2020 IT 21
Global Conference
Digital New Deal
Technology Essentials
디지털 뉴딜 기술 핵심

Session 4-1

Towards Secure AI

이상근 교수 (고려대학교)

[요약문]

AI는 학문적 연구의 범위를 넘어 사회 전 영역에서 사용 가능한 GPT (general purpose technology)로 성장하고 있습니다. 하지만 AI가 GPT로서 사용되기 위해서는 아직 갖추어야할 조건들이 있으며, 특히 스마트 팩토리, 자율주행 자동차, 보안 등 미션-크리티컬한 분야에서 신뢰성이 보장되지 않는 AI를 적용하기는 어렵습니다. 본 강연에서는 오동작 유도 공격 (adversarial attack), 데이터 오염 공격, 모델 탈취 등 현재 알려져 있는 AI의 신뢰성 이슈들에 대해 소개하고, 신뢰성있는 AI를 이루기 위해 앞으로 연구해야 할 방향에 대해 함께 생각해 보고자 합니다.

[발표자 약력]

2003년 서울대학교 컴퓨터공학과 학사 (과수석 졸업)
2005년 서울대학교 전기컴퓨터공학부 석사
2011년 미국 University of Wisconsin-Madision 박사
2011년~2014년 독일 TU Dortmund 대학 협력연구센터 (SFB876) 포닥연구원
2015년~2017년 독일 TU Dortmund 대학 협력연구센터 (SFB876) 프로젝트 리더
2017년~2019년 한양대학교ERICA 소프트웨어학부 조교수
2020년~현재 고려대학교 정보보호대학원 조교수

관심분야 : AI 보안, AI 모델 압축, 산업/의료 영상 분석, Edge-computing 환경 연합/분산 기계학습, AutoML, XAI

AIR LAB

고려대학교 KOREA UNIVERSITY

Towards Secure AI

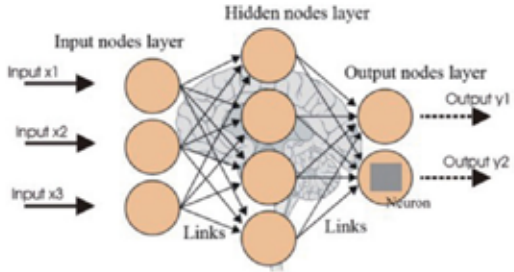
고려대학교 정보보호대학원
Artificial Intelligence Research (AIR) LAB
https://air.korea.ac.kr/

이상근

IT21 글로벌 컨퍼런스 (2020.9.25)

Toward a Connected World

- Success of AI
  - Massive amount of data
  - Scalable computer & software systems
  - The broad accessibility of AI techniques

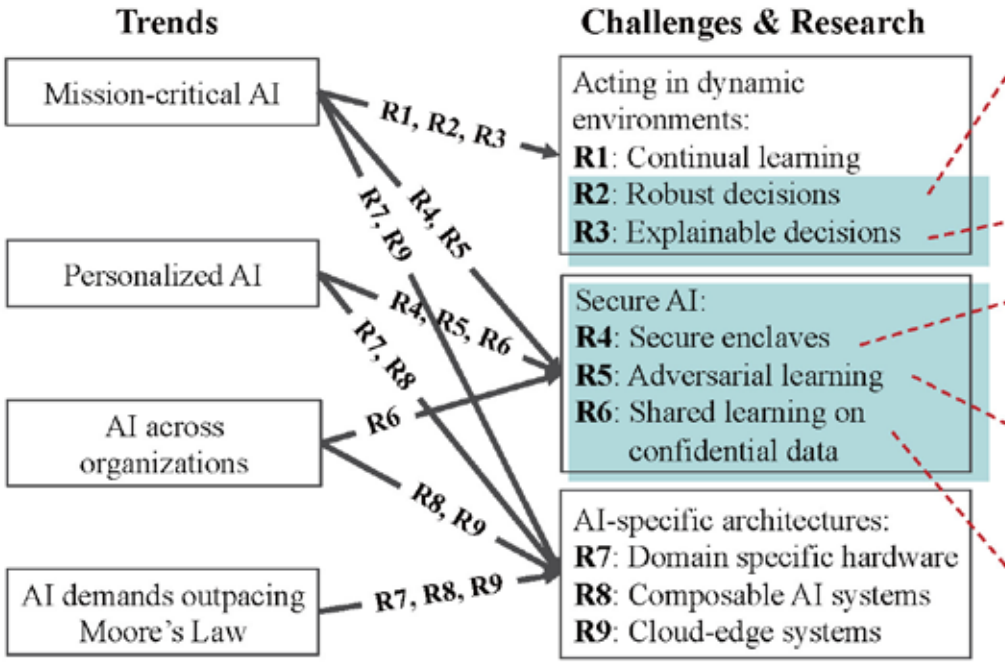


- By 2025 [Global Industry Vision 2025, Huawei]
  - 77% of the global population being connected (100 billion connections)
  - 85% of enterprise applications on the cloud
  - 12% of families use smart home robots

Why AI?

- Artificial Intelligence
  - Before AI: complex deterministic rules crafted by humans
  - AI can learn simple probabilistic decisions from data
- Major pitfalls of legacy systems
  - Based on rules made out of human understanding of things
  - Hard to adapt to constantly changing patterns at large scale
- Ex. AI for cybersecurity
  - AI can adapt to changing patterns, finding new threats on the fly
  - Facilitates automatic defense in scale
  - Cooperation with human experts
- Secure AI is essential
  - AI must be trustworthy and reliable

AI Trends & Challenges



R2: automated decision making
- 데이터의 근원/변화 학습
- 인과관계 학습

R3: XAI

R4: Hardware enclaves
- Intel SGX, ARM TrustZone

R5:
- **Robust learning against adversarial inputs**

R6:
- Federated learning
- Differential privacy
- Homomorphic encryption

[Berkeley View of Systems Challenges for AI, Tech Rep. 2017]

