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

Session 5-1

AI를 속이는 보안 공격과 대응 방안 연구

최대선 교수 (숭실대학교)

[요약문]

AI 기술이 다양한 분야에 적용되어 활용되고 있다. 정보보호 분야에서도 AI 기술을 악성코드 탐지, 침입탐지, 이상거래 탐지 등에 활용하고 있다.

그런데, AI를 대상으로 하는 여러 가지 보안 공격이 존재하며, 데이터에 간단한 변경을 가해서 AI를 속이고 오분류를 유도하는 기만공격이 심각한 문제로 부각되고 있다.

본 발표에서는 AI를 속이는 보안 공격의 다양한 형태와 실제 연구 결과를 소개한다. 얼굴인식기를 속이는 공격, 음성인식을 속이는 공격 등의 실제 원리와 효과를 살펴본다. 또한, 이에 대한 기술적, 절차적 보안 대책의 현황을 살펴본다. 실제 상기 공격에 대응하는 연구 내용과 결과를 소개한다.

[발표자 약력]

2009년 KAIST 전산학과 박사
1999년~2015년 ETRI 정보보호연구본부 인증기술연구실장
2015년~2020년8월 공주대학교 의료정보학과 교수
2020년9월~현재 숭실대학교 소프트웨어학부 교수
2019년~현재 한국정보보호학회 차세대인증연구회장

