Adversarial attacks on Facial Recognition Models

By,
Group 41
Kartik Thakral (PhD19004)
Dimpy Varshni (MT19022)

Introduction

- **Threat to Deep Learning Models**
  As it is known that for any image classification task, deep learning models have always outperformed all the other existing techniques in arena of machine learning. But as like any other thing in this world, these models are also not fully secure and are vulnerable to attacks referred as *Adversarial Attacks.*

- So, **What are Adversarial attacks ?**

  Any changes in the input of the targeted network that are able to fool the network which leads the model to misclassify the inputs and reduce its performance.

- **What's our aim ?**
  In this project, we ought to explore the robustness of face recognition model through black box attacks.

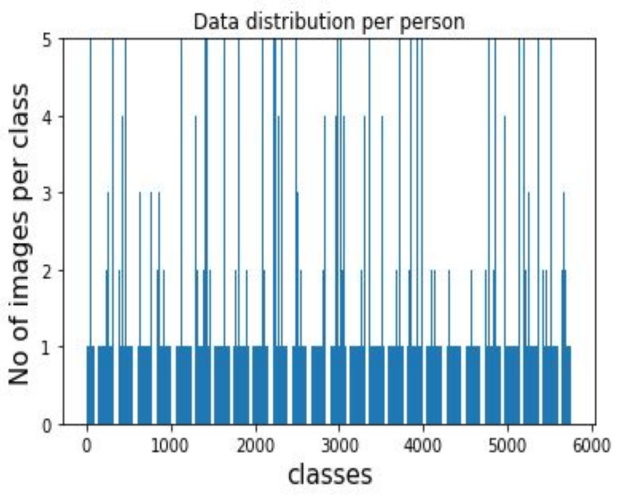
Dataset Description

- The LFW dataset is used for the task which is publicly available on kaggle platform consisting of 13,233 images in total of 5,749 individuals.

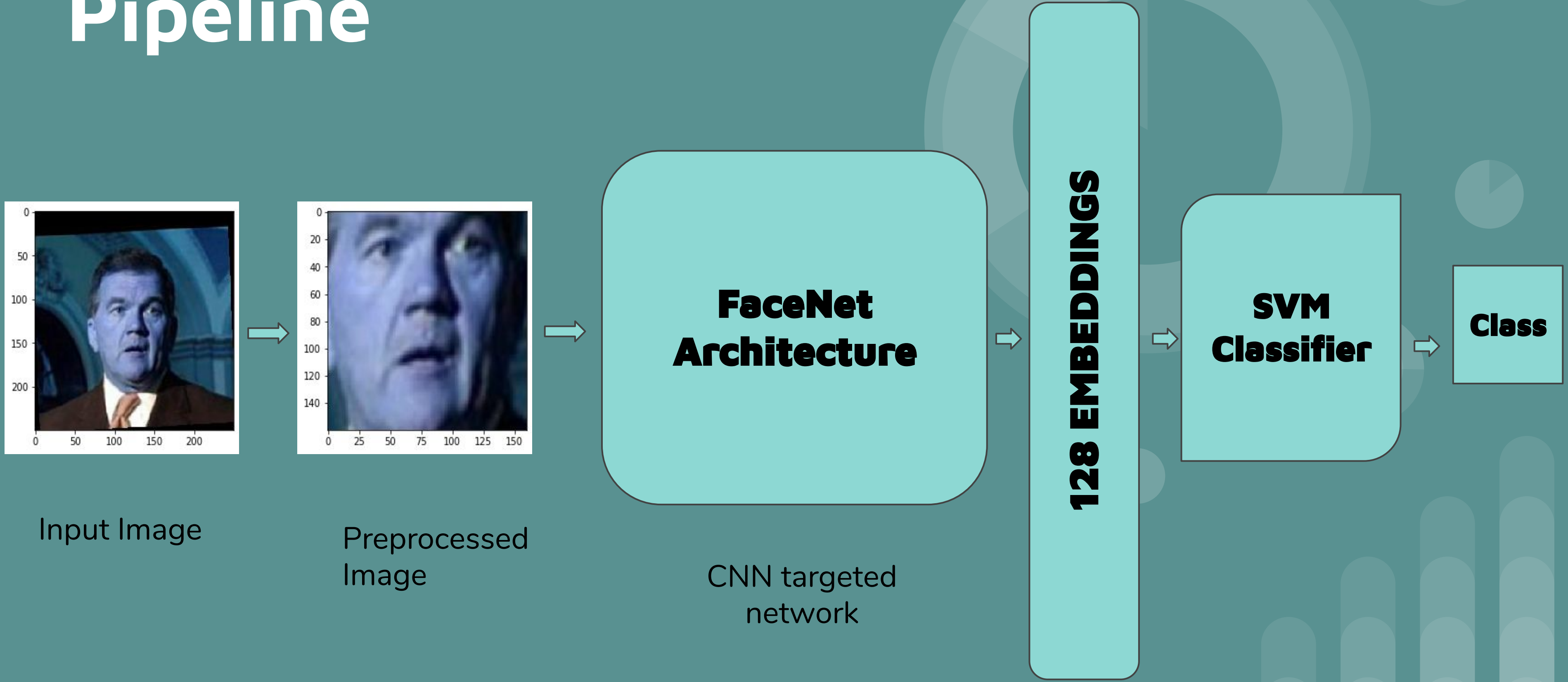
- But the primary task is of classifying face images belonging to each individual using a DNN; while exploring the dataset we found that it is highly imbalanced (as you can see in fig), due to which the dataset was modified and classes having at least 10 images were considered.