Evasion

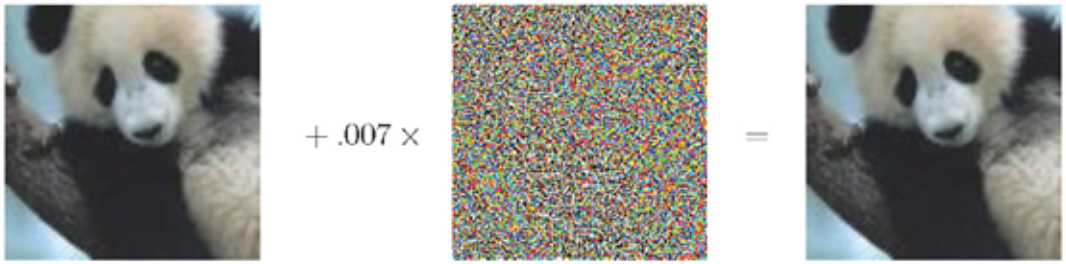
Training:

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, f(w, x_i))$$

Adversarial Attack:

$$\max_{\Delta x} \ell(y_i, f(w, x_i + \Delta x)) \quad \text{s.t. } \|\Delta x\|_p \leq \epsilon$$

$$\min_{\Delta x} \ell(y_{\text{target}}, f(w, x_i + \Delta x))$$



x

"panda"
57.7% confidence

$\text{sign}(\nabla_x J(\theta, x, y))$

"nematode"
8.2% confidence

$x + \epsilon\text{sign}(\nabla_x J(\theta, x, y))$

"gibbon"
99.3 % confidence

[Szegedy et al, Intriguing properties of neural networks, arxiv 2013]

Adversarial Examples: First Appearance

- Szegedy et al., Intriguing properties of neural networks. ICLR, 2013
- Finds a minimum-size adversarial perturbation:

ℓ_2-norm Attack

- Goodfellow, Shlens, & Szegedy. Explaining and harnessing adversarial examples. ICLR, 2015

FGSM

- Goodfellow, Shlens, & Szegedy. Explaining and harnessing adversarial examp les. ICLR, 2015

$$\min_{\Delta x:\|\Delta x\|\leq\epsilon} \ell(y_T, f(w, x_i + \Delta x))$$

- FGSM
  -
  - Consider the 1st order Taylor approximation of the loss w.r.t. :
  - The optimal solution is:

Iterative FGSM

- Based on the projected gradient descent (PGD) optimization:

$$x_{t+1} = \text{Proj}_{\epsilon}[x_t + \eta_t \text{sgn}(\nabla_x \ell(y, f(w, x_t)))]$$

where projection is on the set: (→ simple clipping).

- Madry et al., Towards deep learning models resistant to adversarial attacks. ICLR, 2018
- Raghunathan, Steinhardt, & Liang. Certified defenses against adversarial examples. ICLR, 2018
- Wong & Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. ICML 2018

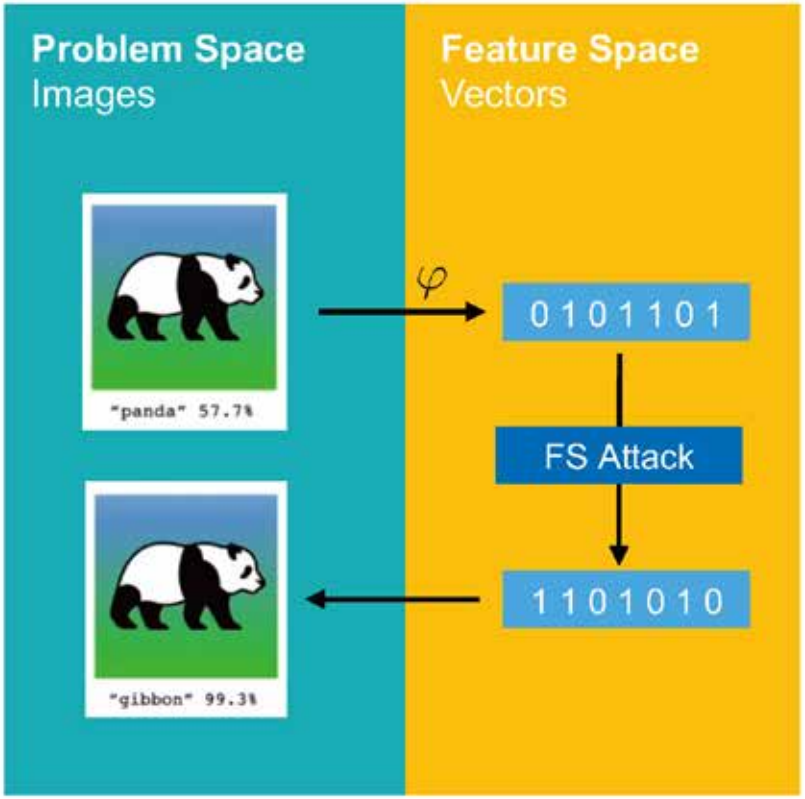
CW

- A problem in FGSM & i-FGSM:

$$\min_{\Delta x:\|\Delta x\|\leq\epsilon} \ell(y_T, f(w, x_i + \Delta x))$$

- Carlini and Wagner. Towards evaluating the robustness of neural networks. S&P 2017

Problem Space vs. Feature Space



Adversarial Text

The APM 20 Lionceau is a two-seat very light aircraft manufactured by the French manufacturer Issoire Aviation. Despite its classic appearance it is entirely built from composite materials especially carbon fibers. Designed by Philippe Moniot and certified in 1999 (see EASA CS-VLA) this very light (400 kg empty 634 kg loaded) and economical (80 PS engine) aircraft is primarily intended to be used to learn to fly but also to travel with a relatively high cruise speed (113 knots). ***Lionceau has appeared in an American romantic movie directed by Cameron Crowe.*** A three-seat version the APM 30 Lion was present-ed at the 2005 Paris Air Show. Issoire APM 20 Lionceau

Figure 4: An adversarial text sample generated by inserting a forged fact (99.9% *Means of Transportation* to 90.2% *Film*).