관심분야 : 정보보호, 사용자 인증, 인공지능 보안, 프라이버시 등

숭실대학교 Soongsil University

AI를 속이는 보안 공격과 대응 방안 연구

최대선
소프트웨어학부
숭실대학교

숭실대학교 Soongsil University

내용

- AI 보안 공격
- Evasion attack
- AI 보안 공격 방어 기술
- Evasion attack 방어 연구 소개

2

숭실대학교 Soongsil University

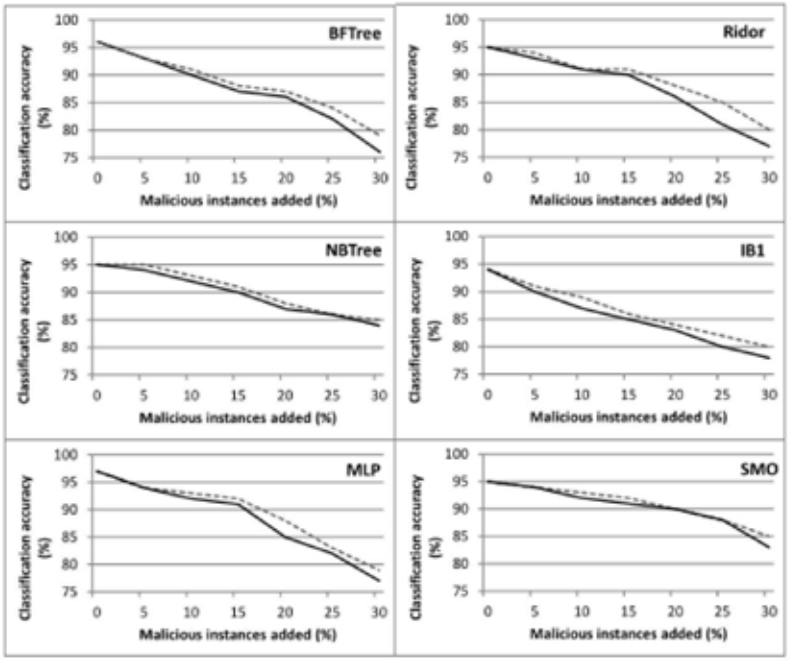
AI 보안 공격

숭실대학교 Soongsil University

POISONING ATTACK

▶ 목표

- 학습데이터에 poison data를 최소한 추가하여 오류 최대화



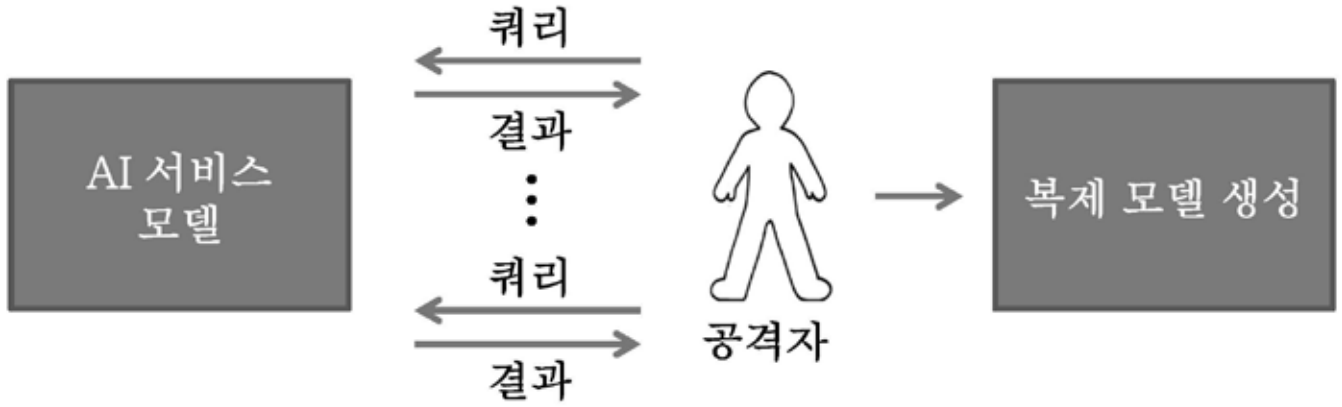
[출처 :M. Mozaffari-Kermani,et.al, Systematic Poisoning Attacks..IEEE Journal of Bio & Health Informatics]

4

MODEL EXTRACTION ATTACK

숭실대학교 Soongsil University

▶ 학습된 모델에 쿼리를 해서 타겟 모델 f 에 가까운 f' 만들기 (복제)



▶ 공격 목적

- 유료서비스 모델 탈취

☞ 다른 공격에 활용

- Inversion attack, Poisoning attack, Evasion attack

5

MODEL EXTRACTION ATTACK

숭실대학교 Soongsil University

▶ Performance

- 100% 흉내 내는데 소요된 쿼리# 및 시간

Service	Model Type	Data set	Queries	Time (s)
Amazon	Logistic Regression	Digits	650	70
	Logistic Regression	Adult	1,485	149
BigML	Decision Tree	German Credit	1,150	631
	Decision Tree	Steak Survey	4,013	2,088

- Decision Tree 재구성 (using incomplete queries)

Model	Leaves	Unique IDs	Depth	Without incomplete queries			With incomplete queries		
				$1-R_{\text{test}}$	$1-R_{\text{unif}}$	Queries	$1-R_{\text{test}}$	$1-R_{\text{unif}}$	Queries
IRS Tax Patterns	318	318	8	100.00%	100.00%	101,057	100.00%	100.00%	29,609
Steak Survey	193	28	17	92.45%	86.40%	3,652	100.00%	100.00%	4,013
GSS Survey	159	113	8	99.98%	99.61%	7,434	100.00%	99.65%	2,752
Email Importance	109	55	17	99.13%	99.90%	12,888	99.81%	99.99%	4,081
Email Spam	219	78	29	87.20%	100.00%	42,324	99.70%	100.00%	21,808
German Credit	26	25	11	100.00%	100.00%	1,722	100.00%	100.00%	1,150
Medical Cover	49	49	11	100.00%	100.00%	5,966	100.00%	100.00%	1,788
Bitcoin Price	155	155	9	100.00%	100.00%	31,956	100.00%	100.00%	7,390

