CS726: Advanced Machine Learning

Morpheous

A Diffusion-based Approach to Morphing Attacks and Defence Systems

May 3, 2023

Anmol Saraf 200070007
Bhavya Kohli 20d070021
Annie John D'souza 20d070028
Amruta Mahendra Parulekar 20d070009

Contents

1 Introduction

Face morphing attacks try to produce photos of faces that can be identified as the faces of two different people simultaneously, which might result in the construction of false identity links during operations like border inspections. Investigating new methods of creating face morphing attacks and identifying the facial recognition systems that are vulnerable to the attacks is necessary to anticipate novel attacks and help mitigate them.

Face morphing attacks can have multiple adversarial uses—if used in association with identity documents, it can allow multiple subjects to "verify" their identity to the information in the authorities' database. This possible false link might enable several illegal actions related to financial transactions, human trafficking, etc. If the method can be proved to be reliably dangerous, targeted methods to mitigate such attacks can be designed.

Usually, image-level interpolation of facial landmarks or representation-level GANs is used for this task.Image-level methods of facial morphing are prone to blending artifacts, which can easily be detected using modern networks, and sometimes even by direct inspection. GAN-based methods are far superior to image-level methods, which work by interpolating encoded images in the latent space to produce a morphed face upon decoding. However, these are constrained by the limited reconstruction fidelity of GAN architectures.

Recent advances in diffusion (autoencoder) models have overcome the limitations in the GAN-based method and have high reconstruction fidelity. The paper, 'MorDIFF: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Diffusion Autoencoders' by Naser Damer et al., discusses the creation of morphing attacks using diffusion autoencoders. The main idea behind the paper is to encode two faces using the diffusion autoencoder and morph them by interpolating between their semantic embeddings. The face detection models are attacked using the new image created from the combined semantic embedding.

2 Aim of the Project

The project has the following objectives:

1. In this project, we will make changes to the MorDIFF approach to generate different kinds of morphed images. First, we will try different interpolations using the diffusion autoencoder to generate the morphed images. After this, we will try morphing 3 faces by morphing two faces and then morphing the resultant images.

2. We will try a defence technique- using morphed images to train a discriminator model that distinguishes between the real face images and ones that have been morphed, to see if such a model is better equipped against morphing attacks.

3 Generation of Different Kinds of Morphed Images

3.1 Different Interpolation Techniques

The paper by Naser Damer et al., 2023, uses only Linear interpolation between the two face image semantic embeddings to generate new semantic embeddings. Meanwhile, it uses Spherical interpolation for the stochastic encodings of the respective images to generate the corresponding stochastic encoding for the morphed image. This new semantic embedding and stochastic encoded image is passed through the decoder to generate the new morphed image. The resulting interpolated images, except the initial images, are passed through the DeepFace, a face recognition model which contains many SOTA face recognition models. These interpolations

are passed through the model to gauge to what extent can the morphed images fool the models. Below are the results of the number of instances that are able to pass for each type of interpolation for the face recognition models based on the distance metric chosen.

3.1.1 Linear-Semantic and Spherical-Stochastic

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

3.1.2 Spherical-Semantic and Linear-Stochastic

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

3.1.3 Linear-Semantic and Linear-Stochastic

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

3.1.4 Spherical-Semantic and Spherical-Stochastic

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

3.1.5 Linear-Semantic and Quadratic-Stochastic

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

3.1.6 Quadratic-Semantic and Spherical-Stochastic

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.

3.2 Three Face Attack

Another implementation we made of the MorDIFF is of the three-face attack. Instead of interpolating two face images, three images are chosen where only one is known and correct. The three images' semantic embeddings are linearly interpolated and their respective stochastic encodings are spherically interpolated as follows,

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

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

where,

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

$\mathbf{z}_s^i$ is the semantic embedding of the ith image and similarly $\mathbf{x}_t^i$ is the stochastic encoding of the same image. λ_1 and λ_2 are varied to create different morphed face images passed through the VGG-Face model using the Euclidean L2 norm criterion for verification. Only linear and spherical interpolations for semantic and stochastic encodings are used, as these are the best interpolating method in the paper, Naser Damer et al., 2023. Below are the results of some morphed image attacks on the model.

(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

Figure 2: 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.