Maisie is a comedy f*li*m property MGM originally purchased for Jean Harlow but before a shooting script could be completed Harlow died in 1937. It was put on hold until 1939 when Ann Sothern was hired to star in the project with Robert Young as leading man. It is based on the novel Dark Dame by Wilson Collison. It was the first of ten films starring Sothern as Maisie Ravier. In Mary C. Maisie (film)

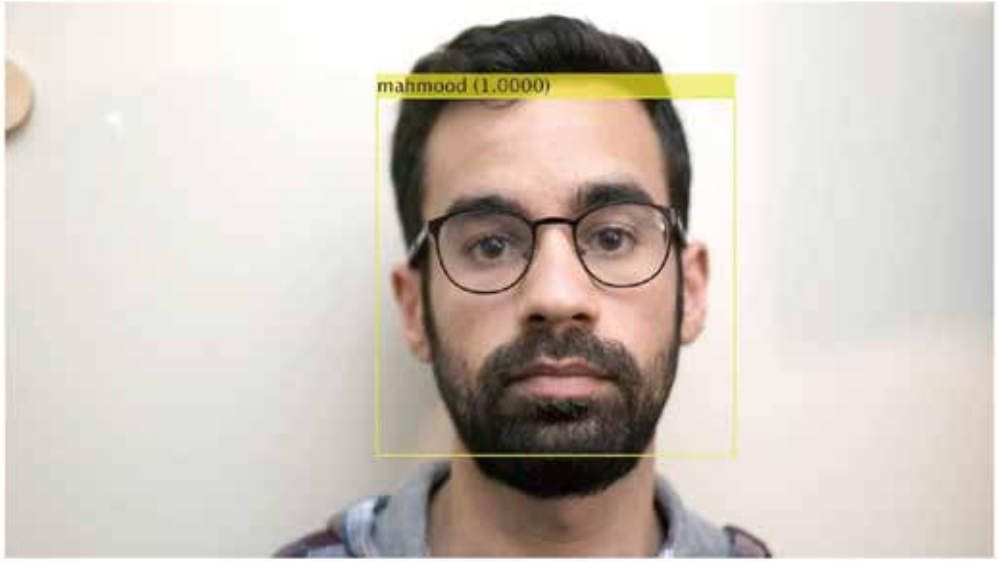
Figure 5: An adversarial text sample generated by introducing a common misspelling (99.6% *Film* to 99.0% *Company*).

[Liang et al., Deep Text Classification Can be Fooled, IJCAI 2018]

Adversarial Images / Videos

[Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR 2018]

[Sharif et al., Adversarial Generative Nets: NN Attacks on SOTA Face Recognition, 2017]



Adversarial Malware

Android malware

- [TDSC`17, ESORICS`17, ACSAC`19, SP`20]

Windows malware

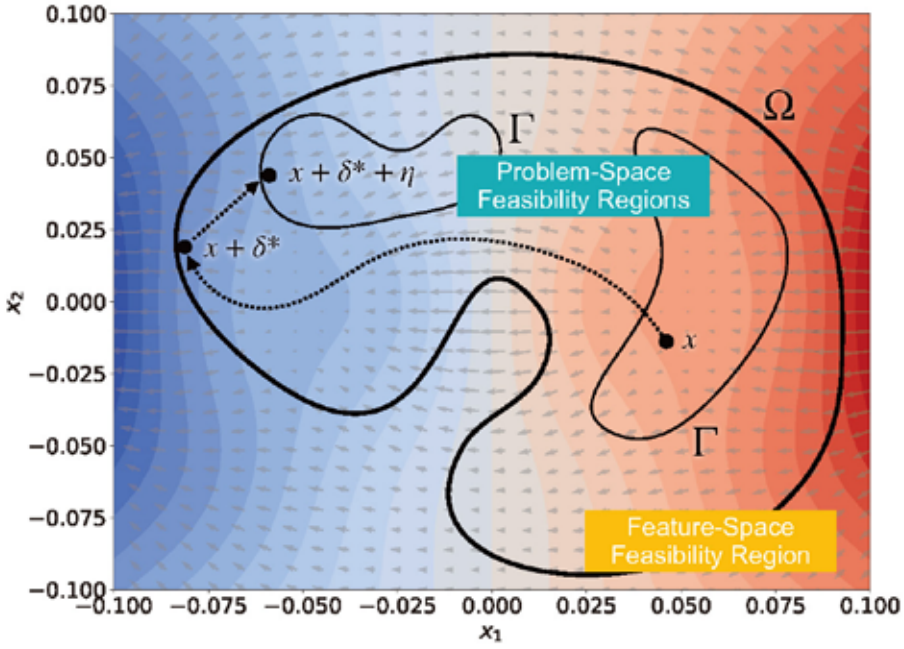
- [RAID`18, EUSIPCO`18]
- AIRLab + NSR`18

PDF malware

- [ECML-PKDD`13, NDSS`16]

Network traffic

- [NCA`18, NCA`19]



[Pierazzi et al., SP 2020]

Data Poisoning

Microsoft Tay chatbot

- Released on Twitter to the public (2016. 3. 23)
- Designed to learn from dialogues, emulating the style of a teenage girl
- Shut down 16 hours of its launch

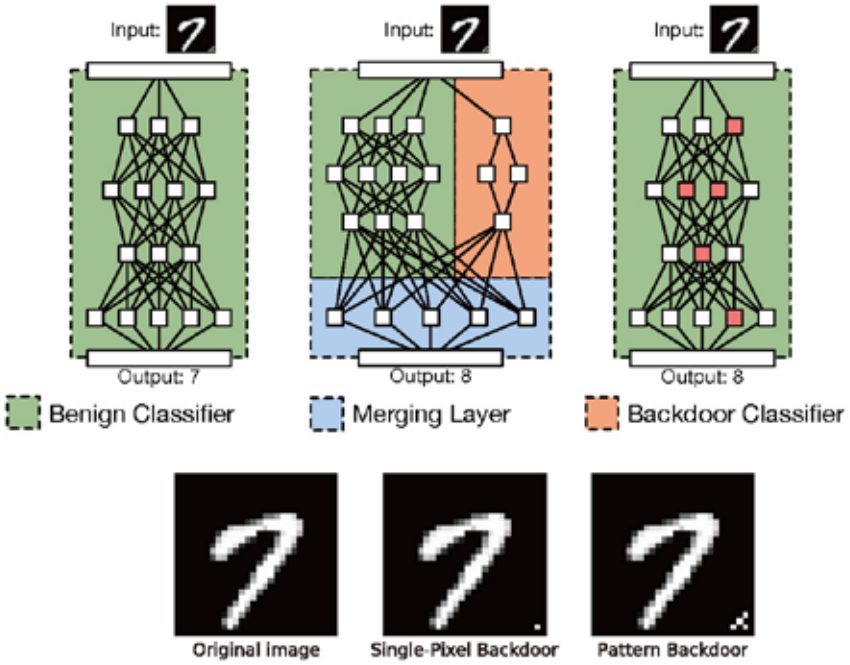
Big Potential Threat

- IDS continuously collect samples and retrain to detect new attacks

Backdoors / Trojans

[Gu, Dolan-Gavitt & Garg
NIPS MLSec Workshop, 2017]