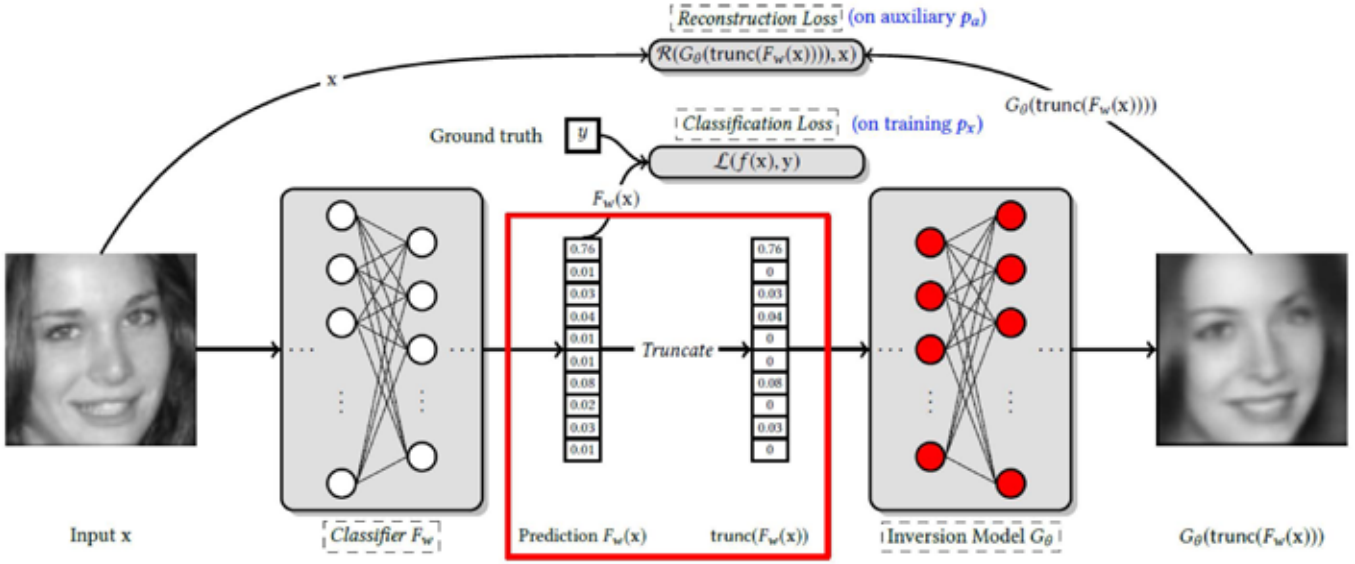
<출처 : Florian Tramer,et.al, Stealing Machine Learning Models via Prediction APIs , Usenix'16>

6

MODEL INVERSION

▶ Model Inversion 공격

- 타겟 모델의 학습 데이터를 추출하는 공격
- Truncation
  - 예측 값이 높은 상위 n개 클래스를 제외한 나머지 클래스의 예측 값을 0
  - 공격에 강건한 모델을 만들어 줌



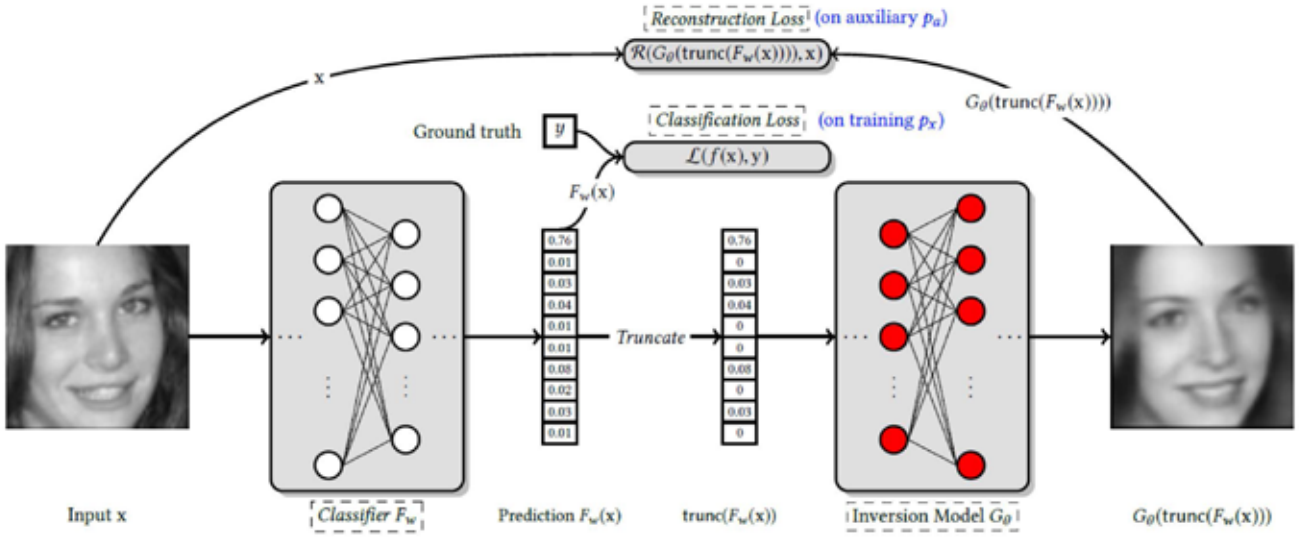
Truncation 적용

<출처 : Z. Yang et al. "Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment." ACM CCS 2019.>

MODEL INVERSION

▶ Model Inversion 공격

- 입력 데이터를 타겟 모델의 학습 데이터와 비슷하게 복원 시켜줌
- Inversion 모델 학습
  - 입력 값: truncated prediction
  - 재복원 loss: 입력 데이터와 재복원 데이터의 차이



