Project Proposal: Schubert and Erler

Highly Descriptive text-to-face Generation to Synthezise Authentic Faces (photofits for criminology purposes) via a GAN

Team

- Daniel Schubert, 16627792, s.schubert@campus.lmu.de
- Max Erler, 11749383, max.erler@campus.lmu.de

Problem description

There are already some papers doing text to face generation:

- Semantic Text-to-Face GAN-ST2FG
- Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions
- TediGAN: Text-Guided Diverse Face Image Generation and Manipulation
- TediGAN: Text-Guided Diverse Face Image Generation and Manipulation
- Zero-Shot Text-to-Image Generation (DALL-E)

We want to train and fine tune an existing model architecture from one of the papers above. The generated images should look authentic and match the level of "photofits" (or phantom images, in German "Phantombilder") of real police workers or lawyers. The input text should include descriptive criteria, e.g. pointy nose, bald, blue eyes, wide mouth, long eyebrows or curly hair.

The first part will be to classifiy / detect important attributes from the text – which is given as an accurate description (i.e. one or more sentences) – and embed them into the latent feature space.

The second part will be the generative part using a VAE or GAN to synthesize images from the textual embeddings.

This task is interesting due to the real use case, that if it works it could support the work of police workers. Also, pretrained networks are not available and training/ building upon existing algorithms and trying to achieve the same or a higher baseline is a challenge we look forward to.

Dataset

We want to use :

- celebA
- celebA HQ
- LSW

During the training we want to combine the datasets or compare each one with our chosen network. Also we want to do some research to find some datasets which maybe fit better to our task due to more

descriptive labels.

Approach

First, we want to use the model from the Semantic Text-to-Face GAN-ST2FG paper and re-implement it with Pytorch since it out performs all the other networks in their related works comparison.

Second, we want to train the unmodified model with the baseline configuration on the different datasets and also on a compostion of them. The goal is to reproduce the baseline results.

Third, we want to "optimize" or fine tune the hyperparameters to see if we can achieve a better score than the existing model.

Fourth, we want to rethink the model based on our observations and try to optimize it in respect to our task – generating authentic photofits.

Moreover, we want to adapt and optimize the textual embeddings to make them appropriate for face descriptions. If we find a better dataset with more accurate labels we want to adapt the text embeddings to those labels and also optimize the GAN as described above.

Evaluation and Expected Results

We will evaluate the baseline and our adapted networks with exiting metrics like FID, LPIPS, BRISQUE and Manipulative precision metric. Moreover we will try to compose a new metric that can evaluate the usability for criminology purposes.

Hardware

We have access to a private machine. We are not sure if it's sufficient. We consider the option to use Azure for Students or the cip pool.