Real and Stealthy Attacks on State-of-the-Art Face Recognition

Mahmood Sharif, Sruti Bhagavatula, and Lujo Bauer, CMU; Michael K. Reiter, UNC

Introduction

Machine learning (ML) is ubiquitous, enables revolutionary technologies:

If ML fails:

Our research questions investigate robustness of ML algorithms:

Can attackers make ML fail? Can attacks be inconspicuous and physically realizable?

Our Approach and Results

Our focus: DNNs for state-of-the-art face recognition [2]

Attack goals:

- Impersonation: being classified as specific target
- Dodging: not being classified as self

We create realizable, inconspicuous attacks by:

1. Limiting perturbation to eyeglass frames
2. Minimizing total variations (TV) btw. adjacent pixels
3. Minimizing “non-printability score” (NPS)
4. Increasing robustness: an attack should fool the system for more than one face image

Objective for impersonation (dodging is analogous):

\[
\arg\min_r \left( \sum_{x \in X} |f(x + r) - l| + \kappa_1 TV(r) + \kappa_2 NPS(r) \right)
\]

Background and Prior Work

ML classifiers (e.g., in intrusion detection, cancer detection, …) are functions from inputs to classes (or probability distributions over classes)

Imperceptible attacks have been demonstrated that confuse deep neural networks (DNNs) [1], by solving:

\[
\arg\min_r \left[ f(x + r) - l \right] + c|r| \\
\text{misclassification imperceptibility}
\]

x is the input image; \( f(\cdot) \) is the classification function (e.g., DNN); \( l \) is the desired output class; \( r \) : perturbation (or change applied to the input).

Our Approach and Results

Alg. to generate accessories

Region to perturb

Images of attacker

Target DNN

Attack generation:

Results: fool DNN trained on 7 subjects + 3 authors

Impersonation

Sruti → Milla Jovovich 88% success

Sruti → Mahmood 88% success

Dodging

Sruti → Lujo 100% success

Sruti → Not Lujo 97% success

In paper: more experiments with larger DNN*