8 <출처 : Z. Yang et al. "Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment." ACM CCS 2019.>

숭실대학교 Soongsil University

MEMBERSHIP INFERENCE

▶ Membership Inference

- 데이터가 타겟 모델의 학습 데이터인지 아닌지 추측

  ➔ 개인의 민감한 정보 노출

  – Evasion attack를 방어하기 위한 Adversarial Training 방법은 Membership Inference 공격에 취약

Benign sample 추론

Adv-exam 추론

Training method	train acc	test acc	adv-train acc	adv-test acc	inference acc ($\mathcal{I}_B$)	inference acc ($\mathcal{I}_A$)
Natural	100%	98.25%	4.53%	2.92%	**55.85%**	54.27%
PGD-Based Adv-Train [33]	99.89%	96.69%	99.00%	77.63%	61.69%	**68.83%**
Dist-Based Adv-Train [50]	99.58%	93.77%	83.26%	55.06%	62.23%	**64.07%**
Diff-Based Adv-Train [66]	99.53%	93.77%	99.42%	83.85%	58.06%	**65.59%**

9

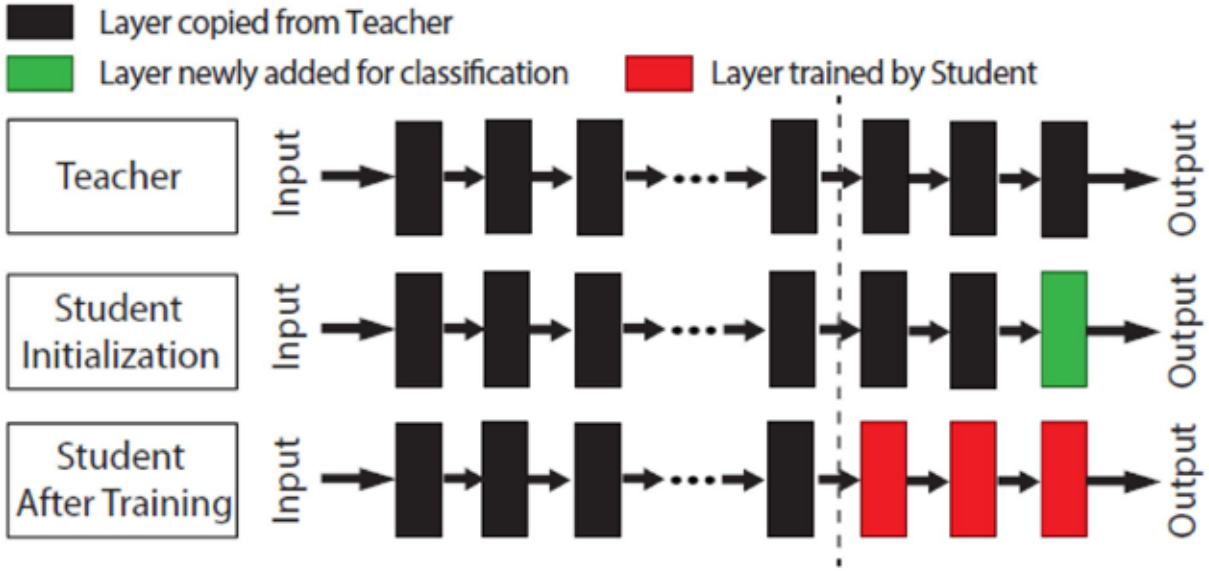
<출처 : L. Song et al. "Privacy Risks of Securing Machine Learning Models against Adversarial Examples." ACM CCS 2019.>

숭실대학교 Soongsil University

BACKDOOR ATTACKS

▶ Transfer Learning에서 backdoor 공격

- Transfer learning
  - 잘 학습되어 있는 teacher 모델을 가져와 사용자 데이터 셋에 맞게 재학습
  - 데이터 양이 많으면 계산량 증가 ➔ transfer learning 사용



