Face Recognition Evasion

Hekmat Saker
Nour Hmeedan
Obada Alnaddaf
Nargiz Aghayeva

Motivation and Goal:

- Face recognition is widely used in security and authentication.
- These systems can be fooled by adversarial attacks with tiny, invisible changes.
- Investigate the effectiveness of current anti-face recognition methods
- Applying PGD and DeepFool algorithms to subtly alter images and evaluate if the model still verifies them correctly.
- This helps understand risks in current face recognition systems and highlights the need for better defenses.

Choosing the Model: FaceNet

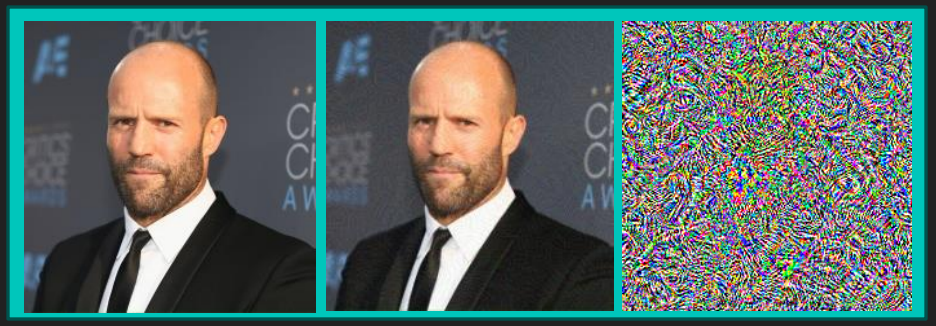
- **What is FaceNet?**
  - A deep learning model designed for face recognition and verification.
  - Converts face images into 128-dimensional numerical vectors called *embeddings*.
  - Instead of classifying faces directly, it compares embeddings using simple distance metrics (like Euclidean distance) to measure similarity.
- **Why Did We Use FaceNet?**
  - **Flexible and well-documented** with open-source implementations in TensorFlow and Keras.
  - **Pre-trained models available**, reducing development time and computational costs.
  - **High accuracy and reliability** in face recognition benchmarks.

Projected Gradient Descent

PGD , it's an adversarial attack method to modify image so that the face recognition model misclassifies it.

Algorithm steps:

- Load Target Image.
- Preprocess for ResNet18.
- Set PGD Parameters/Ep, Alpha, Iteration/.
- Perform PGD Attack.
- Use FaceNet to verify.
- Measure verification and distance.



Challenges:

- Balancing perturbation without strong visual distortion
- Fine-tuning PGD Pparameters.

DeepFool

DeepFool algorithm used to generate minimal perturbations on input images, aiming to fool a face recognition system without visually obvious changes

Algorithm steps:

- Load Target Image.
- Preprocess for ResNet18.
- Apply DeepFool Attack.
- Save and resize adversarial image.
- Use FaceNet to verify.
- Measure verification and distance.

Challenges:

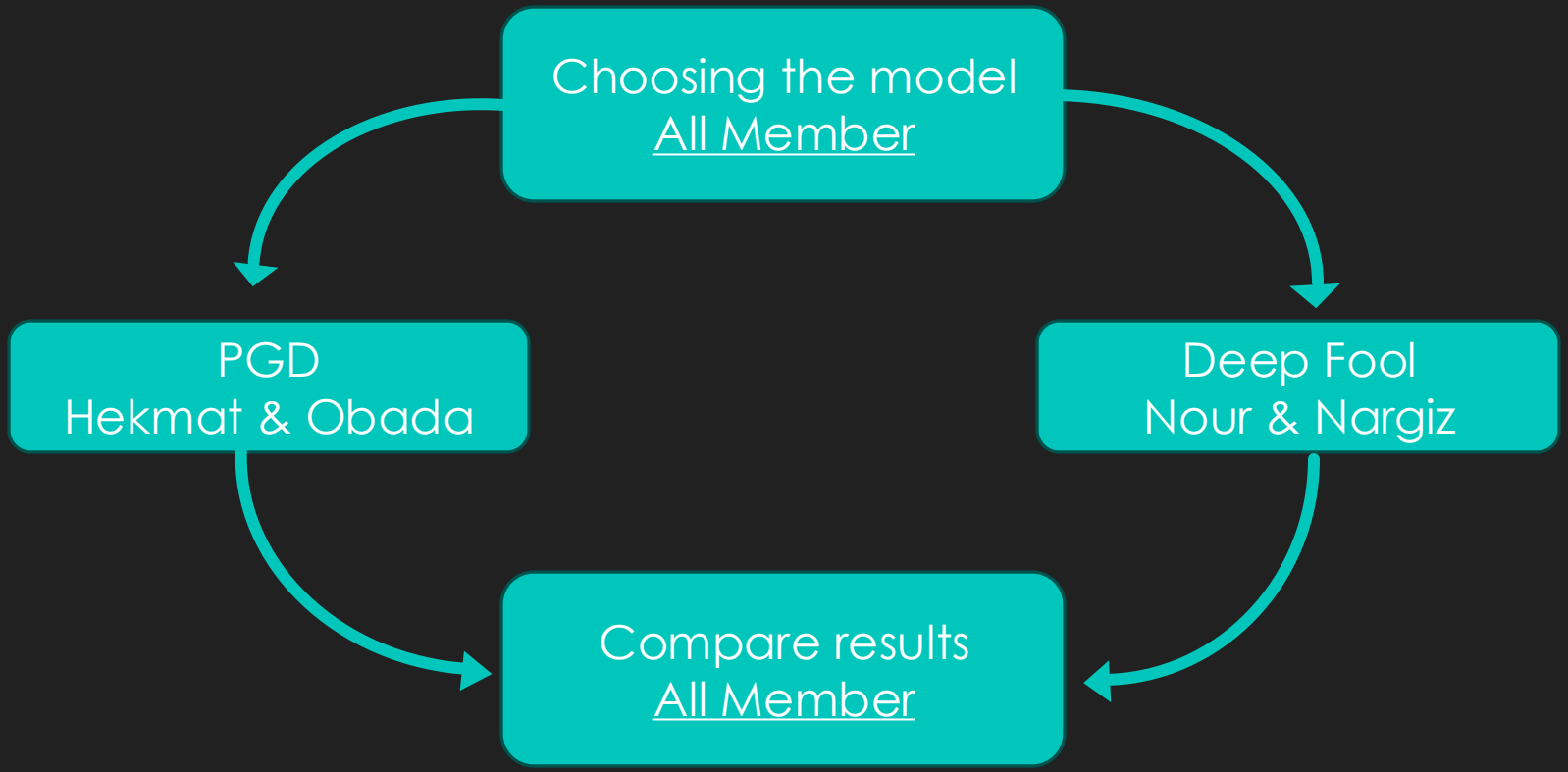
- Balancing perturbation without strong visual distortion
- Fine-tuning DeepFool parameters.

Step	PGD	DeepFool
Initialize	Start with the original image + small random noise (optional)	Start with the original image
Gradient Calculation	Compute the gradient of the loss with respect to the input image	Approximate the classifier as linear around the current image
Perturbation Update	Take a small step (a) in the direction of the gradient	Calculate minimal perturbation needed to cross the nearest decision boundary
Projection/Adjustment	Project the perturbed image back into the allowed perturbation region (bounded by ε)	Apply the minimal perturbation to the image
Iteration	Repeat steps 2-4 for a set number of iterations or until misclassification	Repeat steps 2-4 until the prediction changes
Output	Return the final adversarial image	Return the minimally perturbed adversarial image

Comparision

	PGD	DeepFool
Speed	Slower	Faster
CPU Usage	Higher, esp. with large ε	Lower to moderate
Visual Detection	Sometimes visible/noisy	Barely perceptible

Development Contribution



Conclusion

- Adversarial attacks like PGD and DeepFool can successfully fool state-of-the-art face recognition models such as FaceNet.
- PGD, being an iterative and stronger attack, offers a higher success rate in misclassifying or confusing the model compared to simpler, one-step methods.
- Both methods demonstrated that modern face recognition systems are vulnerable to carefully crafted adversarial examples.

References

- FaceNet: A Unified Embedding for Face Recognition and Clustering (1503.03832)
- Bell B. et al (2024), "Persistent Classification: Understanding Adversarial Attacks by Studying Decision Boundary Dynamics", https://arxiv.org/html/2404.08069v1
- DeepFool: a simple and accurate method to fool deep neural networks 1511.04599
- DeepFace: Closing the Gap to Human-Level Performance in Face Verification: DeepFace: Closing the Gap to Human-Level Performance in Face Verification
- OpenCV Documentation: OpenCV: OpenCV modules
- Github Repo Link: Avoid-Face-Recognition