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

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

Challenges:

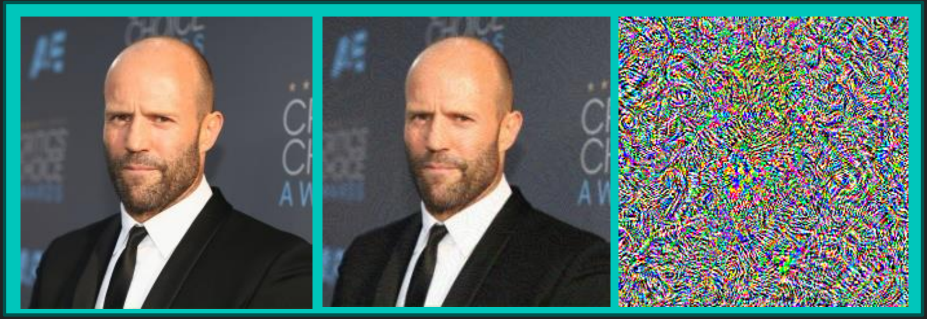
- Balancing perturbation without strong visual distortion
- Fine-tuning DeepFool parameters.

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 DeepFace to verify.
- Measure verification and distance.



Challenges:

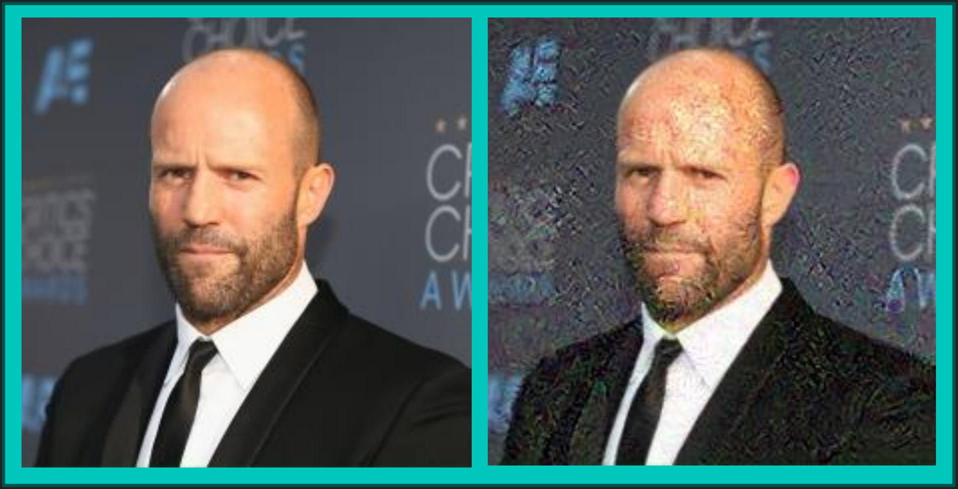
- Balancing perturbation without strong visual distortion
- Fine-tuning DeepFool parameters.

DeepFool

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

Algorithm steps:

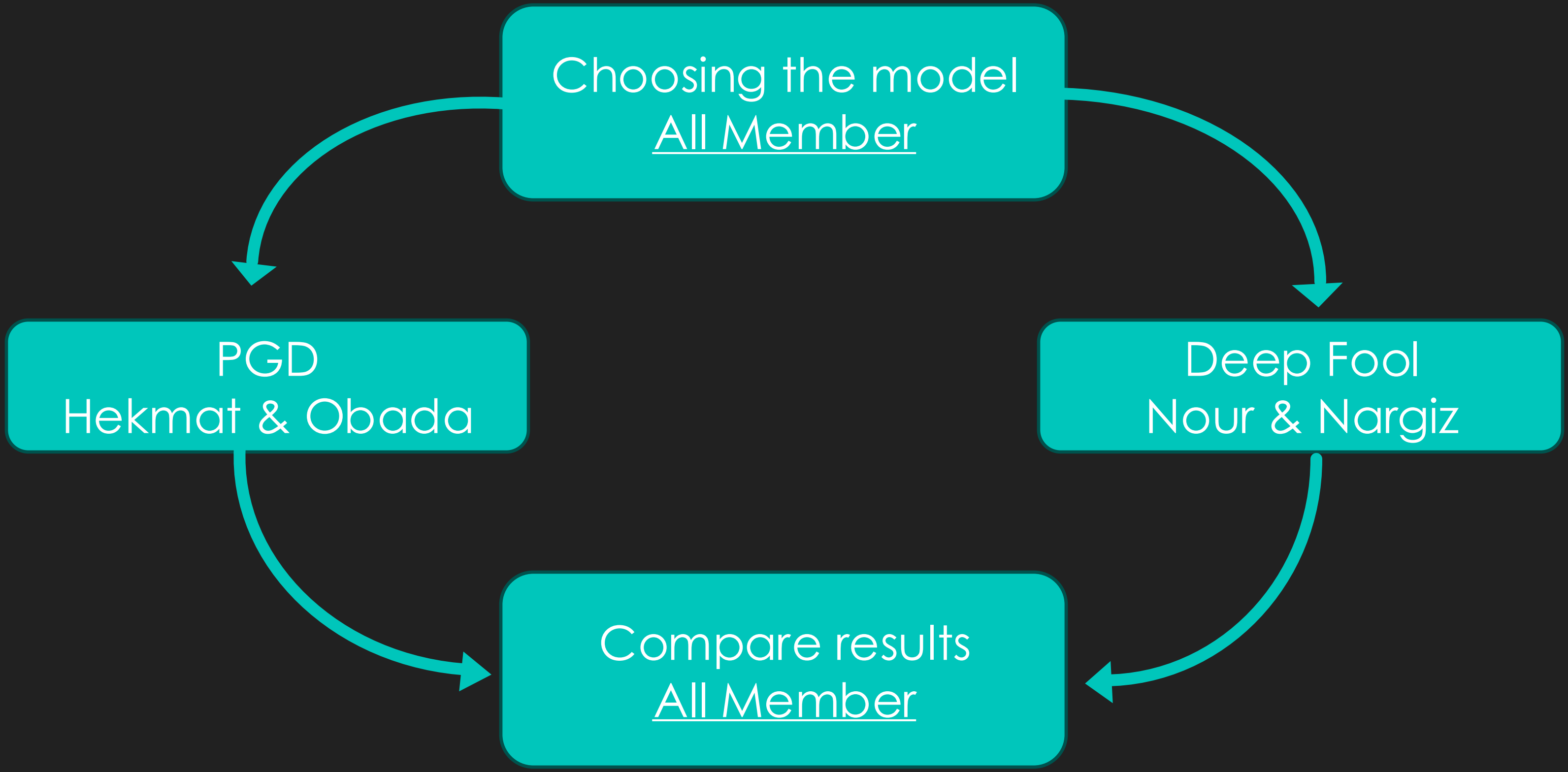
- Load Target Image.
- Preprocess for ResNet18.
- Apply DeepFool Attack.
- Save and resize adversarial image.
- Use FaceNet to verify.
- Measure verification and distance.



Challenges:

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

Development Contribution



Choosing the model
All Member

PGD
Hekmat & Obada

Deep Fool
Nour & Nargiz

Compare results
All Member

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