- Backdoors to NNs are implanted by adding specific neurons



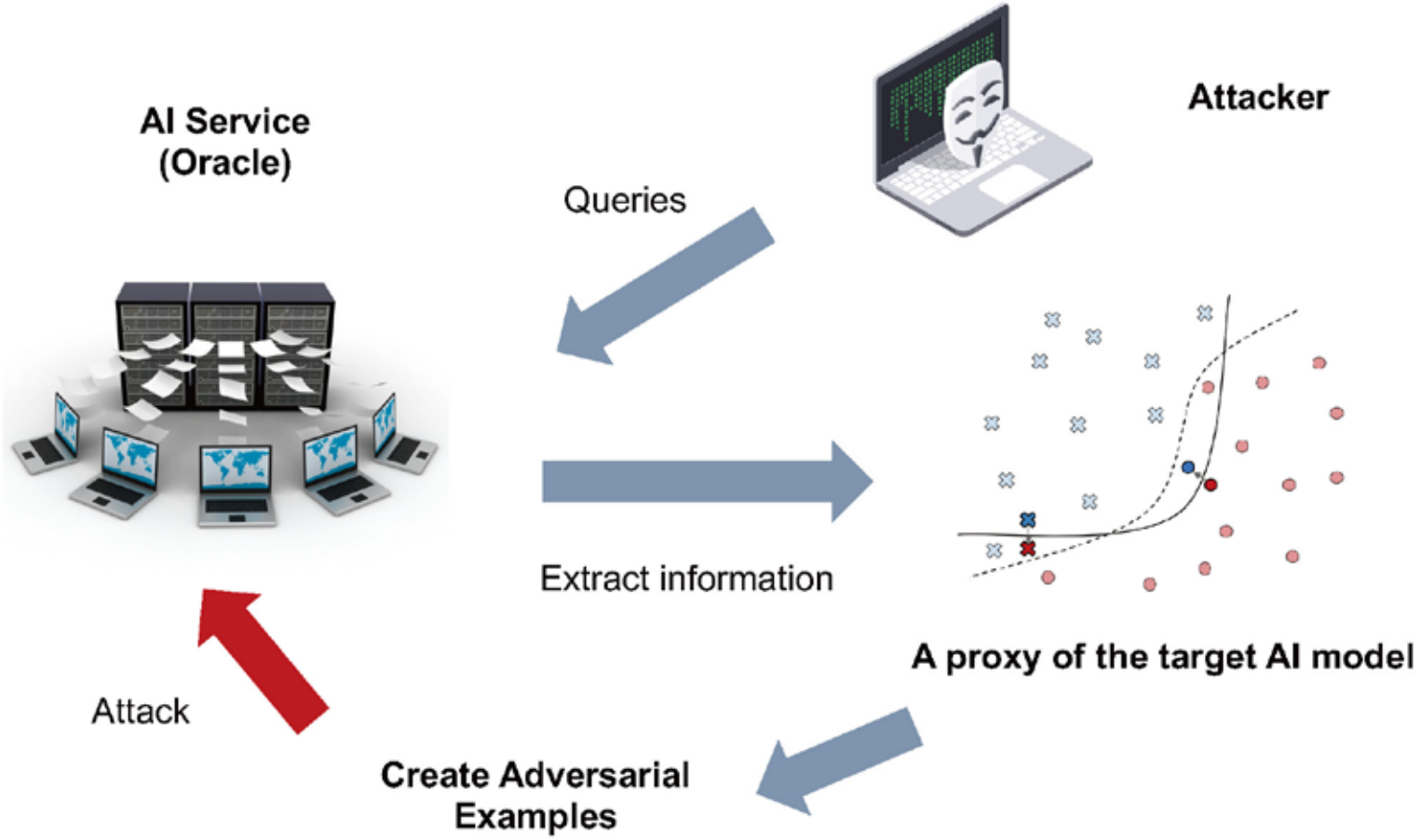
TrojAI

- US Army Research Office
- Intelligence Advanced Research Projects Activity
- CFP: 2019.05.02

Adversary puts a sticky note on a stop sign → AI says it's a speed limit sign. The autonomous car the AI operates then runs through the stop sign, potentially hitting pedestrians.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," ArXiv:1708.06733 [Cs], August 22, 2017, http://arxiv.org/abs/1708.06733.

AI Model Stealing

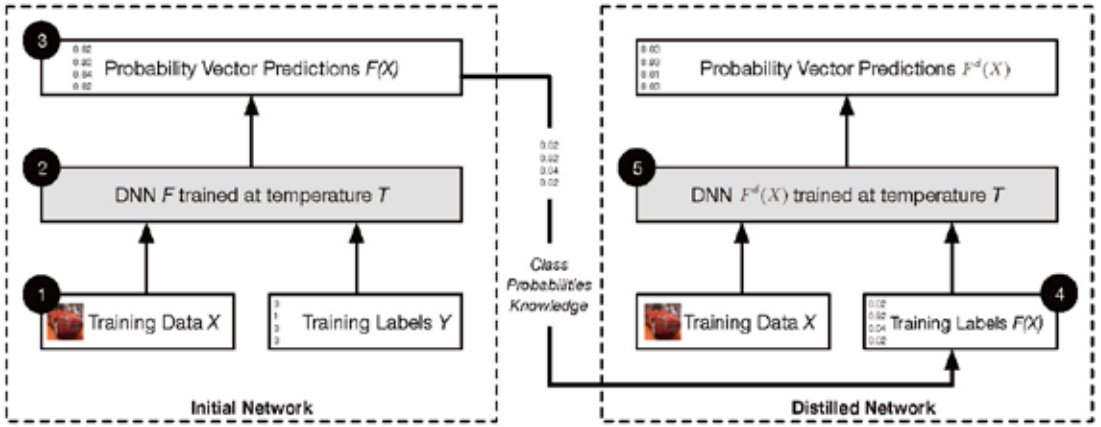


[Papernot, McDaniel & Goodfellow, arXiv, 2016]
[Xu et al., Automatically Evading Classifiers, NDSS 2016]
[Tramèr et al., Stealing Machine Learning Models via Prediction APIs, SECURITY 2016]

