Mitigation Method Against Privacy Violations Attacks on FR Systems

Hackathon Demo

Members: Guy Elovici and Alon Schneider
Mentors: Prof. Asaf Shabtai and Dr. Edita Grolman

CBG Cyber Security Research Center at Ben Gurion University of the Negev

CBG Cyber@ Ben-Gurion University of the Negev

Outline

- Reminder – Project Proposal
- Project Structure
- Initial System Results

Outline

- Reminder – Project Proposal
- Project Structure
- Initial System Results

Outline

We are here

- Reminder – Project Proposal
  - Introduction and Motivation
  - Problem Definition
- Project Structure
- Initial System Results

Reminder – Project Proposal

Outline

- Reminder – Project Proposal
- We are here → Introduction and Motivation
- Problem Definition

Introduction and Motivation

Introduction and Motivation

- We are Guy Elovici and Alon Schneider, 4th year students in data engineering.
- Our mentors are Prof. Asaf Shabtai and Dr. Edita Grolman.
- The field of our project is mitigation method against privacy violations attacks on face recognition (FR) systems.

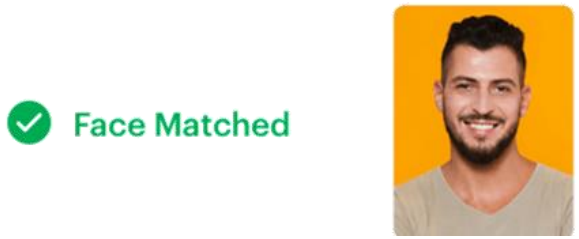
Problem Definition

Outline

- ❑ Reminder – Project Proposal
  - ✓ Introduction and Motivation
- We are here → ❑ Problem Definition

Problem Definition

- Face verification systems, one of the most common computer vision applications, are typically used to compare a photo ID (e.g., passport, driver's license) with an existing photo of an individual.



Problem Definition

- Recent studies have demonstrated the ability to infer various sensitive information related to the dataset set that was used to train the machine learning (ML) model.
- Since FR models are trained on individuals' data (e.g., individuals' images), they are vulnerable to various privacy violation attacks.

Problem Definition

- Most of the FR systems are based on deep neural networks, using publicly available pre-trained components, referred to as backbones, for inducing the FR model (for example, AlexNet, ResNet, and VGG)
- These backbones are commonly used and showed to be vulnerable for adversaries' goals – i.e., privacy attack.
- The goal of this project is to propose a new defense mechanism against privacy violation attacks by masking the backbones influence on the trained model.

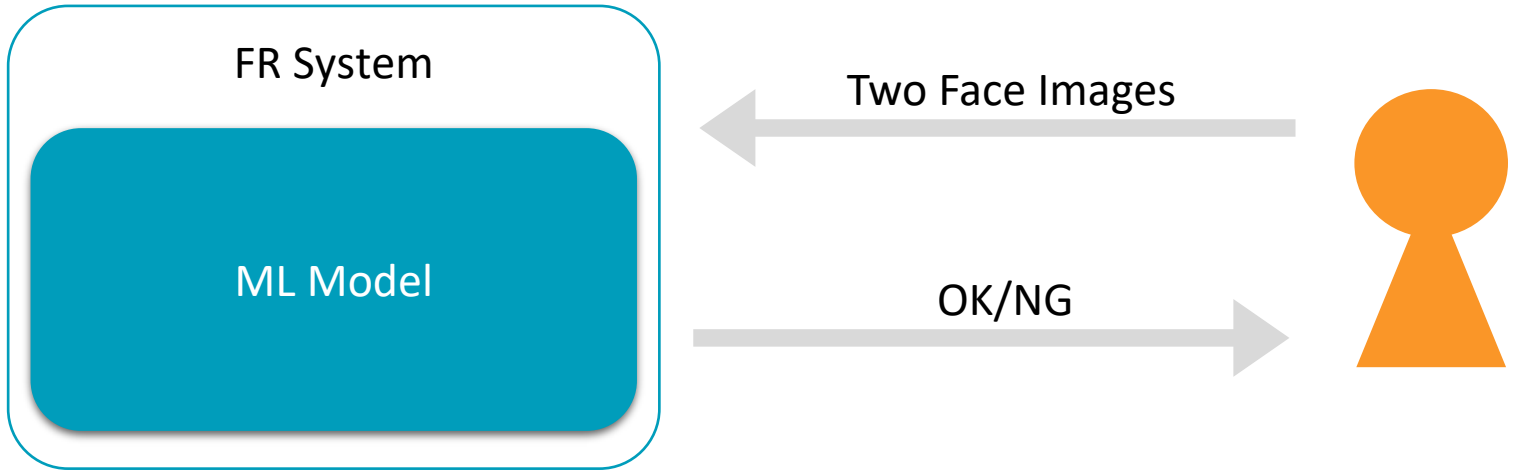
Project Structure

Outline

- ✓ Reminder – Project Proposal
  - ✓ Introduction and Motivation
  - ✓ Problem Definition
  - ✓ Project Content
- We are here → ❑ Project Structure
- ❑ Initial System Results

Selected Use-Case

- Face verification with a photo ID. (e.g., passport, driver's license).

