AI 보안의 문제점과 한계

AI security problems and limitations

인제대학교 컴퓨터공학부 이정훈 (Lee Jung Hun)
영남대학교 컴퓨터공학과 김동현 (Kim Dong Hyeon)
동명대학교 정보보호학과 박수곤 (Park Soo Gon)

INDEX

서론
Introduction

1. 서론

1.1 인공지능보안

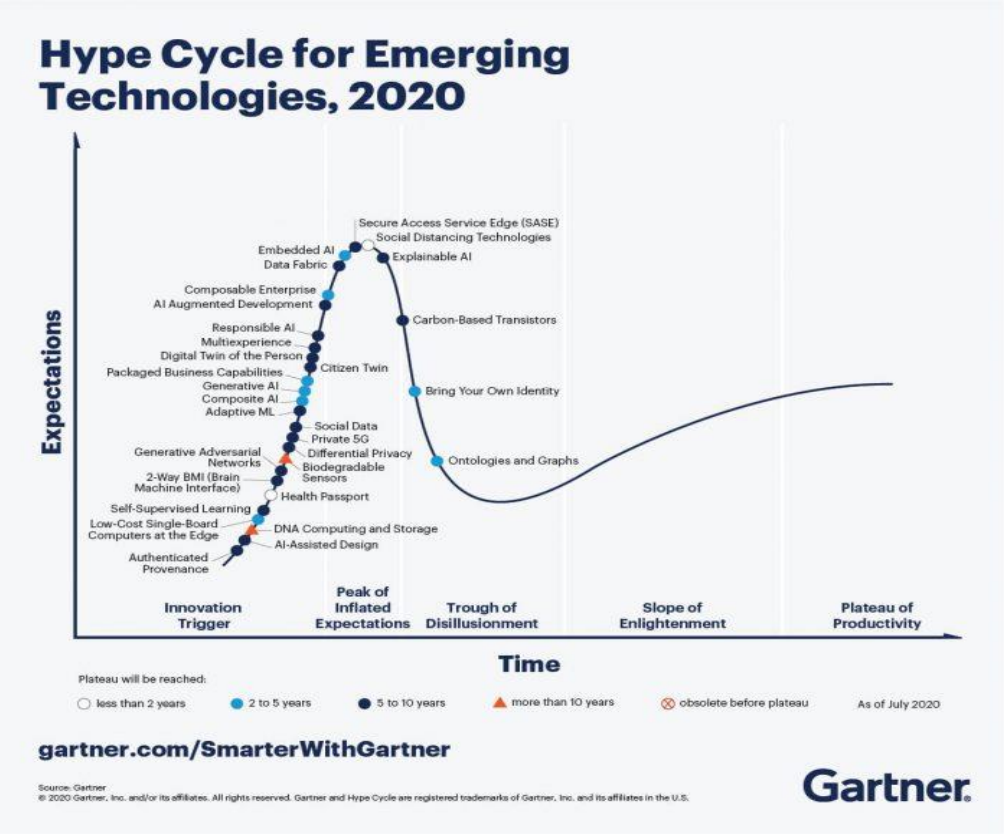
AI 기술이 발전함에 따라서 AI는 제조 및 금융, 의료 산업 및 다양한 분야에서 사용될 것으로 전망

현재까지의 AI의 기술은 전문가를 대체할만한 수준이 안됨

실제로 학습에 걸리는 시간정확도의 효율성에 관해서 기존의 공격은 수동적이었기에, 적대적 AI 학습으로 인한 **AI 보안 모듈이 아닌 CVE나 Payload를 적용하여 자동화를 통한 방어와 AI보안의 문제점 및 한계점에 관해 연구의 필요성이 제기됨**

1. 서론

1.1 인공지능보안



가트너 하이프 사이클
(떠오르는 기술을 비롯해 다양한 기술 분야를 대상으로 매년 발행하는 하이프 사이클)

