Text-to-Face Image Generation

Jongho Han and Gyu Sang Choi

Department of Information and Communication Engineering

Yeungnam University

Gyeogsan, Republic of Korea

Whdgh920423@gmail.com, castchoi@ynu.ac.kr

Abstract— **Active research on text-to-image continues to this day. Most of them focus on improving image quality and semantic consistency between text and image, based on well-established datasets related to animals and objects. However, little research has been done on creating text-to-face images. Therefore, in this paper, we are proposing a model that creates face images from text using ProGAN[2]. In future, we plan to include Korean texts by building such a dataset.**

Keywords-component; Text-to-face; Deep Learning; LSTM; GAN;

I. INTRODUCTION

An interesting field of text-to-image creation has been actively researched and developed since the beginning with StackGAN[3]. Several new models appear and the image quality of the generated images is very good, and recently, attention has been paid to the semantic correspondence between images and text. In addition, the common datasets such as MSCOCO[10], CUB[9], and Oxford-102[11] are well-equipped with data that can be used as an indicator of performance evaluation. However, most of them are natural images or datasets about objects and animals, and datasets on human faces are extremely limited and not well established. Of course, there is celebA[5] as the public data for image compositing except text, but there is not enough research for text-to-face. Therefore, in this paper, we plan to implement the korea language text-to-face image generation model for the first time by creating face image using Face2Text dataset (contain 400 image sample and each face image-5 descrption pair) and constructing Korean dataset.

II. RELATED WORK

The text-to-image generation model has been studied as follows. For the first time, conditional GAN[1] was introduced as an encoder-decoder-based framework, followed by StackGAN, which is currently the most basic. This model consists of two stages of GAN stage after text embedding, Created an image. Later, AttnGAN[7], composed of the attentional generative network and the deep attentional multimodal similarity model, was published and its performance was superior to previous studies. In addition, related studies such as stackGAN++[4], MirrorGAN[8], DCGAN[6], and FTGAN[12] have been published. In this paper, FTGAN, the most recent paper in face image synthesis, was heavily influenced. FTGAN build a dataset SCU-Text2face based on CelebA, which contains 1000 images. For each face images in SCU-Text2face, there are five descriptions given by different persons. It also showed better performance than the existing models. It will be a good baseline for building Korean dataset in the future.

III. PROPOSAL METHOD

The basic structure is a model introduced in the T2F on Github (https://github.com/akanimax/T2F) repository and consists of ProGAN and LSTM network.

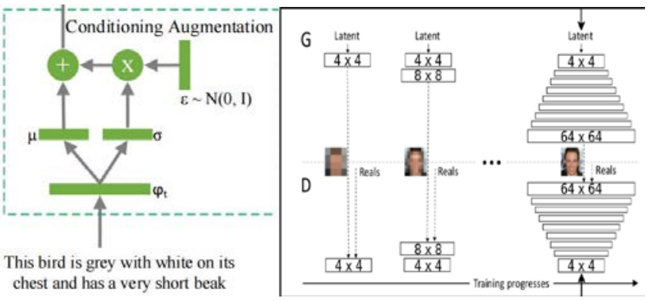
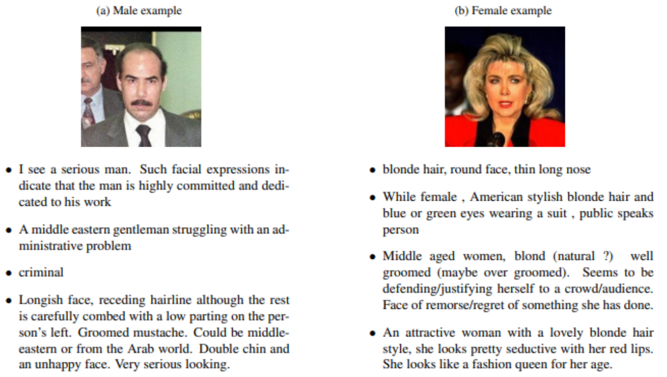

Conditioning Augmentation
ε ~ N(0, I)
μ
σ
φt
This bird is grey with white on its chest and has a very short beak
G
Latent
4 x 4
8 x 8
64 x 64
D
Reals
Training progresses

Figure 1. Overall structure for models based on ProGAN and LSTM networks model.

First, text descrption is encoded and embedded via the LSTM network. The random variable c is then randomly sampled from the independant Gaussian dstribution N in the Conditioning Augmentation step. This is useful for converting text into an image and allowing the same sentence to have various arbitrary expressions. The text latent vector thus generated enters the input of the GAN and gradually increases the resolution of the generated image by gradually adding layers to the generator and discriminator as the training proceeds.