- The resulting dataset consisted of total 4312 images, which is quite smaller than the original version.

- 70%-30% ratio is considered for splitting the dataset into training and testing.



Targeted Face Recognition Pipeline



Input Image

Preprocessed Image

FaceNet Architecture

CNN targeted network

128 EMBEDDINGS

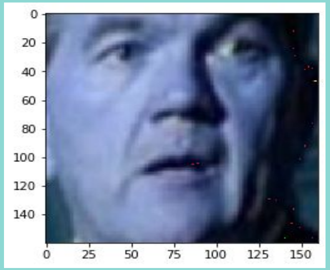
SVM Classifier

Class

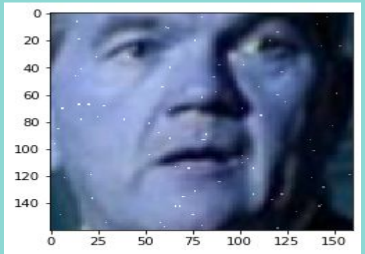
Image Perturbations Techniques

- Adversarial attacks were first performed by generating Gaussian noise, Salt and pepper noise and Random noise.
- When these attacks were not able to show satisfactory drops in accuracy, proper literature review was done and two more attack were performed namely: Grid lines based and eye patch based.

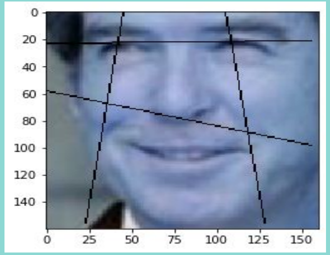
Gaussian Noise with mean as the center of the image and variance 0.5 was constructed first and then added to each test image.



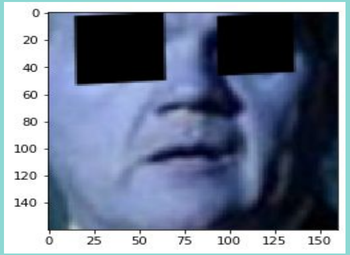
Random noise was created by generating random coordinates in a dummy image of size same as the target image (random coord. = 100).



Grid lines noise was generated by joining 4 points with 2 lines with each point at the boundary of the input image.



Eye patch noise was generated by first extract the face and then extracting the coordinates of the eyes and plotting a black mask over it..



Results & Analysis

Attacks	Mean Euclidean Distance	Perturbed Results	Original Results
Gaussian Noise	6.72	88.47 %	95.260 %
Random Noise	9.51	76.25 %	95.260 %
Grid based noise	10.23	34.49 %	95.260 %
Eye-patch base noise	15.10	55.26 %	95.260 %

Inferences (till interim):

- **Random Noise comes out to be better attack than the Gaussian noise.**
- **Random noise achieved to drop accuracy of the model by 19% while Gaussian noise was able to drop accuracy by 7% only.**

Inferences (after interim):

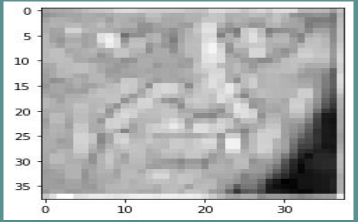
- **None of the attacks performed yet is a proper as the accuracy drop is not significant, as compared to eye patch and grid lines attack where reduction in model performance is clearly visible, 61% drop of accuracy by grid noise and 40% drop of accuracy by eye patch based noise attack**

Notes :

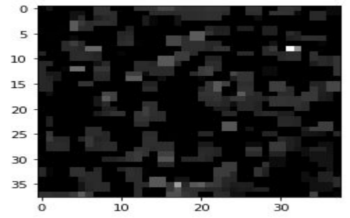
- **The mean Euclidean distance here is the mean of all the euclidean distances between the original & perturbed image 128 feature embeddings obtained from facenet architecture.**

Results & Analysis (cont.)

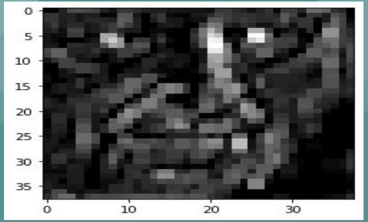
- The analysis of the difference between the success of these two attacks can be done with the help of the euclidean distance of the features obtained from facenet that are fed to SVM of the perturbed and the original images, results of which were shown in previous slide.
- To get the intuition of what's actually going inside the facenet n/w, let's look at one of the feature map generated by the n/w at MAXPOOL_3 layer.
- As the final features obtained at the end of the model in a way is the combination of the features obtained at each layer, this clearly explains the results.



Feature map of original image

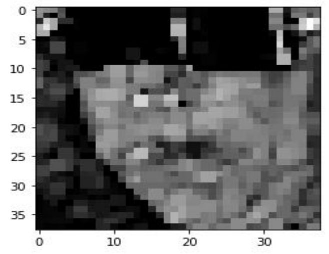


Feature map of random noise

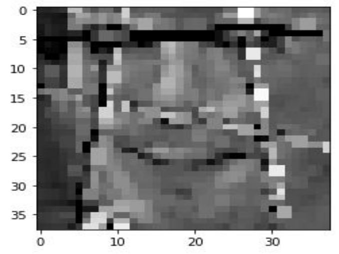


Feature map of Gaussian noise

- Again, these similar effects justifying the results of other two attacks can be seen in similar way.



Feature map of eye-patch based



Feature map of grid-lines based