Adversarial Attack on Soft Biometric Detection

Marrone, Stefano, and Carlo Sansone
2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019
Budapest, Hungary

[Group 12]

Nisarg Doshi 202111033

Utkarsh Pandya 202111026

May 12, 2022

Introduction and Motivation

Introduction:

- The use of biometric-based authentication systems
- Emergence of Soft Biometrics for Detection of Person of Interest.
- Use of Deep Learning in Soft Biometric Detection.

Motivation:

- Defensive application to protect Person Of Interest
- Defensive strategies in the view of a more secure and private AI.

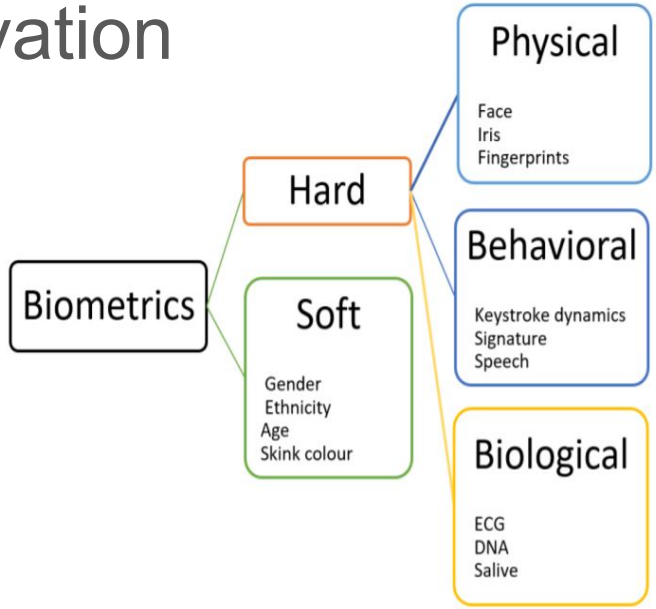

Figure 1. Biometrics Classification[1]

[1] Marrone, S. and Sansone, C., 2019, July. An adversarial perturbation approach against CNN-based soft biometrics detection. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Problem Statement

- To exploit Adversarial Perturbation and mislead state-of-the-art CNNs by injecting a suitable small perturbation over the input image.
- To protect subjects against unwanted soft biometrics-based identification by automatic means.
- We will attempt to generate visible adversarial patches.

Dataset and Architecture

Dataset:

- UTKFace Dataset[2] by University of Tennessee, Knoxville
- Contains 20,000 images with annotations
- Annotations : Age, Gender and Ethnicity & Variety : Pose, Facial Expression, Illumination, Occlusion

Preprocessing:

- Image resizing
- Normalization
- Label Encoding

Model:

- Model to be used : VGG16[1]
- Feature Chosen : Ethnicity
- White, Black, Asian, Indian, Other

Figure 2. Sample Images from UTK Dataset[2]

[2] “UtkFace Dataset,” *Utkface*. [Online]. Available: https://susanqq.github.io/UTKFace/. [Accessed: 26-Mar-2022].

Block Diagram

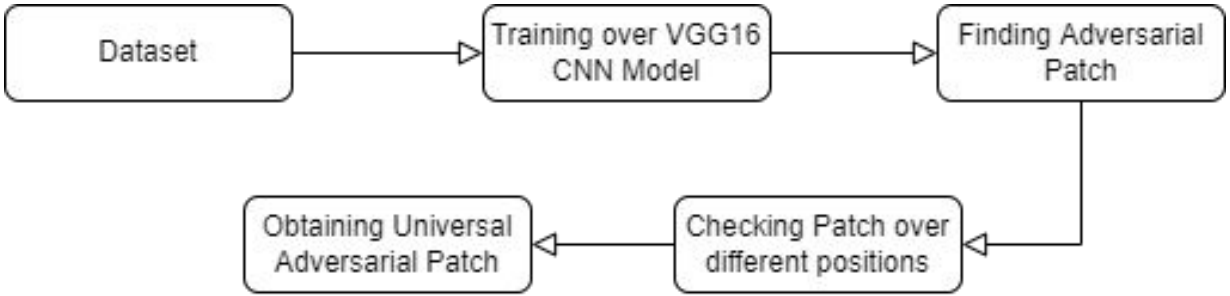

Figure 3. Flow Chart for the Project

Approach used

Training the Model:

- Extracting pretrained VGG16[1] CNN model pretrained on imagenet weights without the top layers(Transfer Learning).
- Adding the final layers for retrofitting the model for ethnicity classification.
- Training on UTKFace Dataset for ethnicity.
- Input parameters for model are set for 100 x 100 x 3 size images.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 keras_layer (KerasLayer)    (None, 7, 7, 512)         14714688

 flatten (Flatten)           (None, 25088)             0

 dense (Dense)               (None, 5)                 125445

=================================================================
Total params: 14,840,133
Trainable params: 125,445
Non-trainable params: 14,714,688
_________________________________________________________________
```

[1] Marrone, S. and Sansone, C., 2019, July. An adversarial perturbation approach against CNN-based soft biometrics detection. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Approach used

Performing the Attack:

- Wrapping the model into a classifier.
- Using ART[3] Toolbox to perform Adversarial Patch instance.
- Low Epochs used and different parameters changed(e.g. Rotation,scaling,mask limits,etc.)
- Trying the adversarial patch on all positions in every image.
- Test with different size of adversarial patch (proportion of patch to image used are 0.1 and 0.4).

[3] GitHub. 2022. *GitHub - Trusted-AI/adversarial-robustness-toolbox: Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference - Red and Blue Teams*. [online] Available at: <https://github.com/Trusted-AI/adversarial-robustness-toolbox> [Accessed 12 May 2022].

Results

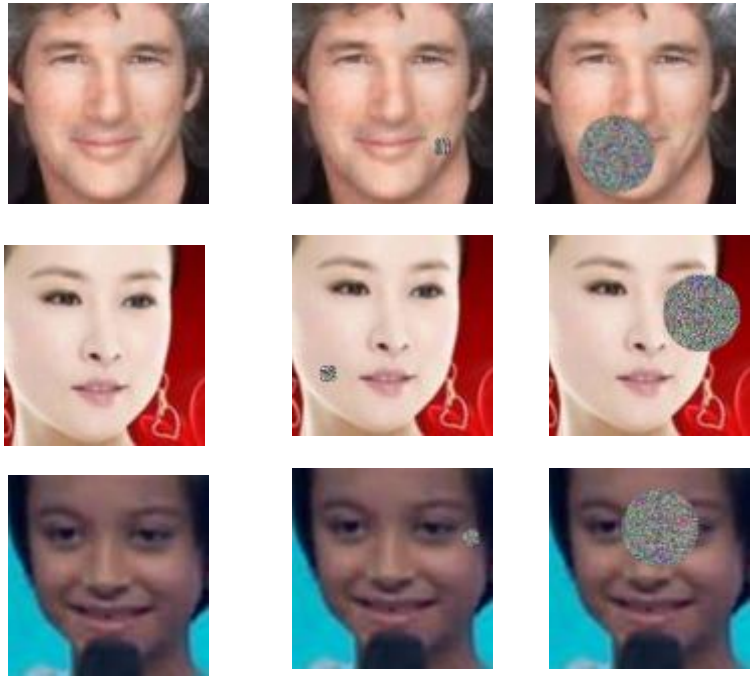

Figure 4. Resulting Images after patch generation

Conclusion

- Adversarial Patches are helpful in creating a real time attack on the model.
- A portable and flexible attack form.
- Easy to carry a printable patch.

Suggestions for improvement

- For issue revolving around crashing the colab. Further techniques to reduce the trainable parameters can be used to reduce computations.
- A better standalone GPU can be used for stable code implementation.
- Patch can be trained across the models.

References

[1] Marrone, S. and Sansone, C., 2019, July. An adversarial perturbation approach against CNN-based soft biometrics detection. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[2] “UtkFace Dataset,” *Utkface*. [Online]. Available: https://susanqq.github.io/UTKFace/. [Accessed: 26-Mar-2022].

[3] GitHub. 2022. *GitHub - Trusted-AI/adversarial-robustness-toolbox: Adversarial Robustness Toolbox (ART) - Python Library for Machine Learning Security - Evasion, Poisoning, Extraction, Inference - Red and Blue Teams*. [online] Available at: <https://github.com/Trusted-AI/adversarial-robustness-toolbox> [Accessed 12 May 2022].

Thank You