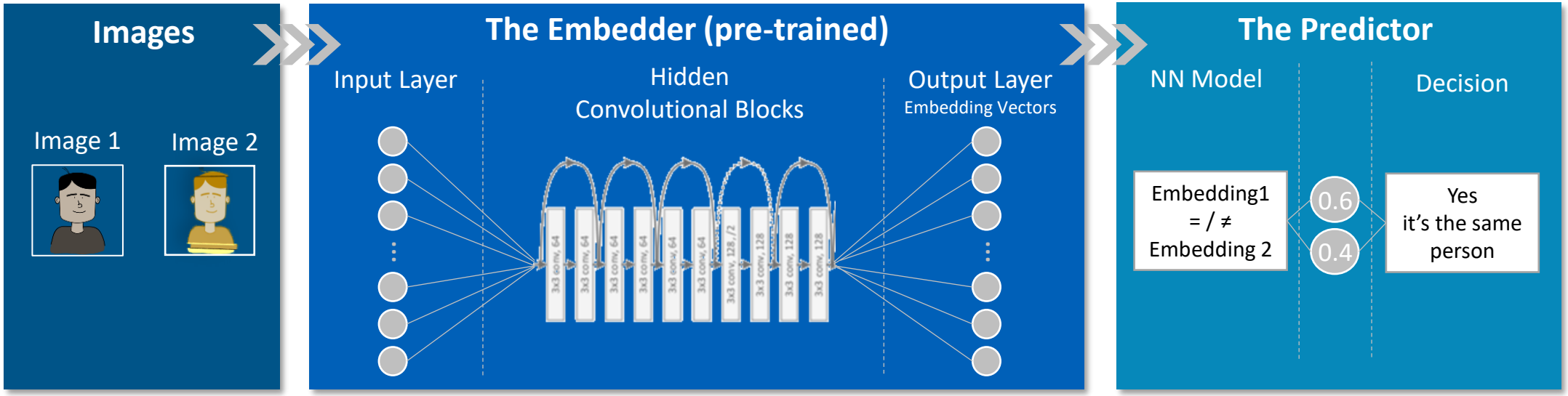


Data Set

- In this demo we use the Labeled Faces in the Wild (LFW) dataset.
  - This dataset contains more the 13,000 face images collected from the web
  - Each image is labeled with the person's name, which makes this dataset appropriate for face verification tasks.

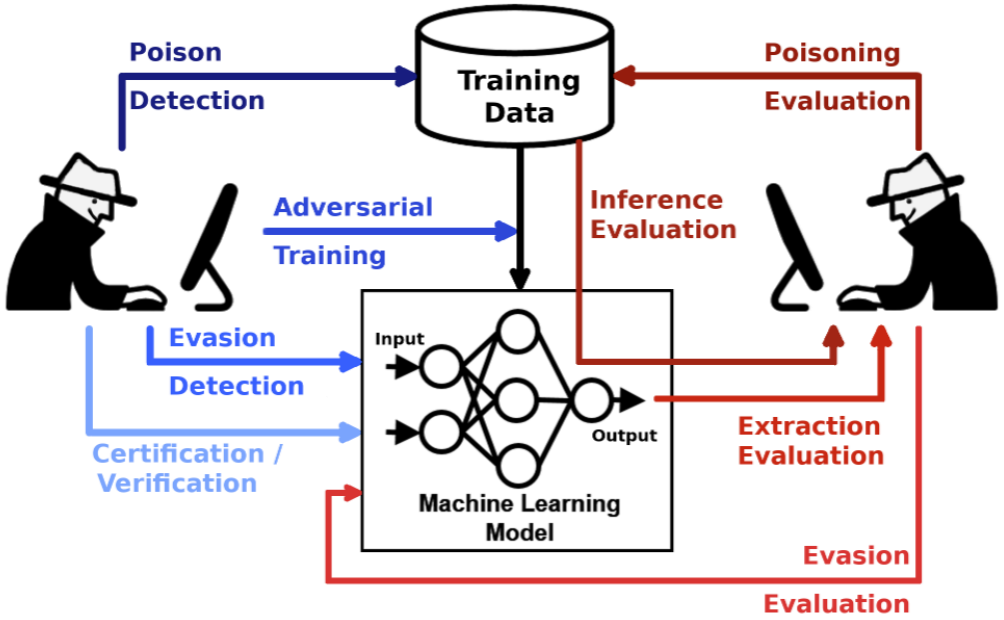
FR System Structure

- This FR system structure will be created using PyTorch to implement the predictor network.



