Morpheus

A diffusion-based approach to morphing attack and defence systems

Aim

Innovate different methods for creating morphed images for diffusion autoencoders based face morphing attacks

Create a discriminator-based defence system that is immune to diffusion autoencoders based morphing attacks

Introduction

- What is face morphing?
- Where is it used?
- Face morphing techniques used in the past
- The MorDIFF technique

Experiment 1A– Generation using Interpolation

Methods:

- Using different interpolation techniques for creating the morphed images such as linear, spherical and quadratic interpolation.
- Passing these morphed images through SOTA Face Recognition models namely VGG-Face, FaceNet, FaceNet512, OpenFace.
- Using different evaluation metrics for the face recognition function, namely cosine, Euclidean distance and the Euclidean L2 norm.

Results:

Linear – Semantics and Spherical – Stochastic Embeddings

Distance Metric	VGG-Face	Facenet	Facenet512	OpenFace
cosine	3	3	3	2
euclidean	5	3	5	3
euclidean_l2	3	3	4	3

Spherical – Semantics and Linear – Stochastic Embeddings & Other

Distance Metric	VGG-Face	Facenet	Facenet512	OpenFace
cosine	8	4	5	2
euclidean	8	6	8	3
euclidean_l2	6	4	8	3

Figure 1: Different interpolations between two images passed through the various face recognition models. Each row's first and last images are the original images used as the base image in the face recognition model. The interpolations from top to bottom are (a)linear-spherical, (b)spherical-linear, (c)linear-linear, (d)spherical-spherical, (e)linear-quadratic, (f)quadratic-spherical, respectively.

Experiment 1B – Three Face Attack

Methods:

- The three face attack is performed by using a linear interpolation for semantic and spherical interpolation for stochastic encodings.

$$Lerp(\mathbf{z}_s^1, \mathbf{z}_s^2, \mathbf{z}_s^3; \lambda_1, \lambda_2) = \lambda_1 \mathbf{z}_s^1 + \lambda_2 \mathbf{z}_s^2 + (1 - \lambda_1 - \lambda_2)\mathbf{z}_s^3$$

$$SLerp(\mathbf{x}_t^1, \mathbf{x}_t^2, \mathbf{x}_t^3; \lambda_1, \lambda_2) = \frac{\sin\theta\lambda_1 \cdot \mathbf{x}_t^1}{\sin\theta} + \frac{\sin\theta\lambda_2 \cdot \mathbf{x}_t^2}{\sin\theta} + \frac{\sin\theta(1 - \lambda_1 - \lambda_2) \cdot \mathbf{x}_t^3}{\sin\theta}$$

where,

$$\theta = \arccos\left(\frac{1}{3} \cdot \frac{\mathbf{x}_t^1 \cdot \mathbf{x}_t^2}{\|\mathbf{x}_t^1\|\|\mathbf{x}_t^2\|} + \frac{1}{3} \cdot \frac{\mathbf{x}_t^2 \cdot \mathbf{x}_t^3}{\|\mathbf{x}_t^2\|\|\mathbf{x}_t^3\|} + \frac{1}{3} \cdot \frac{\mathbf{x}_t^3 \cdot \mathbf{x}_t^1}{\|\mathbf{x}_t^3\|\|\mathbf{x}_t^1\|}\right).$$

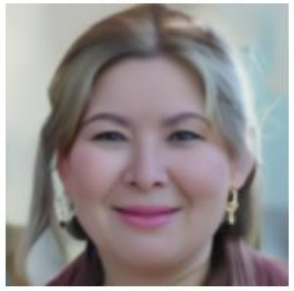
Results:

(a) $\lambda_1 = 1, \lambda_2 = 0$
Verified = True
Distance = 0.0

(b) $\lambda_1 = 0, \lambda_2 = 1$
Verified = False
Distance = 1.0102

(c) $\lambda_1 = 0, \lambda_2 = 0$
Verified = False
Distance = 0.9426

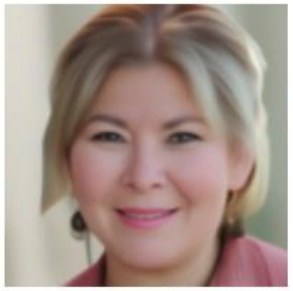
(d) $\lambda_1 = 0.6, \lambda_2 = 0.2$
Verified = True
Distance = 0.5782



