different colors.

[3] CONCLUSION AND FUTURE WORK

We have proposed a novel method for image synthesis using textual descriptions which achieves high accessibility, diversity, controllability, and accurateness for facial image generation and manipulation. Through the proposed multi-modal GAN inversion and large-scale multi-modal dataset, our method can effectively synthesize images. This proposed work has a huge impact on security related domains like forensic analysis and public safety domain etc. Also, as generating images from text is the opposite process of image captioning and image classification, where text and caption are generated from images. Just like the image captions, text to image generation helps to find context and relationship between the image and the text along with exploring human visual semantics.

Finally, we demonstrated why inception distance used to measure the performance of GANs [1] fails to evaluate their performance on our dataset.

We plan on extending the work in the following directions: -

1) Improve the selection of the wrong image. Currently, we randomly select images from the dataset as wrong image. One possibility is to select the wrong caption for real image rather than selecting the wrong image. This could be done by selecting the caption having the lowest cosine similarity with the caption of the real image.
2) Propose a better evaluation metric to capture the semantic similarity of the generated faces with their captions, without using the classes.
3) Improving the resolution of the generated faces e.g., 128 × 128 and 256 × 256 faces.

REFERENCES

[1] (Goodfellow, 2014) Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3. 10.1145/3422622.

[2] (Yandong Guo, 2016) MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9907. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_6.

[3] (Huang, 2008) Huang, Gary & Mattar, Marwan & Berg, Tamara & Learned-Miller, Eric. (2008). Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Tech.

[4] (Ziwei Liu, 2015) Liu, Ziwei, Ping Luo, Xiaogang Wang and Xiaoou Tang. “Deep Learning Face Attributes in the Wild.” 2015 IEEE International Conference on Computer Vision (ICCV) (2015): 3730-3738.

[5] (Mehdi Mirza, 2014) Mirza, Mehdi and Simon Osindero. “Conditional Generative Adversarial Nets.” ArXiv abs/1411.1784 (2014): n. pag.

[6] (Zisserman, 2008) Nilsback, Maria-Elena and Andrew Zisserman. “Automated Flower Classification over a Large Number of Classes.” 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing (2008): 722-729.

[7] (Tao, 2019) Qiao, Tingting, Jing Zhang, Duanqing Xu and Dacheng Tao. “MirrorGAN: Learning Text-To-Image Generation by Redescription.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 1505-1514.

[8] (Chintala, 2016) Radford, Alec, Luke Metz and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” CoRR abs/1511.06434 (2016): n. pag.

[9] (Reed, 2016) Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. & Lee, H.. (2016). Generative Adversarial Text to Image Synthesis. Proceedings of the 33rd International Conference on Machine Learning, in Proceedings of Machine Learning Research 48:1060- 1069 Available from https://proceedings.mlr.press/v48/reed16.html.

[10] (Ting-Chun Wang, 2018) Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8798-8807.

[11] (Hua, 2018) Bao, Jianmin, Dong Chen, Fang Wen, Houqiang Li and Gang Hua. “Towards Open-Set Identity Preserving Face Synthesis.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 6713-6722.

[12] (He, 2017) Huang, Rui, Shu Zhang, Tianyu Li and Ran He. “Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis.” 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 2458- 2467.

[13] (Peng, 2018) Chen, Xiang, Linbo Qing, Xiaohai He, Jie Su and Yonghong Peng. “From Eyes to Face Synthesis: a New Approach for Human-Centered Smart Surveillance.” IEEE Access 6 (2018): 14567-14575.

[14] (Patel, 2018) Di, Xing and Vishal M. Patel. “Face Synthesis from Visual Attributes via Sketch using Conditional VAEs and GANs.” ArXiv abs/1801.00077 (2018): n. pag.

[15] (Jiahui Yu, 2018) Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, Thomas S. Huang. Generative Image Inpainting with Contextual Attention. 10.1109/CVPR.2018.00577.

[16] (Zhang, 2016) Li, Mu, Wangmeng Zuo and David Zhang. “Convolutional Network for Attribute-driven and Identity-preserving Human Face Generation.” ArXiv abs/1608.06434 (2016): n. pag.

[17] (Tenenbaum, 2015) Kulkarni, Tejas D., William F. Whitney, Pushmeet Kohli and Joshua B. Tenenbaum. “Deep Convolutional Inverse Graphics Network.” NIPS (2015).

[18] (Liu, 2017) Tran, Luan, Xi Yin and Xiaoming Liu. “Disentangled Representation Learning GAN for Pose-Invariant Face Recognition.” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1283-1292.

[19] M. Z. Khan et al., "A Realistic Image Generation of Face from Text Description Using the Fully Trained Generative Adversarial Networks," in IEEE Access, vol. 9, pp. 1250-1260, 2021, doi: 10.1109/ACCESS.2020.3015656.

[20] Li, Bowen & Qi, Xiaojuan & Lukasiewicz, Thomas & Torr, Philip. (2020). ManiGAN: Text-Guided Image Manipulation. 7877-7886. 10.1109/CVPR42600.2020.00790.

Author[s] brief Introduction

Purna Nandiboina

Purna Nandiboina is a Computer Engineering Research Scholar at New Horizon Institute of Technology and Management, Thane. Her research interests include Machine Learning, Deep Learning and Artificial Intelligence. She is the member of CSI.

Akshata Salian

Akshata Salian is a Computer Engineering Research Scholar at New Horizon Institute of Technology and Management, Thane. Her research interests include Machine Learning, Deep Learning and Front-end Development. She is the member of CSI.

Shweta Akhadmal

Shweta Akhadmal is a Computer Engineering Research Scholar at New Horizon Institute of Technology and Management, Thane. Her research interests include Deep Learning.

Megha V. Gupta

Mrs. Megha V Gupta is a Computer Engineering Research Scholar at the Research Centre, Datta Meghe College of Engineering in Airoli, Navi Mumbai. She is the Vice-Principal at the New Horizon Institute of Technology and Management, Thane, besides being an Assistant Professor at the Dept. of Computer Engineering. Her research interests include Artificial Intelligence and Machine Learning. She has also published a patent and over 15 published research papers in international journals/conferences. Mrs. Gupta pursued her engineering education at the University of Mumbai and RTM Nagpur University from where she has ME, Computer Engineering, and B.E Computer Technology degrees respectively. She is a member of IEEE and CSI.