Transfer Learning

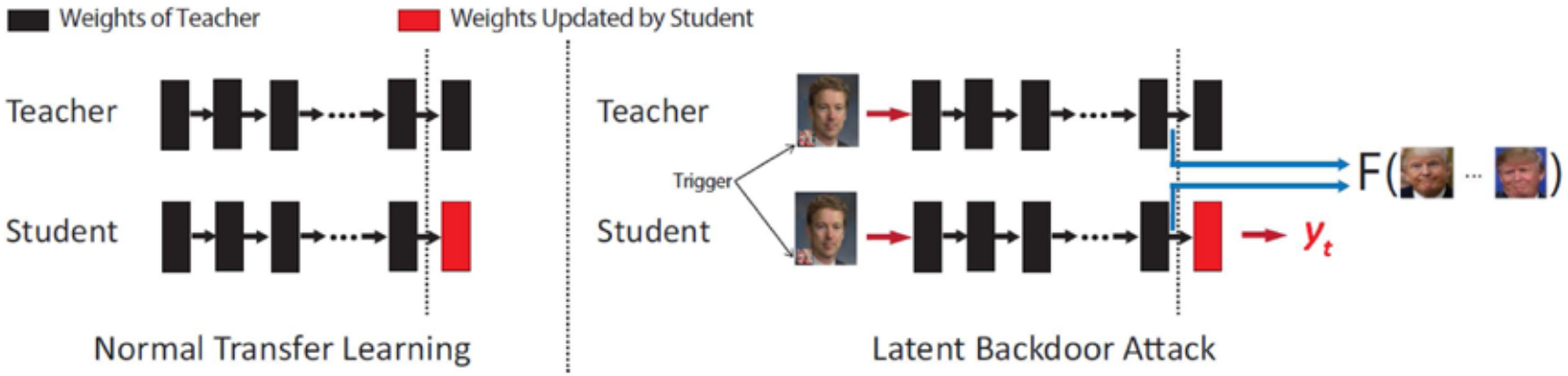
10 <출처 : Y. Yao et al. "Latent Backdoor Attacks on Deep Neural Networks." ACM CCS 2019.>

숭실대학교 Soongsil University

BACKDOOR ATTACKS

▶ Transfer Learning에서 backdoor 공격

- 모델 학습 시, 데이터에 trigger 삽입
  - 인물 분류 시, trigger가 있는 이미지는 특정 인물로 분류하도록 학습
    (ex. trigger 이미지 ➔ 트럼프)
- Trigger 이미지로 학습된 모델을 무료 배포
  - 누군가 모델을 가져다가 transfer learning



11 <출처 : Y. Yao et al. "Latent Backdoor Attacks on Deep Neural Networks." ACM CCS 2019.>

숭실대학교 Soongsil University

EVASION ATTACK

숭실대학교 Soongsil University

EVASION ATTACK

▶ 이미지 변조

- 사람이 보기엔 별 문제 없는 [최소 변조] 이미지
- 4%만 변조해도.. 97%는 잘못 분류 [최대 오류]

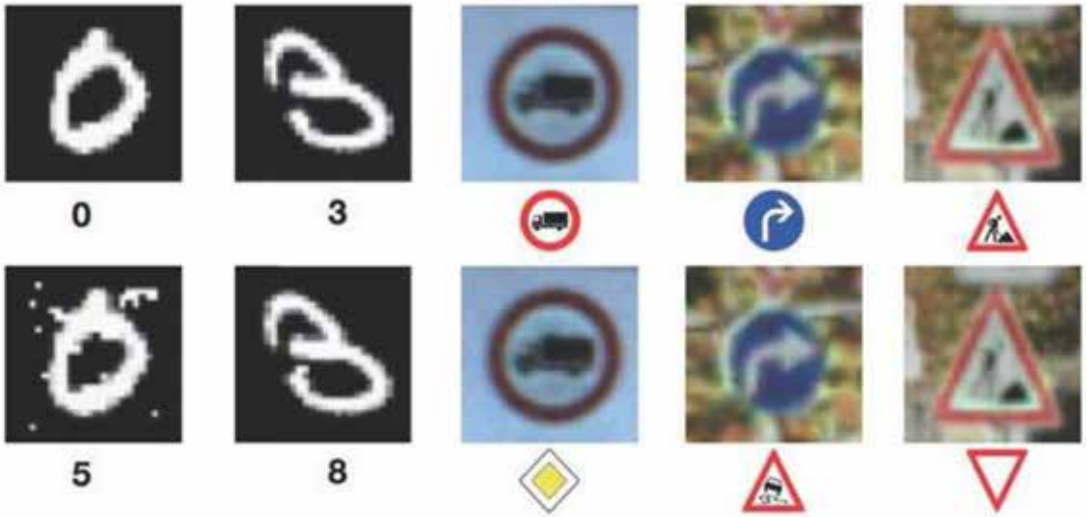
<출처 : 구글-뉴욕대, 2014>

13

EVASION ATTACK

숭실대학교 Soongsil University

- 심각하지 않다고요?
  - ☞ 좌회전 표시를 보고 우회전하는 자율 주행차?



