CS5353 Final Project Proposal

1 Team Members

name	uid
Herbert Wright	u1331078
Aaron Schindler	u1117629

2 Description of Problem

Generative Adversarial Networks (GANs) [1] tend to require a lot of data. We wish to construct a network that can, given a few images of an unseen face, construct face images/deepfakes that are similar to that one, even if it has not seen that person before. [2] is an example of face swapping, but without the few-shot learning of new faces.

3 Why Deep Learning

Deep learning is the traditional method of generating and detecting deepfakes [3]. While some computer graphics methods can be used, they lack the same ability of deep learning architectures to capture and learn complex functions efficiently. Because the images generated by deepfake architectures are very realistic, most humans are not able to distinguish between real and fake images. We can use a different (or same in some cases) model to learn the subtle differences between real and fake images, like differences in noise or counts of pixel colors, to effectively decide if an image is real or not.

4 Description of Approach

Overview. We will train an encoder-decoder network using a discriminator network. The decoder and discriminator network will be conditioned on a latent variable that gives facial information. The discriminator will give two values corresponding to whether or not the reconstructed image is fake and whether the image corresponds to the latent variable it is conditioned on.

Networks. Let (W, H) be the size of images, n be a small integer denoting number of faces used for conditioning. Our Architecture will consist of 4 different networks:

1. face encoder network: $f_F : \mathbb{R}^{n \times W \times H} \to \mathbb{R}^{h_F}$.
   - input: n images of the same persons face
   - output: latent representation of that person's face
2. image encoder network: $f_I : \mathbb{R}^{W \times H} \to \mathbb{R}^{h_I}$
   - input: an image of a persons face.
   - output: latent representation of the image
3. image decoder network: $f_G : \mathbb{R}^{h_F + h_I} \to \mathbb{R}^{W \times H}$
   - input: latent representations of a face and an image
   - output: a reconstructed image decoded from the image encoder
4. discriminator network: $f_D : \mathbb{R}^{W \times H} \times \mathbb{R}^{h_F} \to [0, 1]^2$
   - input: an image and a latent vector representation of a face
   - output: two probabilities: (1) probability of image being fake (not original); (2) probability of image being a different person than the faces encoded into the latent vector

Here is a visual of the architecture:

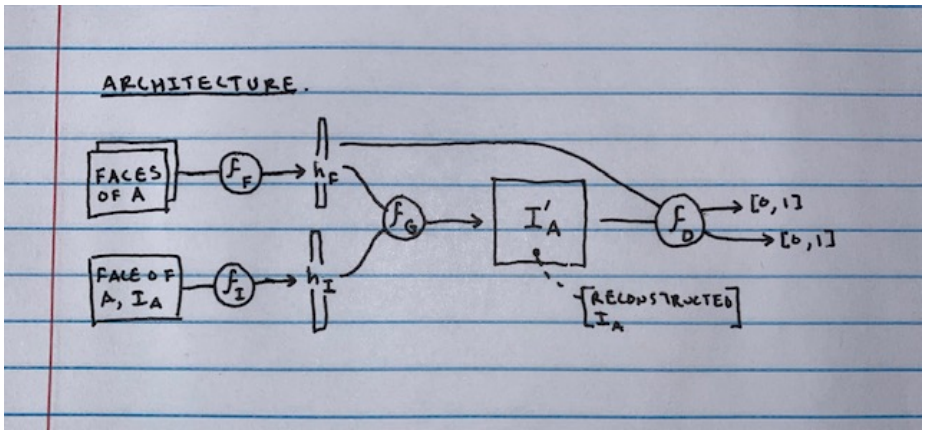

Figure 1: Our proposed architecture

Training. We plan to use the MS-Celeb-1M dataset [4]. Given a batch $\mathcal{B} = (F_A, I_A, I_B)$ (n faces of person A, N other faces of person A, and N faces of people other than A), we compute the following quantities:

1. $h_F = f_F(F_A) \in \mathbb{R}^{h_F}$
2. $h_I = f_I(F_A) \in \mathbb{R}^{N \times h_I}$
3. $h_B = f_I(F_B) \in \mathbb{R}^{N \times h_I}$
4. $I'_A = f_G(h_F, h_I) \in \mathbb{R}^{N \times W \times H}$
5. $I'_B = f_G(h_F, h_B) \in \mathbb{R}^{N \times W \times H}$
6. $(R_{A'}, C_{A'}) = f_D(I'_A, h_F) \in [0, 1]^{N \times 2}$

7. $(R_A, C_A) = f_D(I_A, h_F) \in [0,1]^{N \times 2}$
8. $(R_{B'}, C_{B'}) = f_D(I'_B, h_F) \in [0,1]^{N \times 2}$
9. $(R_B, C_B) = f_D(I_B, h_F) \in [0,1]^{N \times 2}$
10. $D_A = \|I_A - I'_A\|$
11. $D_B = \|I_B - I'_B\|$

We optimize f_D by maximizing R_A, R_B, C_A and minimizing $R_{A'}, R_{B'}, C_B$. We optimize f_F by maximizing $C_A, C_{A'}$, and potentially $C_{B'}$ and minimizing C_B and potentially D_A. We optimize f_G, f_I by maximizing $R_{A'}, R_{B'}, C_{A'}, C_{B'}$ and minimizing D_A, D_B. This will be done using the gradient descent algorithm.

Inference. We can use f_F, f_I, f_G to perform face swaps. We can use f_D to detect fake data vs real data and facial recognition given a face encoding h_F. We can use f_F, f_G to generate new images of a known face.

5 Research Papers

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] B.-S. Lin, D.-W. Hsu, C.-H. Shen, and H.-F. Hsiao, "Using fully connected and convolutional net for gan-based face swapping," in *2020 ieee asia pacific conference on circuits and systems (apccas)*, 2020, pp. 185–188.

[3] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, 2016, pp. 87–102.