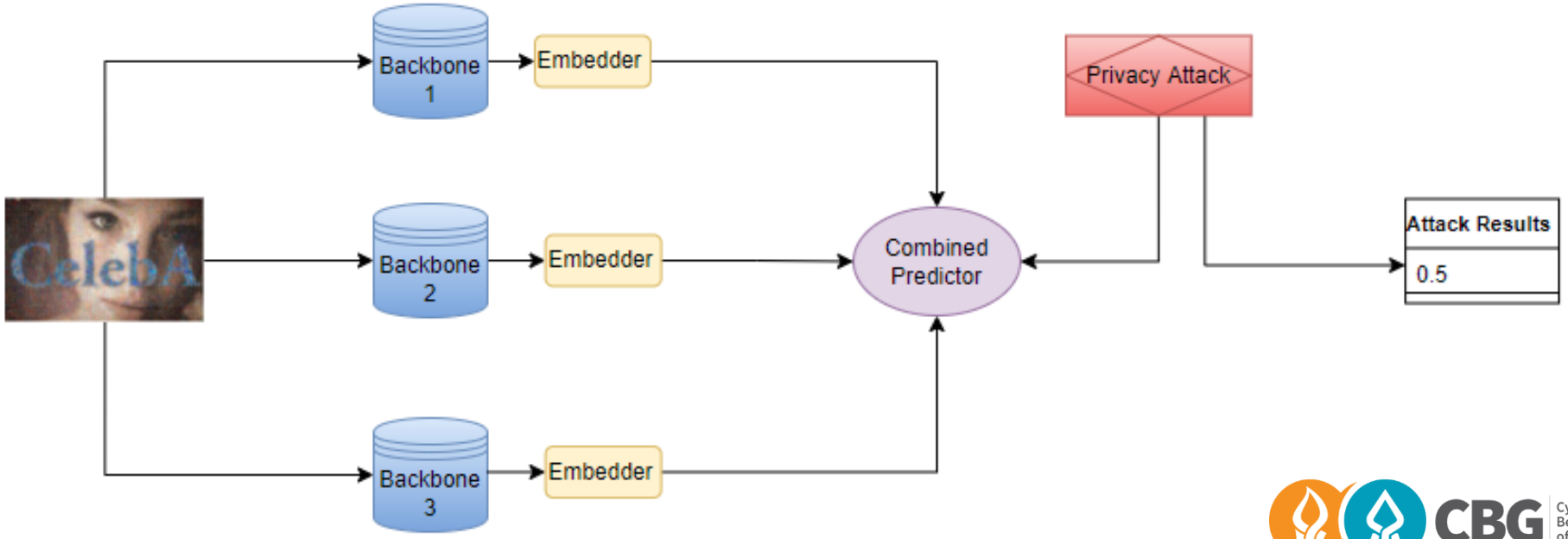
Privacy Violation Attacks Structure

- The implementation of existing privacy violation attacks will be done using the Adversarial Robustness Toolbox (ART) repository.
  - This repository contains existing privacy violation attacks with support for PyTorch models.



Mitigation Method Structure

- The structure of the mitigation method is creating a target model which contains multiple backbones.
  - This is done by creating a predictor which trains on embedding data of multiple backbones.
  - The result is a predictor which trains on different backbones, and this is masking the backbone's influence.



Outline

- ✓ Reminder – Project Proposal
  - ✓ Introduction and Motivation
  - ✓ Problem Definition
- ✓ Project Structure
- ❑ Initial System Results

We are here →

Initial System Results

Initial System Components

- This demo contains the implementation of the FR system and a simple illustration of the mitigation method.
- In addition, we've implemented an existing membership inference attack which attacked our FR system in this demo.
- Overall, the demo has two main components:
  - Demonstrating our system on face verification tasks.
  - Demonstrating our mitigation method against an existing privacy violation attack.

Initial Results

- Our FR model achieved an accuracy score of 0.86 on the LFW dataset.
- When we used the mitigation method, the accuracy score improved to 0.89.

Initial Results

- When performing the membership inference attack on our FR model, the attack accuracy is 0.97.
- However, the attack accuracy on the mitigation method decreased to 0.82.