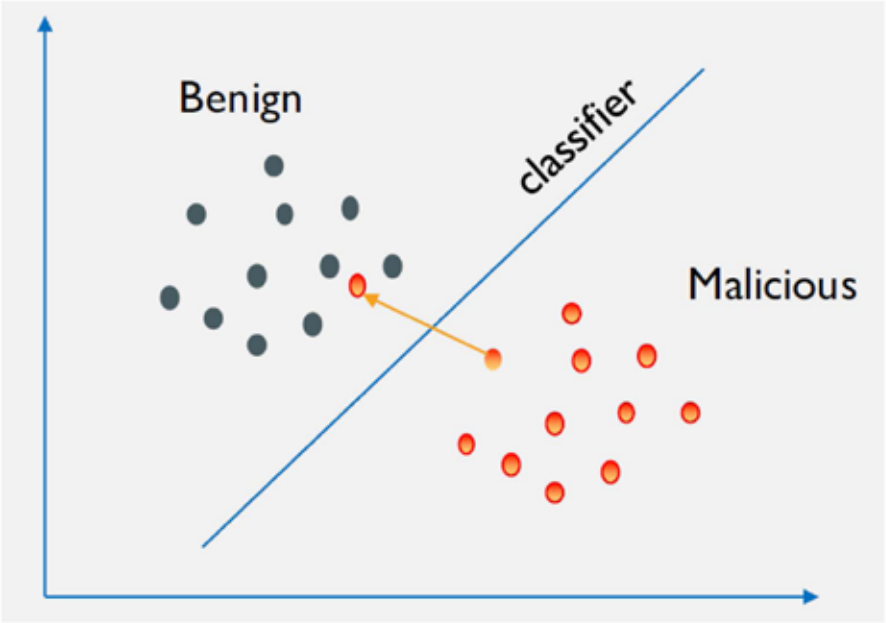
<출처 : PSU, google, WSU, 2016>

14

숭실대학교 Soongsil University

활용단계-EVASION ATTACK

▶ 최소한의 변조 (feature perturbation)로 다른 class로 인식되는 적대적 예제(adversarial example)를 찾는 것