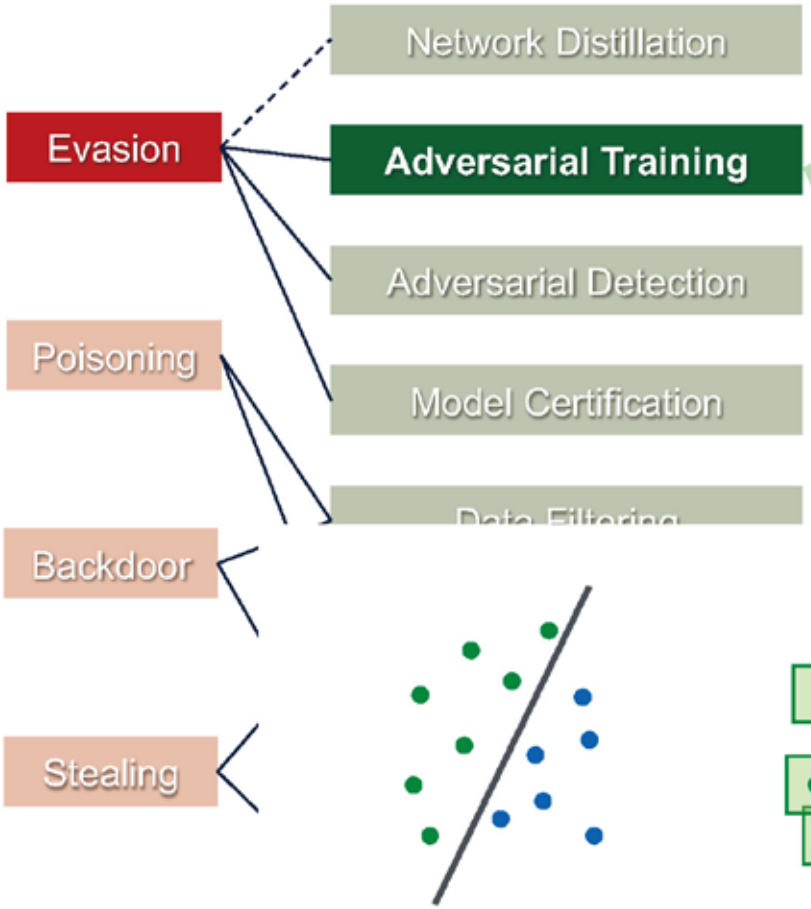
Defense

- Evasion
  - Network Distillation
  - Adversarial Training
  - Adversarial Detection
- Poisoning
- Backdoor
- Stealing
- Model Watermarking

- [Papernot et al., Distillation as a defense to adversarial perturbations against deep neural networks, SP 2016]
- Not effective [Carlini and Wagner. Towards evaluating the robustness of neural networks. SP 2017]



Defense



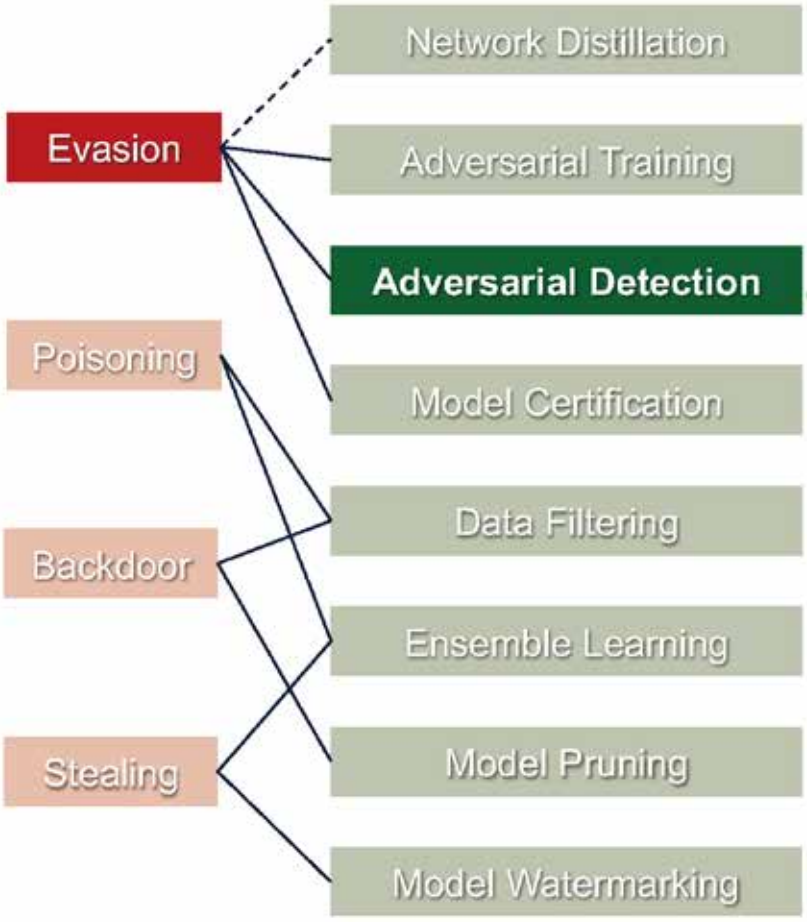
- [Goodfellow, Shlens, & Szegedy, Explaning and Harnessing Adversarial Examples, ICLR 2015]
- Robust optimization approach

$$\min_{\theta} \sum_{i \in \mathcal{D}} \max_{\eta_i : \|\eta_i\| \leq \epsilon} \ell(F(x_i + \eta_i, \theta), y_i)$$

- Can be criticized, since an attacker may not perturb **all** training examples.

[Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018]

Defense

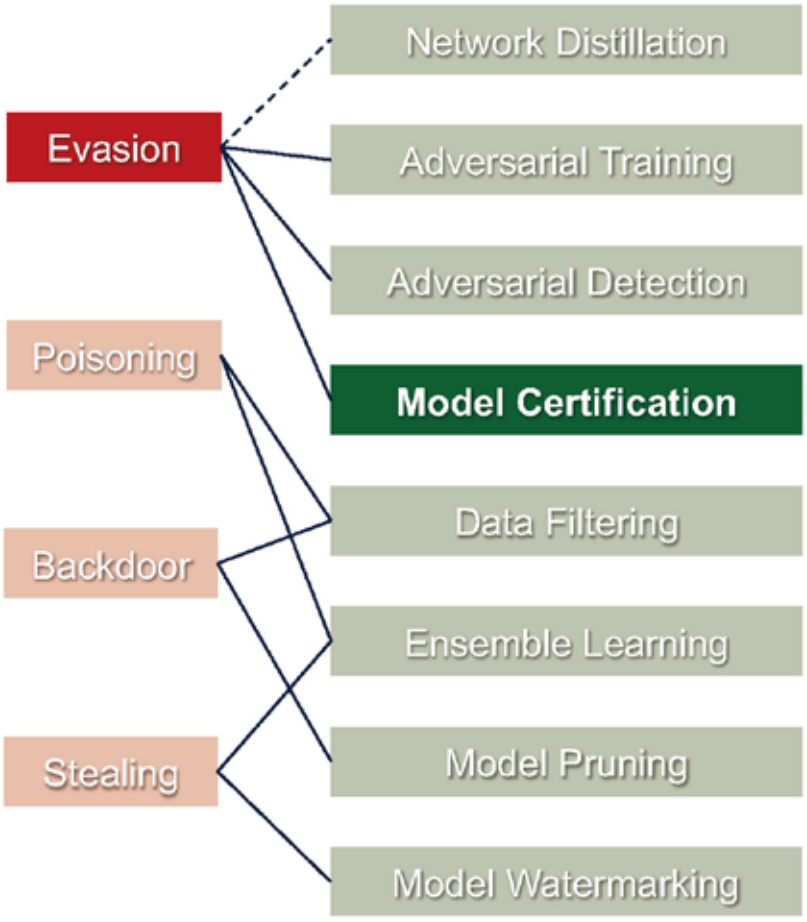


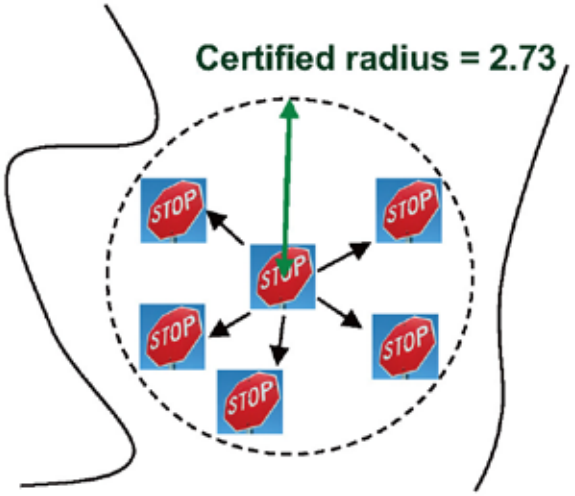
- In essence, anomaly detection

Very active research area:

- [Pang et al., Towards Robust Detection of Adversarial Examples, NIPS 2018]
- [Hu et al., A New Defense Against Adversarial Images: Turning a Weakness into a Strength, NIPS 2019]

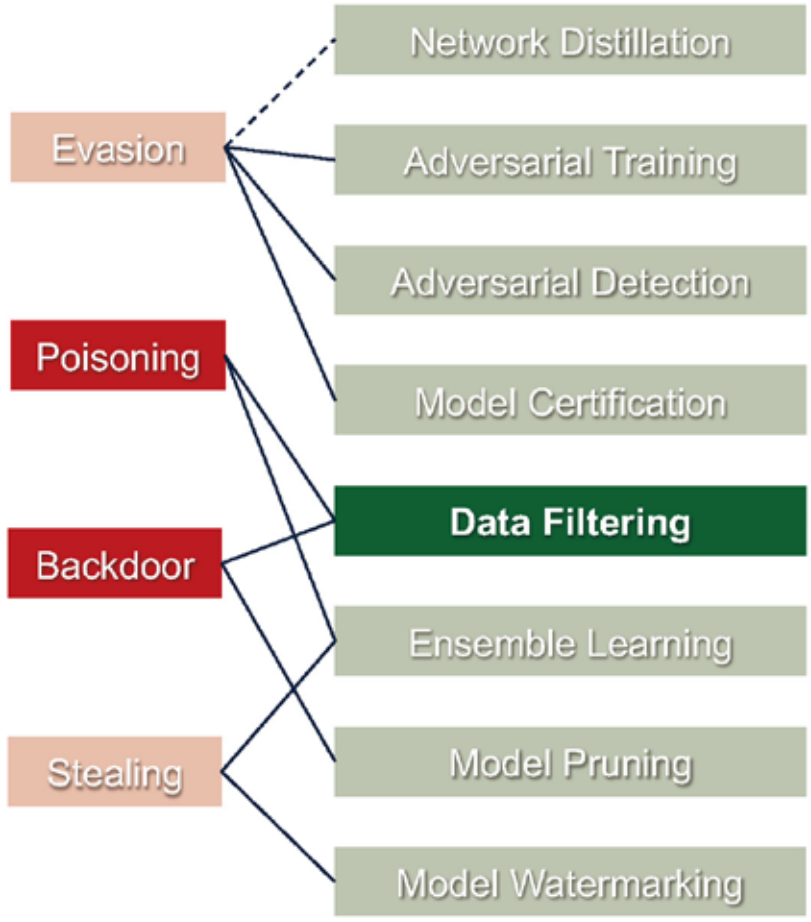
Defense





- [Raghunathan, Steinhardt & Liang, Certified Defenses against Adversarial Examples, ICLR 2018]
- [Ghiasi, Shafahi & Goldstein, Breaking Certified Defenses: Semantic Adversarial Examples with Spoofed Robustness Certificates, ICLR 2020]
  - Certified ≠ Robust