IV. Experiments

In this paper, we experiment with creating a simple face image and show information about the dataset.



(a) Male example

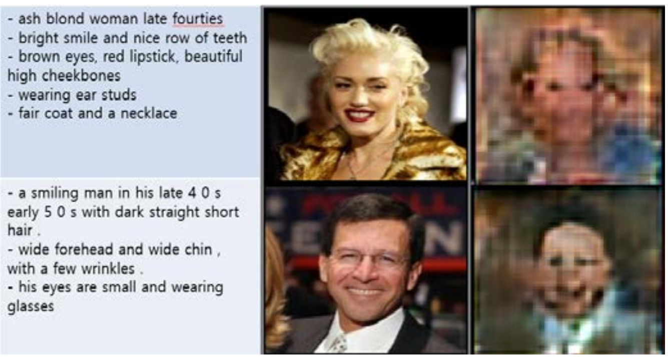
- I see a serious man. Such facial expressions indicate that the man is highly committed and dedicated to his work
- A middle eastern gentleman struggling with an administrative problem
- criminal
- Longish face, receding hairline although the rest is carefully combed with a low parting on the person's left. Groomed mustache. Could be middle-eastern or from the Arab world. Double chin and an unhappy face. Very serious looking.

(b) Female example

- blonde hair, round face, thin long nose
- While female , American stylish blonde hair and blue or green eyes wearing a suit , public speaks person
- Middle aged women, blond (natural ?) well groomed (maybe over groomed). Seems to be defending/justifying herself to a crowd/audience. Face of remorse/regret of something she has done.
- An attractive woman with a lovely blonde hair style, she looks pretty seductive with her red lips. She looks like a fashion queen for her age.

Figure 2. Information about the dataset[13] used to train the model (400 image, 5 descriptions on average for each image)

Figure 2 shows the dataset used to create the text-to-face image. It consists of 400 images and five descriptions on average for each image. As there is no data currently built, there was a limit to selecting good performance.



- ash blond woman late fourties
- bright smile and nice row of teeth
- brown eyes, red lipstick, beautiful high cheekbones
- wearing ear studs
- fair coat and a necklace

- a smiling man in his late 4 0 s early 5 0 s with dark straight short hair .
- wide forehead and wide chin , with a few wrinkles .
- his eyes are small and wearing glasses

Figure 3. As a result of the experiment, the actual explanation from the left, real image, generated image

Therefore, it was difficult to expect the ability to catch only a certain shape like figure 3 and distinguish people properly. In addition, the image of the image is also much lower than the existing studies, so not only do we need to construct new data, but also modify the model.

V. Conclusion

The field of text-to-image generation itself is a very interesting topic and is very likely for future development. Research on the new image creator itself has been steadily being published with high quality models. In line with this period, this paper proposed a model for text-to-face image generation. Although the field is still under study, the model itself is the basis of many existing studies, and its performance is much lower than the papers published so far. In the future, however, we will increase the 400 datasets by 1000 to produce better quality images, and we will continue to update the GAN model. In addition, we will build a new dataset and modify it to a model that can be embedded in Hangul.

Acknowledgment

This research was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program. No.10063130 and MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2019-2016-0-00313) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

[1] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv: Learning, 2014.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In international conference on learning representations, 2018.

[3] H. Zhang, T. Xu, and H. Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In international conference on computer vision, pages 5908–5916, 2017.

[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2018.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In international conference on computer vision, pages 3730–3738, 2015.

[6] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In international conference on learning representations, 2016

[7] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In computer vision and pattern recognition, pages 1316–1324, 2018.

[8] T. Qiao, J. Zhang, D. Xu, and D. Tao. Mirrorgan: Learning text-toimage generation by redescription, 2019.

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. J. Belongie. The caltech-ucsd birds-200-2011 dataset. Advances in Water Resources, 2011.

[10] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft coco: Common objects in context. In european conference on computer vision, pages 740–755, 2014.

[11] M. E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In Conference on Computer Vision, Graphics & Image Processing, 2009.

[12] Xiang Chen1 , Lingbo Qing1 , Xiaohai He1 , Xiaodong Luo1 , Yining Xu1. FTGAN: A Fully-trained Generative Adversarial Networks for Text to Face Generation. 2019

[13] A. Gatt, M. Tanti, A. Muscat, P. Paggio, R. A. Farrugia, C. Borg, K. P. Camilleri, M. Rosner, and L. V. Der Plas. Face2text: Collecting an annotated image description corpus for the generation of rich face descriptions. language resources and evaluation, 2018.