<출처 : Vorobeychik, Florian Tramer,et.al, Adeversarial AI>

min $d(x, x')$ s.t.: $d(x, x') \leq cost\ budget$, x' classified as benign

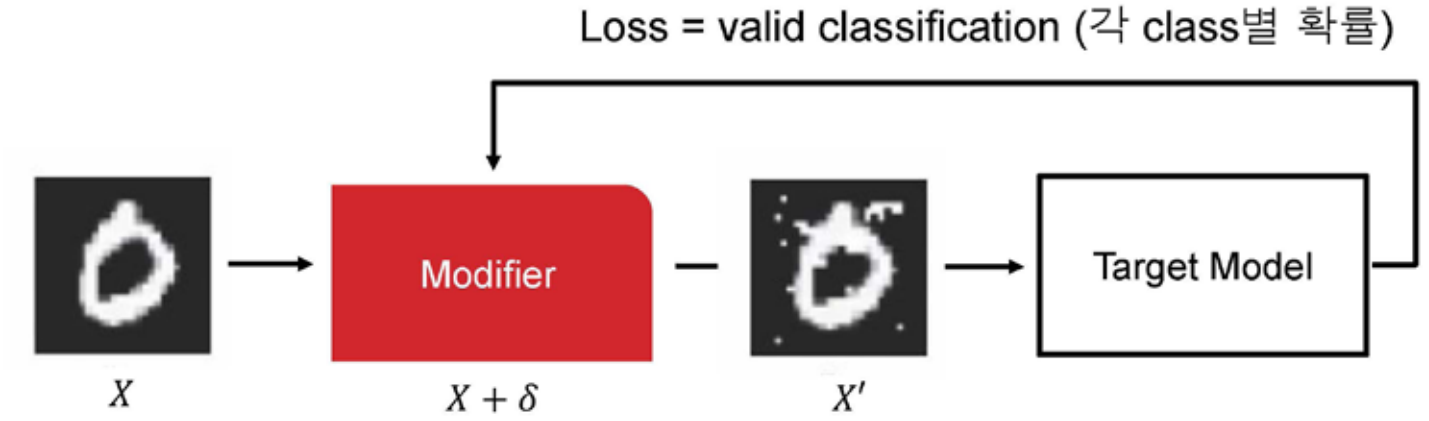
15

숭실대학교 Soongsil University

ADVERSARIAL ATTACK 원리

▶ Target classifier를 속일 수 있도록 변조

- 머신러닝과 동일한 원리



▶ 변조율 최소를 목표 : 사람에 의한 탐지를 방지

- $\min \delta$ & $\min L$

16

숭실대학교 Soongsil University

EVASION ATTACK

▶ Procedural Noise Adversarial Examples

- 노이즈 패턴 이미지를 생성하여 타겟 블랙박스 모델 공격
  - Black-box 공격 ➔ Bayesian optimization 사용해서 타겟 모델 공격
  - Target model: Inception V3
  - ImageNet의 Validation dataset 에서 랜덤하게 5,000장 사용하여 학습
  - 타겟 모델의 출력 값에서 top label만 요구 (대체 모델 X)

tabby	0.706	shower_curtain	0.236
tiger_cat	0.221	tabby	0.157
Egyptian_cat	0.046	quilt	0.140
window_screen	0.002	tiger_cat	0.122
Persian_cat	0.001	Egyptian_cat	0.075

17 <출처 : K.T. Co et al. "Procedural Noise Adversarial Examples for Black-Box Attacks on Deep Convolutional Networks." ACM CCS 2019.>

숭실대학교 Soongsil University

FRIEND-SAFE EVASION

- Enemy는 틀리게 인식하고, Friend (아군, 동맹국)은 올바르게 인식하게 만드는 adv. exam. 생성

우회전으로 인식

좌회전으로 인식

아군 자율 주행차

적군 자율 주행차

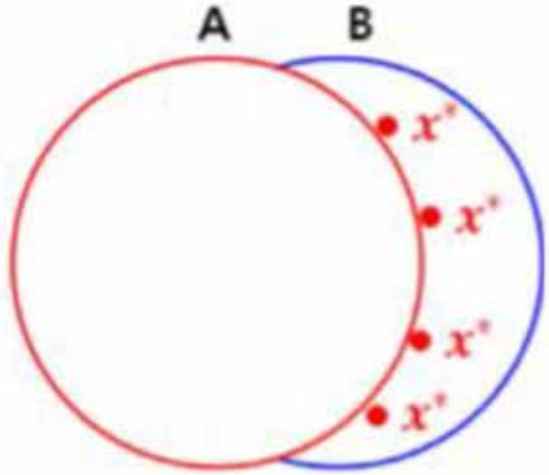
<Hyun Gwon, et al. "Friend-safe evasion attack: An adversarial example that is correctly recognized by a friendly classifier", Computers & Security, 2018.>