Defense



- [Laishram & Phoha, Curie: A method for protecting SVM classifier from poisoning attack, arXiv:1606.01584, 2016]
- [Liu, Yang & Ankur, Neural trojans, IEEE ICCD , 2017]
  - Using AE-based input reconstruction as pre-processing, they can avoid 90 % of trojan triggers

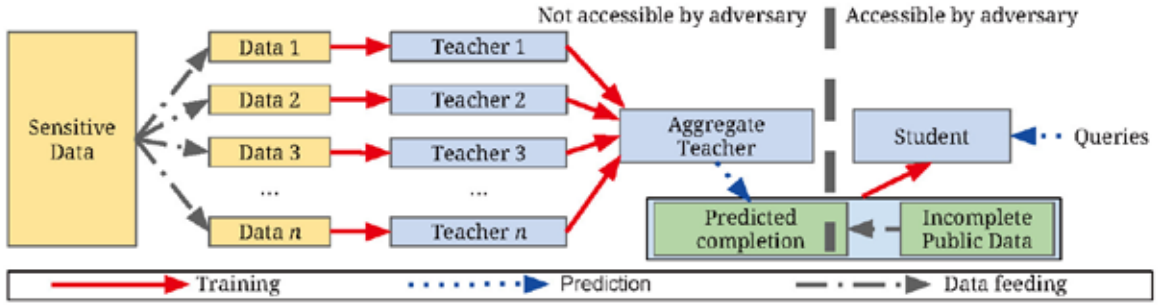
Defense

Evasion

Poisoning

Backdoor

Stealing



Data Filtering

Ensemble Learning

- [Papernot et al., Semi-supervised knowledge transfer for deep learning from private training data, arXiv:1610.05755, 2016]

Model Pruning

- [Liu, Brendan & Siddharth, Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks, arXiv:1805.12185, 2018]

Model Watermarking

- Implant neurons to change prediction on special inputs

Challenges

AI model verification

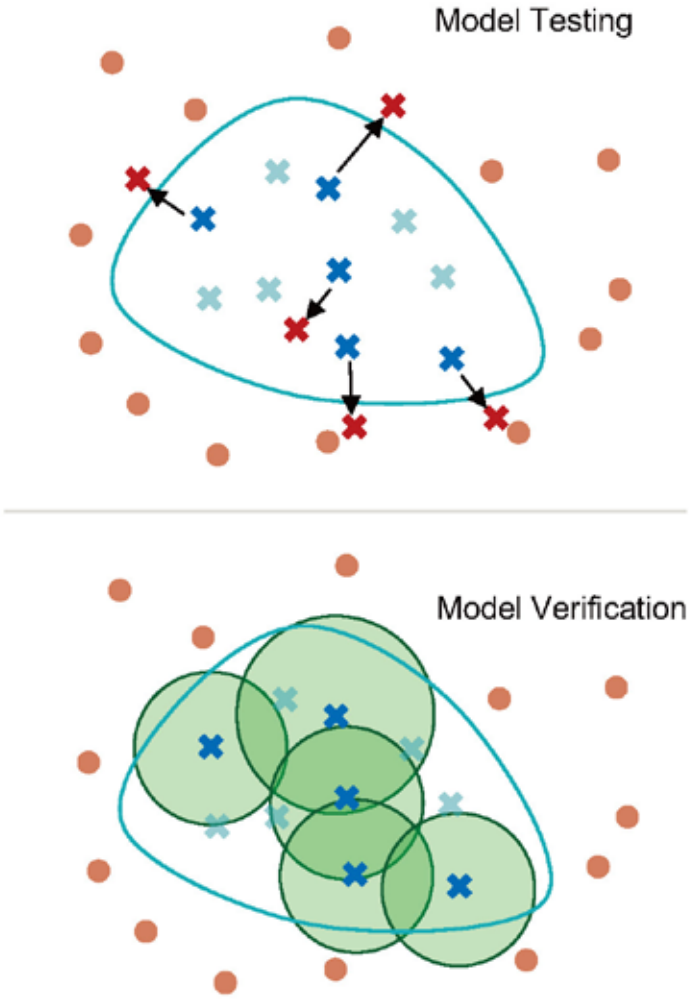
- Certified verification of AI models
- Essential for mission-critical AIs
  - E.g., autonomous driving, IDS, etc.
- Requires better understanding of AI models

AI system verification

- Integrated AI systems consist of perception, l earning, decision, and actions
- Needs to verify AI components independently and in concert

Model explanation

- Critical in automated decision making
- Required to analyze fairness, logic vulnerabili ties, blind spots of data, etc.



Other Issues of Secure AI

Privacy / Data Security

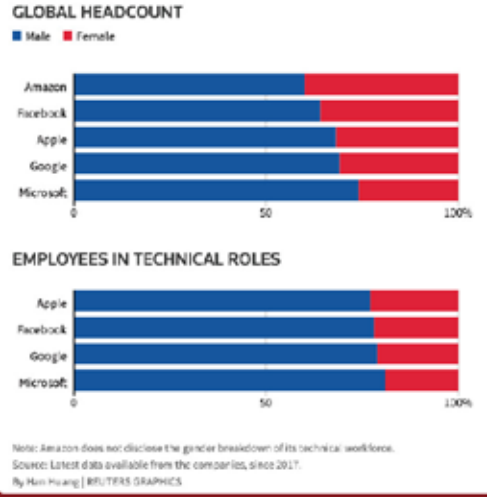
- One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority, New York Times, 2019
- Berkeley Bans Government Face Recognition Use, Joining Other Cities, Bloomberg Law, 2019
- California Just Blocked Police Body Cam Use of Face Recognition, ACLU, 2019

