자율주행차에서 군사 감지까지: 양자 컴퓨팅과 AI 시스템

(2023.06.07., 양자정보연구지원센터)

□ 머신러닝 모델의 취약성에 대한 잠재적 솔루션

○ 머신 러닝 알고리즘은 악의적인 공격에 얼마나 안전한가

- 인공 지능 알고리즘으로 강력한 보안이 필요한 많은 시스템은 안면 인식, 은행 업무, 군사 표적 애플리케이션, 로봇 및 자율 차량 등 포함
- 이러한 모델에 양자 컴퓨팅 통합하여 적대적 공격에 강력한 복원력을 가진 새로운 알고리즘 생성 가능성 제안

○ **데이터 조작 공격의 위험**

- 머신 러닝 알고리즘은 이미지 기능 분류 및 식별에 유용하지만, 심각한 보안 위험 초래할 데이터 조작 공격에 매우 취약함
- 이미지 데이터의 미세한 조작 포함하는 데이터 조작 공격은 알고리즘 훈련에 사용되는 훈련 데이터 세트에 손상된 데이터 혼합하여 공격 시작
- 조작된 데이터는 AI 시스템이 사용 중, 기본 알고리즘을 계속 훈련하는 경우 테스트 단계 중 주입될 수 있음, 예) 자율 주행 자동차가 손상된 머신 러닝 알고리즘을 사용하는 경우, 도로에 사람이 있음에도 없다고 잘못 예측할 수 있음

○ 양자 컴퓨팅이 도움 되는 방법

- 양자 컴퓨팅을 머신 러닝과 통합, **양자 머신 러닝 모델**(quantum machine learning model)이라는 보안 알고리즘 생성 방법
- 조작이 쉽지 않은 이미지 데이터의 특정 패턴을 찾을 수 있는 특수 양자 속성 활용하도록 설계, 그 결과 강력한 공격에도 안전한 알고리즘 사용 가능

○ 동작 원리

- 고전 물리학 법칙을 따르는 고전 컴퓨터와 달리, 양자 컴퓨팅은

양자 물리학 원리를 따르며, 정보는 0, 1 또는 둘의 조합으로 동시에 존재할 수 있는(중첩 상태) 큐비트에 저장되고 처리됨, 양자 컴퓨터는 이 속성을 이용한 알고리즘 설계에 사용

- 양자 머신 러닝 모델은 많은 민감한 응용 프로그램에 중요한 보안 제공

ㅇ 극복해야 할 한계

- 오늘날의 양자 컴퓨터는 상대적으로 작고(500큐비트 미만) 오류율이 높음, 큐비트의 불완전한 제작, 제어 회로의 오류 또는 환경과의 상호 작용 통한 정보 손실(양자 결맞음)을 비롯한 오류 발생
- 양자 하드웨어 및 소프트웨어에서 엄청난 발전을 보임, 최근 양자 하드웨어 로드맵에 따르면 양자 장치는 수백에서 수천 큐비트 가질 것으로 예상됨

□ 적대적 기계 학습(adversarial machine learning)

ㅇ 적대적 공격 유형

- 중독 공격(poisoning attacks): 모델 훈련에 사용되는 데이터에 초점, 공격자는 기존 데이터 변경하거나 잘못 지정된 데이터 도입
- 회피 공격(evasion attacks): 모델 자체에 초점, 합법적인 것처럼 보이지만 잘못된 예측으로 이어지도록 데이터 수정 작업 포함
- 모델 도용(model stealing): 훈련된 모델에 집중, 공격자는 모델의 구조 또는 모델 교육에 사용되는 데이터에 대해 알고 싶어함

ㅇ 공격에 대처하는 방법

- 머신러닝 시스템 방어 방법은 모델 유형에 따라 다름, 선형 회귀 또는 로지스틱 회귀 같은 간단한 모델로 많은 문제를 해결할 수 있음
- 적대적 훈련(adversarial training), 모델 전환(switching models), 일반화 모델, 일반 보호 조치 및 책임 있는 AI

(원문)

1. https://phys.org/news/2023-05-self-driving-cars-military-surveillance-quantum.html