Adversarial attacks on Facial Recognition Models

Kartik Thakral
PhD19004

Dimpy Varshni
MT19022

1. Introduction

Deep Neural Networks (DNNs) have always been known to perform well for any image classification task, so is the case of facial recognition models. However, deep learning based models are highly vulnerable to adversarial attacks. Recent researches have been focused on exploiting the robustness of deep learning based models through these attacks which acts as the primary motivation for our project. But in comparison to the current researches going on this topic specifically for facial recognition models which utilizes prior information of internal DNN architecture, our project just aims to use addition of noises as adversarial samples for analyzing the sensitivity of deep learning based facial recognition models. For this, we use a pipeline for facial recognition model which uses open-source facenet [4] features fed to SVM network. The primary goal of the project is to demonstrate the effects of basic image processing based adversarial attacks on the mentioned pipeline.

2. Related Work

As mentioned, a lot of researches have been done and are still going on the chosen topic. Goswami et. al. [2] showed the effects of image level and face level distortions on open-source DNN face recognition models. Face spoofing has also been one of the explored attacks. Also, the researches have been done on mitigation of adversarial samples.

3. Dataset Description

The dataset used is publicly available Labelled faces in the Wild (LFW) [3] dataset that can be used for various tasks in face biometrics such as face recognition, face verification etc. The dataset originally consisted of 13,233 images of 5,749 individuals. But as our task is aligned to face recognition only, while exploring dataset we noticed that original dataset is highly unbalanced for the task which can be seen in Figure 1. From Fig 1, it is clearly visible that high number of classes contains only one image which shows the instability in the dataset. To overcome this, the dataset is modified with the classes consisting minimum of 10 images per class. The resulting dataset is comparatively smaller consisting a total of 4312 images, out of which 70% is used for training and 30% testing. Because of this instability, we wish to explore new face dataset also after finishing with this one (if time permits).

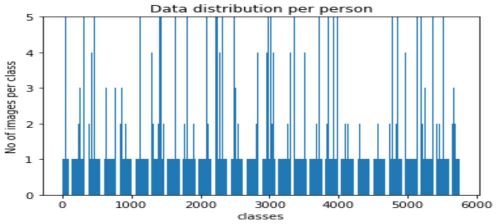

Figure 1. To show the amount of instability of dataset in terms of images per individuals for face recognition, the graph is limited up-to 5 images per class.

As we are using pre-trained facenet model architecture for features feeding to SVM network, validation dataset is not kept for the purpose.

4. Methodology

As our main task is to fool deep learning based face recognition model, the inclination of our work is towards generating adversarial sample. Therefore, for the task the pipeline of face recognition model is chosen which uses pre-trained facenet model for feature extraction part and SVM for classification/recognition part (Cole Murray [5] pipeline is referred). Also, the necessary pre-processing steps were taken before feeding the images to the network including segmentation, cropping and alignment using dlib library. Adversarial attacks were performed by generating Gaussian noise, Salt and pepper noise and Random noise. As definition of Gaussian noise and salt and pepper noise is usually known, we only discuss random noise generation method in this section.

4.1. Random Noise

Random noise (inspired from [1])was created by generating random co-ordinates in a dummy image of size same as the target image. On those random co-ordinates, pixel value was set to 255 for all the three channels resulting in white dots in the output noise image.

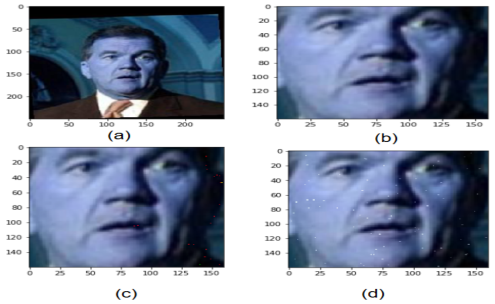

Figure 2. a represents the sample image from the dataset, b represents the pre-processed aligned image, c and d represents a sample of Gaussian noise and random noise adversarial sample.

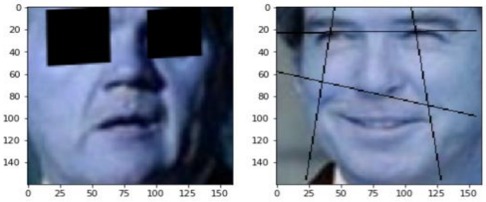

Figure 3. Figure representing eye-patch based and grid lines based generated adversarial samples

4.2. Grid based occlusion

was performed as described below:

1. First, randomly generate four random points at the boundary of the image, then
2. construct two lines passing through them such that the randomly generated co-ordinates lie on the boundary of the image.

4.3. Eye-patch based occlusion

For this we made use of 'Haar-Cascade' which is used to locate face and other facial attributes, so we first detect and crop the face out of the image, then we locate eyes. After getting the coordinates of eyes, we simply built a black solid rectangle from the coordinates over each eye.

5. Results, Analysis and Progress

In this section, we briefly discuss the effect of adversarial samples on the performance of the mentioned pipeline. Starting with the noise generation, Figure 2 shows both the original and perturbed image for Gaussian noise attack as well as perturbed image for random noise attack. The success of both the attacks are shown in amount of accuracy both of the attacks were able to drop. From Table 1, it is clearly visible that Gaussian noise attack was able to drop accuracy only up to 7% while random noise attack was able to drop accuracy up to 19 %.