(e) $\lambda_1 = 0.5, \lambda_2 = 0.5$
Verified = False
Distance = 0.9017

(f) $\lambda_1 = 0.4, \lambda_2 = 0.4$
Verified = True
Distance = 0.7755

(g) $\lambda_1 = 0.33, \lambda_2 = 0.33$
Verified = True
Distance = 0.7554



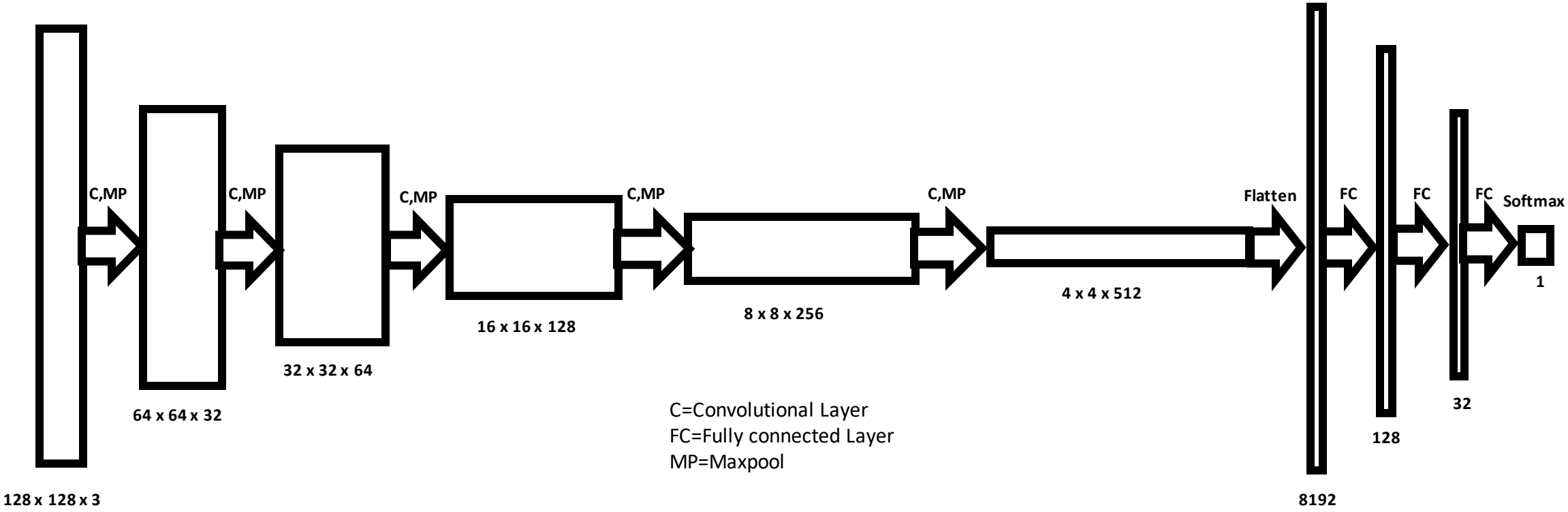
(h) $\lambda_1 = 0.2, \lambda_2 = 0.6$
Verified = False
Distance = 1.0102

(i) $\lambda_1 = 0.2, \lambda_2 = 0.2$
Verified = True
Distance = 0.8236

The base image chosen is (a), and the various interpolations are presented with their success rate on passing them through the VGG-Face facial recognition model with the distance metric set as the Euclidean L2 norm.

Experiment 2- Defense Strategy Against Face Morphing Attacks

Model:



The model was trained using the Adam optimizer with a learning rate of 0.001.
Binary cross entropy loss was the loss function used.
It was trained on the FFHQ dataset

Methods:

- We trained the model once using morphed images along with normal images and once with only normal images. Then we tested on a morphed image that had one sub-image inside the dataset and one outside.

Results:

- Training without morphed images was unsuccessful, giving a test loss of 0.69. However, training with morphed images gave good convergence and a test loss of 1.69e-20

Conclusions

Thus, spherical interpolation for semantics and linear interpolation for stochastic information produced the best morphing attack.

Changing the interpolation methods further did not produce much change in performance.

Morphing three faces together also produced good attack results.

The hypothesis that training the discriminator on both original and morphed images would give better results and this will be a good defence strategy against face morphing attacks was proven to be true.

Contributions

Annie & Anmol - Worked with different interpolation techniques to create different morphed images to attack SOTA face recognition models and implemented a three-face morphing strategy to attack the same model.

Amruta & Bhavya - Worked on the defence strategy against face morphing attacks by making the generator model, training it and conducting different attack and defence experiments.

Future Work-

1. Using Morphs Made Using Different Interpolation Techniques for training discriminator

2. Using 3-face Morphed Images for training discriminator