4 Defence Strategy Against Face Morphing Attacks

4.1 The Model

4.1.1 Architecture

The facial recognition model was a discriminator model that we created. It classified faces into two categories- ones that are in the database and ones that aren't. For this, we used five successive convolutional layers, followed by three fully connected layers. Finally, softmax activation was used to classify the image.

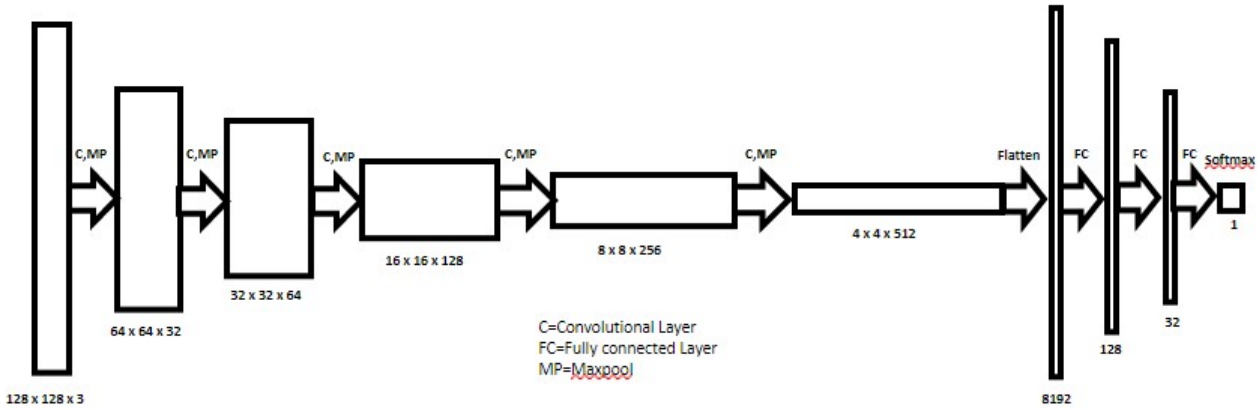

Figure 3: The Discriminator Architecture

4.1.2 Training Details

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.

4.2 The Attack

The proposed attack is to use the morphed image of a person in the database and a person not in the database. If this is marked as positive by the model, then the attack is said to be successful. In this project, three attacks were conducted, one using two face morphed images, one using three face morphed images and one using morphs made with different interpolation techniques.

4.3 The Defence Strategy

We propose training a discriminator model with *both* real images and morphed images using pairwise combinations of a percentage of people in the complete dataset. The idea is that the model trained in such a manner will be more robust than if it was trained only on the original data. If data is sparse, i.e., we only have a few pictures of the members, this method also serves as a way to effectively increase the size of the dataset, improving its learning. Upon giving a morphed image (one face in the database and one outside) the model should declare it as a "0" for the defence to be successful.

4.4 Experiment

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.

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

5 Conclusion

5.1 Generation of Morphed Images

5.1.1 Different Interpolation Techniques

The generation of morphed images using different interpolation methods showed that there was a certain threshold to which the morphed image could pass through the model for different facial recognition models. This can be seen as the number of passed morphed interpolations become constant for all interpolations apart from the linear-spherical interpolation. Furthermore, the least secure model was found to be the VGG-Face amongst the four models as it passed through mostly all the interpolated morphed images, even when the actual image representation was quite low. The more secure metric for the facial recognition model seemed to be the cosine metric, as it outperformed the other two in thresholding the morphed images.

5.1.2 Three-face Attack

For the three-face attack, it was found that if the initial image representation is relatively higher than the other two, it passes the model more frequently. This can be seen in Figure 2. (d), (f), (g) cases where even though the representation is not that high per se but since it's relatively the highest, it dominates and passes as a positive through the model. Another inference that can be made from this is that we can add more amount of similar-looking faces in the morphed image as they pass through the model. This could be due to the semantics of the two images being similar to each other and thus creating a good morph image when their proportions are high. This can be seen in Figure 2. (a) and (c) is somewhat similar to (a) and (b), and thus even with high values of (c) as in (g), (i) the morphed images pass easily but a high value of (b) does not reflect the same as in (e) and (h).

5.2 Defence Strategy Against Face Morphing Attacks

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.

6 Work Distribution

1. Annie and 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.

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

7 Future work

We can also train our model on morphed images that we have generated using three faces and those generated using different interpolation techniques.