The analysis of the difference between the success of these two attacks can be done with the help of the euclidean distance of the features obtained from facenet that are fed to SVM of the perturbed and the original images. The mean of the euclidean distance of the same is mentioned in Table 1, from which it is clear that euclidean distance of adversarial samples generated with Gaussian noise is comparatively less than the one generated with random noise, which means that the feature maps generated for both the original images at the intermediate layer are quite similar in case of Gaussian attack but this is not the case in random noise attack.

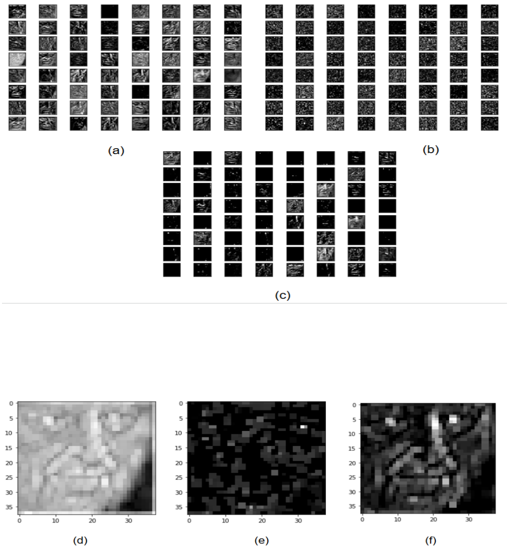

Figure 4. a, b, c are the feature maps obtained at maxpool3 layer of the Facenet model of the original images, Gaussian attack and random noise attack perturbed images. to bring out the difference between the same. d, e and f represents one feature map only taken respectively from a, b and c. It is visible that d and f are more similar than d and e.

From the results obtained, it can be observed that face recognition model is vulnerable to the adversarial attacks. One of the main reason behind the possibility of these attacks is the mystery of the features learnt by each layer within a deep neural network. We have to just feed the network with images and all the learning is done implicitly by a network, we can't really control the learning of features within the network. The final features obtained at the end of the model in a way is the combination of the features obtained at each layer. Although, we use pre-trained facenet model for our task, but the difference in the feature maps should still be visible as the weights of facenet is learned on face dataset only. For instance, in case of random noise addition, even the euclidean distance between the perturbed and original images is within the range of 0 to 10, (shown in table) the feature maps obtained at each layer of facenet is highly affected evident by Figure3 which shows one such

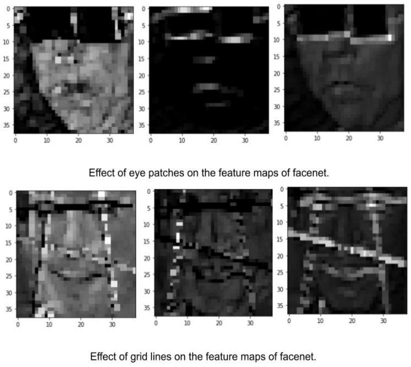

Figure 5. Effect of attacks on feature maps of facenet (taken from MAXPOOL 3 layer)

difference in feature maps of perturbed and original images at MAXPOOL layer 3, which in result obviously affects the overall features obtained at the end. This clearly shows why there is a drop of accuracy on perturbed images.

5.1. Performance of grid-lines based and eye-patches based attacks.

The model was again attacked with a grid based noise and eye-patch based noise which was added to the input images. The drop in accuracy witnessed with grid based and eye patch was significant i.e. accuracy achieved was around 55% and 45%. The probable explanation for this trend is justified by visualizing the feature maps. If we analyze them, we can see the amount of noise in them has caused the distortion of the features to learnt. This drop in accuracy can also be accounted by the fact that deeper layers of the network tend to learn about minute features like eyes, lips etc. So, adding noises like eye patches, beard patches or grid lines in to the input image will definitely decrease the classification accuracy as noises like eye patch or grid lines directly distort these features which can easily be seen from the feature maps. These noises directly attack the features that are acceptable to be learnt by a face recognition deep learning model.Same is shown in Figure 5. The limitation of these noises is that these are not disguisable noises, i.e. on adding these noises to the input image, the visual appearance of the image will change and be easily noticed. Considering Gaussian and Salt and Pepper noise, these were a great failures as they didn't make any changes in the visual appearance of the image, but at the same time grid noise is seen to perform best among all added noises.

Attacks	Mean ED	Perturbed Results	Original Results
Gaussian Noise	6.72	88.47%	95.260%
Random Noise	9.51	76.25%	95.260%
Grid based	10.23	**34.49**%	95.260%
Eye-patch based	15.10	**55.26**%	95.260%

Table 1. Results of the attacks performed on mentioned face recognition pipeline, here ED stands for Euclidean Distance

6. Individual Contributions

6.1. Deliverables

- All the promised deliverables were met at the time of interim review.
- Individual Contribution: MT19022 did two attacks - gaussian and grid lines based and Phd19004 performed other two attacks. Preprocessing was done by Phd19004 and feature-extraction was done by MT19022. Analysis of all the written code and results given are done by both.
- As asked all the proper references of the code is attached in a file named referemces.txt along with the code submitted.

References

[1] Michael Karr Andrew Milich. Eluding mass surveillance: Adversarial attacks on facial recognition models.

[2] Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Thirty-Second AAAI on Artificial Intelligence*, 2018.

[3] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[4] Google Inc. Google's face recognition model.

[5] Cole Murray. " building a facial recognition pipeline with deep learning in tensorflow. *Accessed on October*, 5, 2017.