Text-to-Face Generation with StyleGAN2

D.M.A. Ayanthi* and M.K.S. Madushika

Department of Computer Science, Faculty of Science, University of Ruhuna, Sri Lanka
**Corresponding Author E-mail: ayanthi9896@usci.ruh.ac.lk, TP: +94713054568*

Synthesizing images from a text description has become an active research area with the advent of Generative Adversarial Networks. It is a flexible way of generating images in a conditioned environment and has made significant progress in the recent years. The main goals of these models are to generate photo-realistic images that are well aligned with the input descriptions. Text-to-Face generation is a sub-domain of Text-to-Image generation that has been less explored because of its challenging nature. This is difficult because facial attributes are less specifically mentioned in descriptions and also because they are complex and has a wide variety. Although few works have been done in this domain, it has a variety of applications like in the fields of criminal investigation. But still there is the need to improve the image quality and how well the generated images match the input description. In this paper, we propose a novel framework for text-to-face generation using the state-of-the-art high-resolution image generator, StyleGAN2. For this task it is required to learn the mapping from the text space to the latent space of StyleGAN2. We chose BERT embeddings to encode the input descriptions. The text embedding mapped to the latent space, in turn was input to the StyleGAN2 model to generate facial images. We train and evaluate our model on the Text2Face dataset containing descriptions with at most 40 attributes for the images in the CelebA dataset. Our novel framework generates photo-realistic images by adopting StyleGAN2 and also improves the semantic alignment with the use of BERT embeddings that better capture the content of the description and the perceptual loss calculated using a pretrained VGG16 model. In the initial training we obtained a FID score of 370.57, Face Semantic Distance of 25.57 and a Face Semantic Similarity score of -0.002. With further training we believe the images could be made more realistic and semantically matching the input description.

Keywords: Text-to-Image Synthesis; Text-to-Face Synthesis; StyleGAN2; High-resolution; Semantic Alignment