精读

On the Resilience of Biometric Authentication Systems against Random Inputs (NDSS'2020)

Summary

This paper analyzes the volunerability of the machine learning-based biometric authentication systems that the attacker has the probability to bypass these systems by taking a small number of attempts (i.e., uniform random inputs) if he knows the feature space of these authentication system. The authors provide detailed analysis about this volunerability on four different biometric modalities and four different machine learning classifiers. The results indicate that the success rate of the attack can be higher than the FPR of the biometric system, so the authors suggest that FPR and FRR cannot alone to assess the security of the system.

Assumptions

- The attacker has access to the biometric system via a blackbox API and can launch a series of random inputs to the system.
- The attacker has knowledge about the length of the input feature vector and permissible range of each feature value (if not normalized).
- Other assumptions not specifically demonstrated: binary classification, min-max normalization, the identifiers of legitimate user (e.g., username).

Evaluation

- Four biometric schemes
    - Face recognition
    - Voice authentication
    - Gait authentication
    - Touch (swipe) authentication
- Classifiers
    - SVM (Linear and radial)
    - Random forests
    - DNN

Strengths

- The paper talks about a random input attack on biometric authentication system and gives detailed analysis with experiments.
- The idea of this paper is clear and well-demonstrated: the region of accepting region is significant larger than the true positive region of a specific user.
- The paper is well-orgnized and gives a full analysis with extensive experiments.

Drawbacks

- The ability of acquiring the propoties of the input feature vector is a strong assumption considering that most of biometric authentication system is not the case, e.g., the size of image for a face authentication system.
- The attacker has to send a series of random input to the authentication system so as to acquire the boundary of the classifier which may not permitted by the authentication system with attack detection.
- The random input attack based on the fact that the attack has knowledge about the classification boundary which is not easy to require. The authors do not have a discussion about the effects the attack has to take about this.
- The authors only consider the min-max normalization without other normalizing methods (e.g., zero-score normalization). The theoretical model may be different.
- Alien detection can mitigate this attack considering that it only works on binary classification?

泛读

Practical Black-Box Attacks against Machine Learning (Asia CCS'17)

Summary

This paper introduce the first practical demonstration of an attacker controlling a remotely hosted DNN with no such knowledge. The only capa- bility of our black-box adversary is to observe labels given by the DNN to chosen inputs. The attacker can training a local substitude to generate adversaral samples that are misclassified by the user's model with high probability (about 84%).

Strength

- The first blackbox attack on DNN models via a novel substitute training algorithm using synthetic data generation, to craft adversarial examples misclassified by black-box DNNs.
- Practical attack on COST platforms, e.g., MetaMind, and models deployed on Google and Amazon.

Stealing Machine Learning Models via Prediction APIs (Usenix'16)

Summary

The authors propose a model extraction attack motivated by the tension between model confidentiality and public access. The adversary with blackbox access, but no prior knowledge of an ML model’s parameters or training data, can duplicate the functionality of the model with near-perfect fidelity(the meaning of "steal" in this paper). The authors demonstrate these attacks against the online services of BigML and Amazon Machine Learning. The results show that the natural countermeasure of omitting confidence values from model outputs still admits potentially harmful model extraction attacks.

Category of Extraction mothods

- Extraction with confidence values
- Extraction with given class abels only

Target models

- Linear binary models
- Multiclass LR models
- Neural networks
- RBF kernel SVMs