가트너에서 선정하는 **"10대 미래 전략 기술"**에도 인공지능이 매년 자리하고 있음

기존의 보안 솔루션들은 인공지능을 접목하는데 주목

인공지능으로 모든 것을 다 처리 한다면 그에 따른 문제점과 한계가 존재함

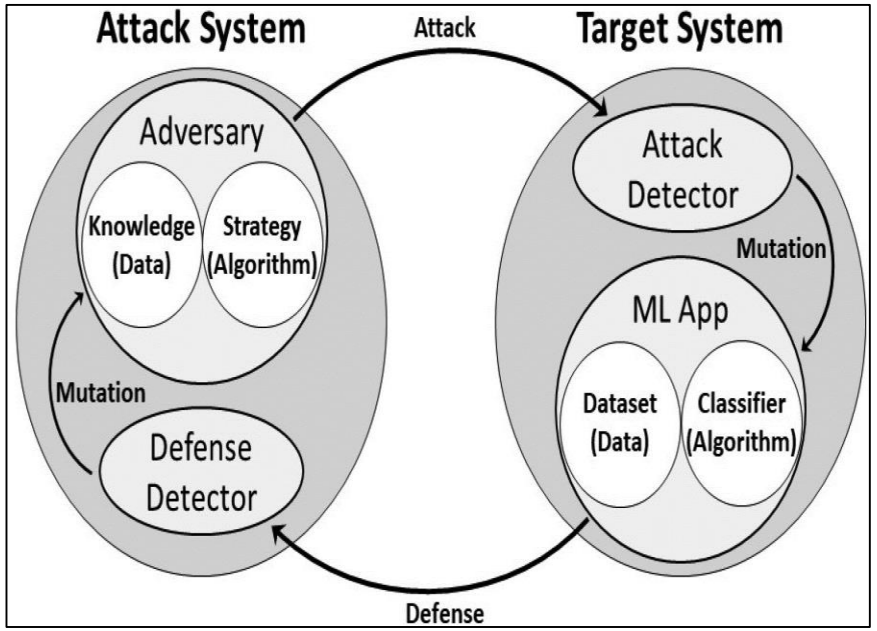
본론
Main subject

2. 본론

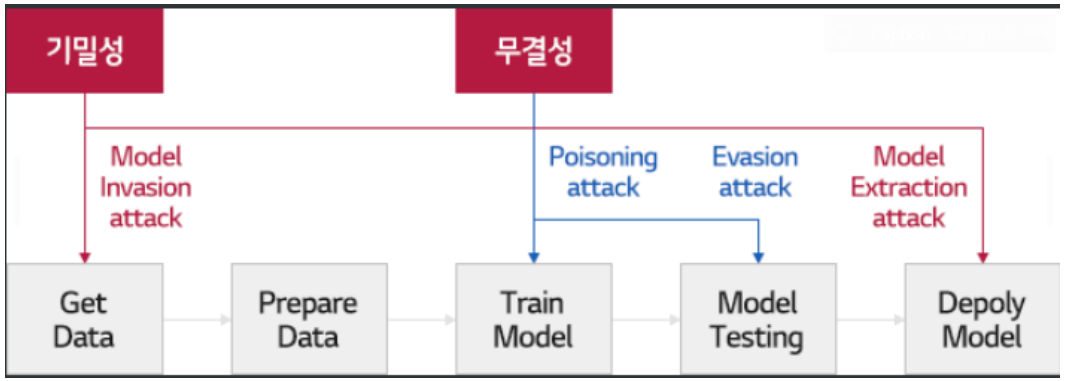
2.1 적대적 기계학습

- **Adversarial-machine-learning :** 의도 되지 않는 Input으로 모델을 속이는 머신러닝 기술

Adversarial Machine Learning Cycle



머신러닝 학습과 적대적 공격 유형



CISC-W'20
Conference on Information Security and Cryptography-Winter 2020

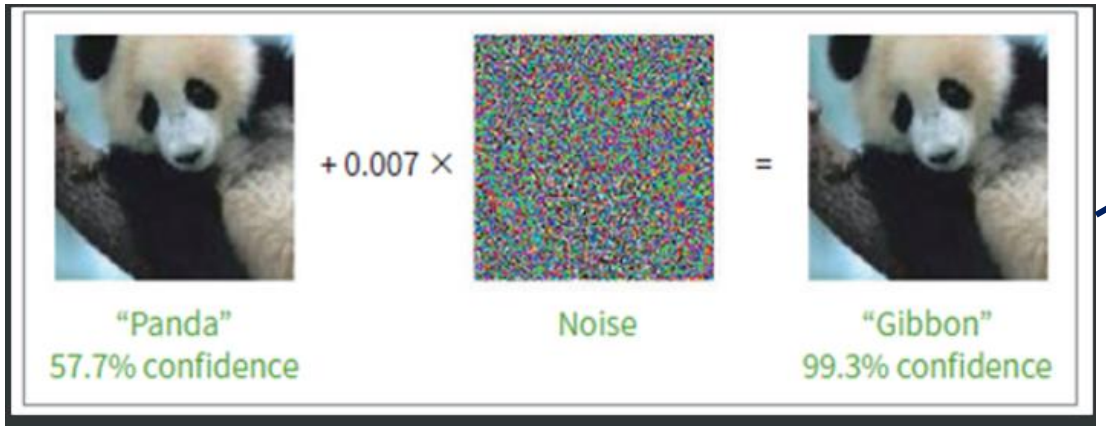
2. 본론

2.2 활용예시

Fast Gradient Sign Method (FGSM)

적대적 공격을 위한 샘플을 만드는데
가장 간단하고 계산적으로 효율적인 방법

연구에서 벤치마킹을 위한 baseline 모델로서 주로 사용



노이즈를 통하여 판다를 긴팔원숭이로 인식

2. 본론

2.2 활용예시

One-Step Target Class Methods

FGSM의 대체 알고리즘 중의 하나 표적공격을 하므로 Targeted FGSM

한 특정 레이블 y target에 속할 가능성이 적은 어떠한 이미지 X에 대해서 그 이미지가
특정 레이블에 속할 확률 p (y target | X)을 최대화하는 방향으로 perturbation을 추가하는 학습을 진행함
유사하지 않거나 전혀 연관성이 없어 보이는 레이블로 오인하도록 표적공격을 수행