NY Times, 2019


American Civil Liberties Union, 2019

Fairness in AI

- Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 2018
- This is how AI bias really happens - and why it's so hard to fix, MIT Technology Review, 2019
- What Does Fairness in AI Mean?, Forbes, 2020



Rethinking Privacy for the AI Era

FUTURE of AI Act (US Congress, 2018)

- Fundamentally Understanding The Usability and Realistic Evolution of Artificial Intelligence
- AI의 개인정보 남용을 막기 위한 기본 틀

The General Data Protection Regulation (GDPR, EU, 2018.5.25)

- 설명 가능성 (Explainability)
  - 법적인 또는 유사한 효과를 갖는 자동화된 결정에 대해 요구됨
    - 예: 고용, 신용평가, 보험
  - 대상이 되는 사람에게 자동화된 결정에 대한 설명을 들을 권리를 보장
- 위험 평가 (Risk Assessment)
  - 선제적인 개인정보 침해 위험도 평가의 필요성
  - 특히, 최신 기술을 사용하는 경우나 개인정보 침해 가능성이 큰 경우

California Consumer Privacy Act (Jan. 1, 2020)

- 개인 정보 삭제의 권리, 수집에서 탈퇴할 권리, 수집된 본인의 개인 정보에 접근할 권리

GDPR Considerations (1)

Purpose

- Scientific research vs. application
  - Static (offline) vs dynamic (online) models
  - 데이터 사용 목적이 불분명하거나, 시간이 흐름에 따라 변하는 경우도 있음
- Data Minimization:
  - "More data is better. So give me all your data"
  - → 필요 이상의 개인정보 데이터 사용을 제한해야 함
- Limit re-purposing of data

Transparency

- 데이터 수집 대상자에게는 개인정보가 어떻게 사용될 것인지를 알려야 함
- Challenges
  - 발전된 기술은 설명하거나 이해하기 어려운 경우가 많음
  - AI ≈ black box: 정보가 AI 내에서 어떻게 사용되는지 설명 불가할 수 있음
  - AI 모델에 대한 정보는 자칫 핵심 기술 유출로 이어질 수 있음

GDPR Considerations (2)

Individual automated decisions

- 기계화된 프로세스에 기반한 개인에 대한 의사 결정
  - 예를 들어, 속도 위반 카메라에 기반한 자동화된 벌금 부여

- GDPR에서는 기본적으로 허용하지 않음

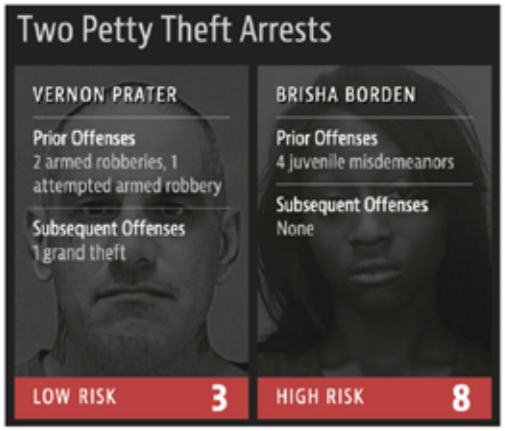
- 예외 조건:
  - 자동화된 의사 결정이 계약을 체결하는데 있어 필요 조건이거나, 법에 의해 허용되거나, 또는 데이터 수집 대상자의 명시적 동의가 있는 경우 허용됨
  - 데이터 수집 대상자가 동의하거나 법에 의해 허용된 경우 민감한 개인정보를 다루는 자동화된 의사 결정이 허용됨

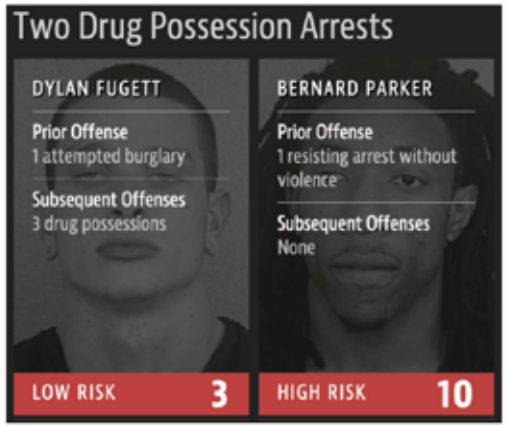
- 주의점: 다른 형태의 개인 정보의 결합이 민감한 정보를 유출할 수 있음
  - 간단한 survey와 Facebook의 "likes"를 조합한 정보로 사용자의 성적 경향을 88%, 인종을 95%의 정확도로 예측 가능함 [Kosinski, Stillwell, and Graepel, PNAS 2013]

Fairness in Machine

Risk scores of future crime

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software
  - 피고인이 다른 범죄를 저지를 위험도 평가
  - 미국 Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin주의 법정에서 사용됨
- 미국 법무부 국립교정연구소 (National Institute of Corrections)는 재판의 모든 단계에서 위험도 평가 점수를 고려하도록 권유하고 있음





Machine Bias, ProPublica, 2016

Top10 Strategic Tech Trends for 2020

Gartner

AI Security

- The increase of AI solutions and potential points of attacks (e.g., IoT devices & highly connected services) creates a true security challenge

- Protecting AI-powered systems
  - Securing AI training data, training pipelines and ML models
- Leveraging AI to enhance security defense
  - Use of ML to understand patterns, uncover attacks and automate parts of cybersecurity processes
- Anticipating nefarious use of AI by attackers
  - Identifying attacks and defending against them

Through 2022, 30% of all AI cyberattacks will leverage training-data poisoning, AI model theft or adversarial samples to attack AI-powered systems

Conclusion

AI is a new, efficient, general-purpose technology

- Computing power, data volume, advanced algorithms enable more powerful AI
- AI is a double-edged technology

Secure AI is a key for the safe use of AI

- More understanding and advances are required toward secure AI
- Proper monitoring and integration is required
  - Human-in-the-loop

Collaborated effort is a key for successful implementation of AI

- Secure data sharing: homomorphic encryption, federated learning

감사합니다.