인공지능 보안 이슈

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/e239af44-5330-4294-83c7-b9c079f2d7bc/%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5_%EB%B3%B4%EC%95%88_%EC%9D%B4%EC%8A%88.pdf

박소희, 최대선 - 2017

1. 서론

- 패턴 인식이 필요한 보안 문제에서도 기계학습 기술을 활용한 연구가 많이 등장
- 기계학습 중심으로 인공지능 보안 이슈 살펴볼 예정
- Possion attack, evasion attack 등 인공지능 속이는 공격 형태
- 인공지능 모델 자체를 탈취할 수 있는 model extraction attack
- 학습된 모델에서 데이터 추출해내는 inversion attack

2. adversarial ai

a. posion attack

- 잘못된 데이터를 제공하여 잘못된 분류 결과를 내도록 하는 것
- 최소한의 데이터로 최대한의 오작동 목표
- 가정 : 공격자가 학습 알고리즘을 알고 있는 경우 원 데이터에 접근 가능하여 시뮬레이션 가능

1. posion data 생성

- gradient ascent 방법 통해 모델의 loss 를 극대화 하는 x를 구하는 방식
- 각 class 속성 값 분포 분석 후 비율 높은 분포로 학습 데이터 생성 후 다른 클래스로 학습
- posion data 추가될수록 정확도가 떨어짐

b. evasion attack

- 학습데이터가 아닌 입력 데이터를 약간 변조하여 다른 class로 인식되도록

- 학습된 이후 입력 데이터 조작 통해 속이는 공격이므로 적용 범위가 훨씬 넓음

1. image
    - 사람의 눈으로는 거의 차이점을 발견하지 못할 정도로 약간의 변조
    - 그러나 전혀 다른 모양으로 인식해 다르게 분류
2. voice
    - 변조한 음성을 사용자는 알아들을 수 없음
    - 음성인식 시스템에서는 올바른 명령으로 인식되도록 목표
    - 사용자가 알아들을 수 없는 소리에 의해 사용자 의도와 무관하게 휴대전화의 특정 기능이 작동
    - Audio Mangler 통해 변조
    - Audio Mangler는 음성 명령 MFCC 특징 추출 후, MFCC 벡터 변조하여 역 MFCC 적용하여 변조된 음성 생성

c. model extracion attack

- 모델 모사는 고비용 구축 모델 쉽게 탈취 가능하며 다른 attack 용이
- 반복적으로 쿼리 x 입력하고 답인 f에 가까운 f^만드는 것을 목표
- class와 confience value 제공 경우가 많아 이를 활용해 모델 내부 parameter 추정해 f^ 생성

3. data privacy

a. inversion attack

- 모델에 쿼리를 하여 학습 데이터 재현해내는 공격을 inversion attack
- 공격자가 알고 싶은 feature의 모든 값을 시험해 y 예측 오차가 가장 작은 값 선택

b. attack on data sanitizing

- data sanitizing : 비정형 데이터의 경우 제거나 암호화 하는 등 마스킹 작업
- 노출있는 데이터를 학습 데이터로 활용해 미탐 사례만을 탐지하는 모델 만들 수 있음

의문

1. model extracion attack과 inversion attack은 같은 것인지?
2. 미탐 사례?가 무엇이고 어떻게 만드는 것인지
3. posion data 에서 저 논문에서 말하는게 FGSM 인건지?

느낀 점

1. IoT 시장 커지므로 음성 변형 중 단어만 변형 시킬 순 없는 건지? (turn off → turn on)