DeepFool

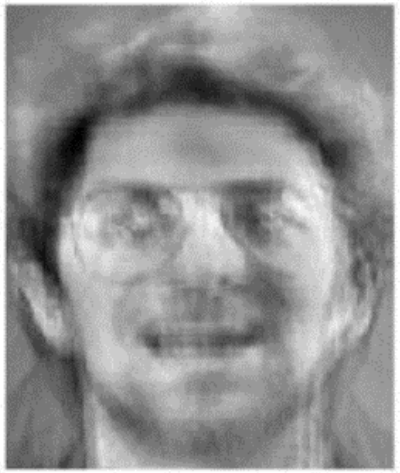
탐욕 알고리즘이기 때문에 다소 느리지만
Jacobian-Based Saliency Map Attack 알고리즘보다 빠르고, FGSM
알고리즘보다 정확하기 때문에 유용하게 활용

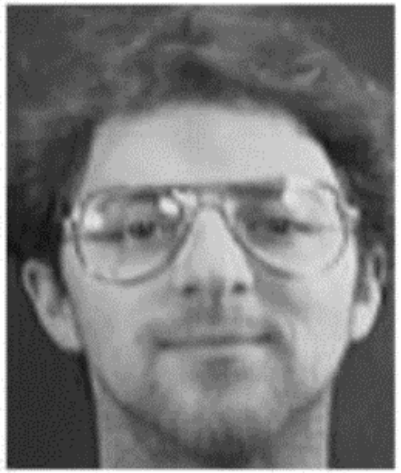
2. 본론

2.2 활용예시

Jacobian-Based Saliency Map Attack (JSMA)

주의 맵(attention map)인 saliency map을 gradient 를 기반으로 구성





실제 학습 데이터,

왼쪽 - Inversion attack을 이용해 재현된 이미지

2. 본론

2.2 활용예시

NewtonFool

입력 영상에서 화소 단위로 특정 레이블에 높은 확률로 속하는 화소와, 반대로 해당 레이블에 높은 확률로 속하지 않는 화소를 뉴턴 알고리즘을 통해 찾는 방법

이미지 교란으로 STOP 표지판 오인식 유도

결론
Conclusion

3. 결론

3.1 불확실성과 적대적 공격

중독 공격 (Poisoning)

적대적 사례가 모델 학습 단계에 적용되어 훈련 중 모델을 방해하려고 시도하는 경우

회피 공격 (Evasion)

모델의 추론(inference) 단계에서도 고의적으로 모델이 오작동을 일으키도록 사용

3. 결론

3.1 불확실성과 적대적 공격

딥러닝 학습모델에 대한 공격 시나리오는 공격자가 공격의 대상이 되는 모델에 대해 가진 정보의 양에 따라 다름 (화이트 박스 / 블랙 박스)

화이트박스(white box)

학습 데이터를 알고 있는 상황에서 적대적 공격을 수행

블랙 박스(black box)

공격과 공격자가 모델에 대한 내부구조를 파악하지 못한 상태에서 수행

참고 문헌
Reference

4. 참고 문헌

권현, 윤현수, 최대선 (2018). Evasion attack에 대한 인공지능 보안이슈. 정보과학회지, 36(2)

http://news.khan.co.kr/kh_news/khan_art_view.html?art_id=201609211731001 | 자율주행 전기차, 해킹에 '원격조종' 당했다.

이상근, 인공지능의 오동작 유도 및 방어 (금융보안원 전자금융과 금융보안 20제17호)

김휘영, 정대철, 최병욱, 딥러닝 기반 의료 영상 인공지능 모델의 취약성 : 적대적 공격 (대한영상의학회지 2019.3)

권현, 방승호, 인공지능(AI)과 정보보호 측면에서의 군 도입시 고려사항 (월간국방과 기술 2019.8) http://bemil.chosun.com/nbrd/bbs/view.html?b_bbs_id=10008&num=180

국경완, 공병철, 인공지능을 활용한 보안기술 개발 동향

Battista Biggio, giorgio fumera, Paolo Russu, Luca Didaci, Fabio Roli, Adversarial Biometric Recognition : A review on biometric system security from the adversarial machine-learning perspective, IEEE Signal Processing Magazine (Volume: 32 , Issue: 5 , Sept. 2015)

BattistaBiggio, Fabio Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Received 8 December 2017, Revised 29 June 2018, Accepted 16 July 2018, Available online 21 July 2018.

Sadeghi, K., Banerjee, A., Gupta, S.K.S., A System-Driven Taxonomy of Attacks and Defenses in Adversarial Machine Learning, IEEE Transactions on Emerging Topics in Computational Intelligence IEEE Trans. Emerg. Top. Comput. Intell. Emerging Topics in Computational Intelligence, IEEE Transactions on. 4(4):450-467 Aug, 2020

https://www.sedaily.com/NewsVIew/1L07C0W5EU
서울경제 AI를 속이는 방법

Thank You