18

숭실대학교 Soongsil University

FRIEND-SAFE EVASION

▶ 목표

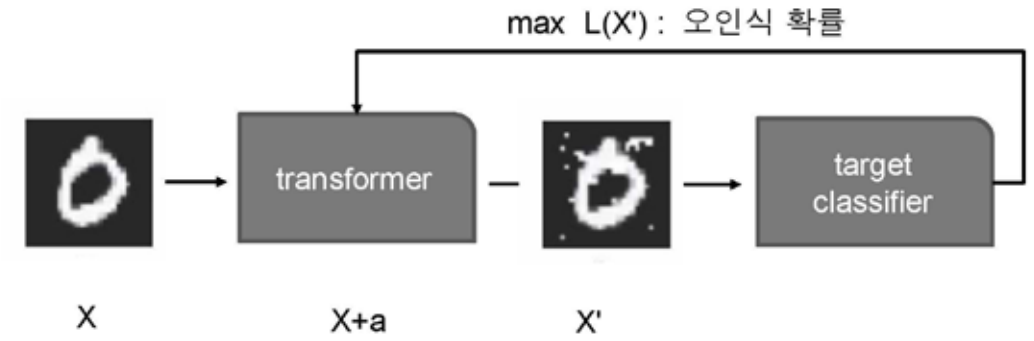


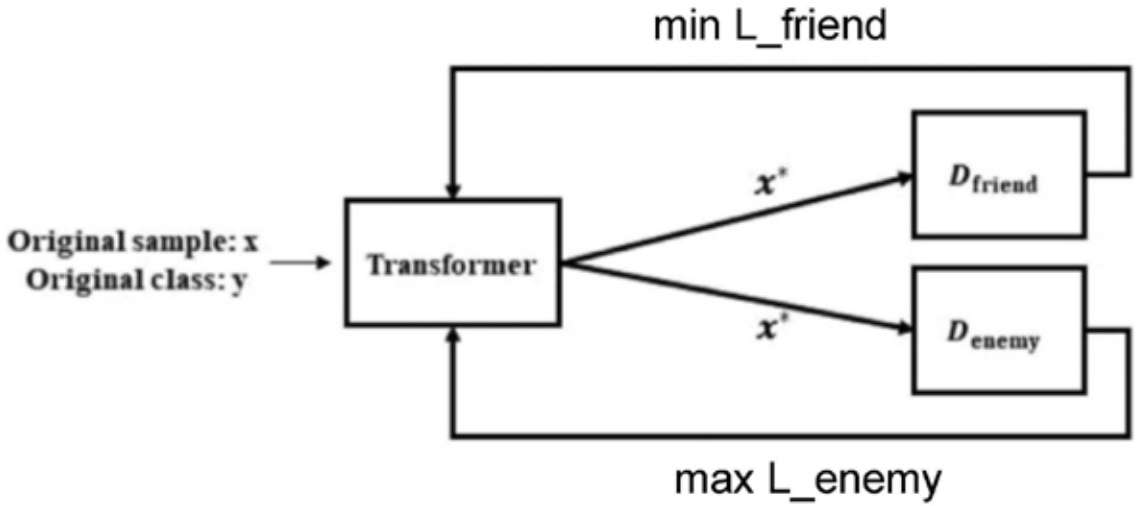
19

숭실대학교 Soongsil University

FRIEND-SAFE EVASION

▶ 방법



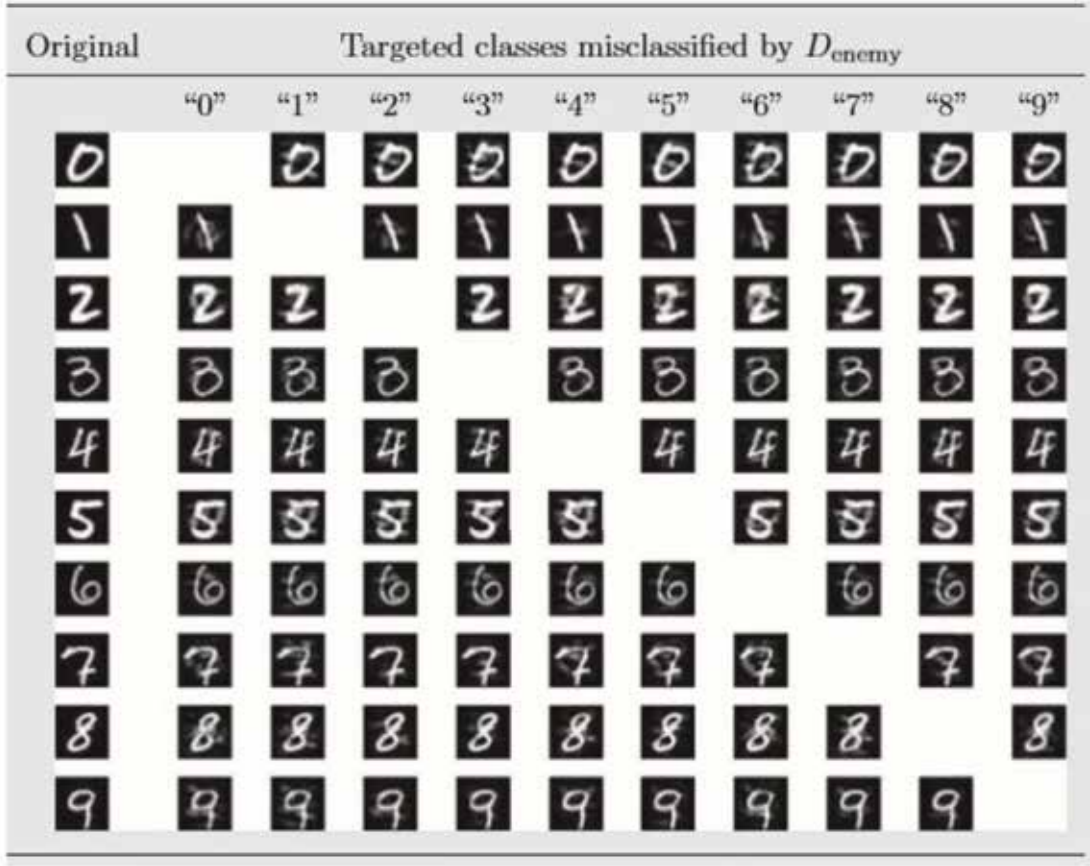


20

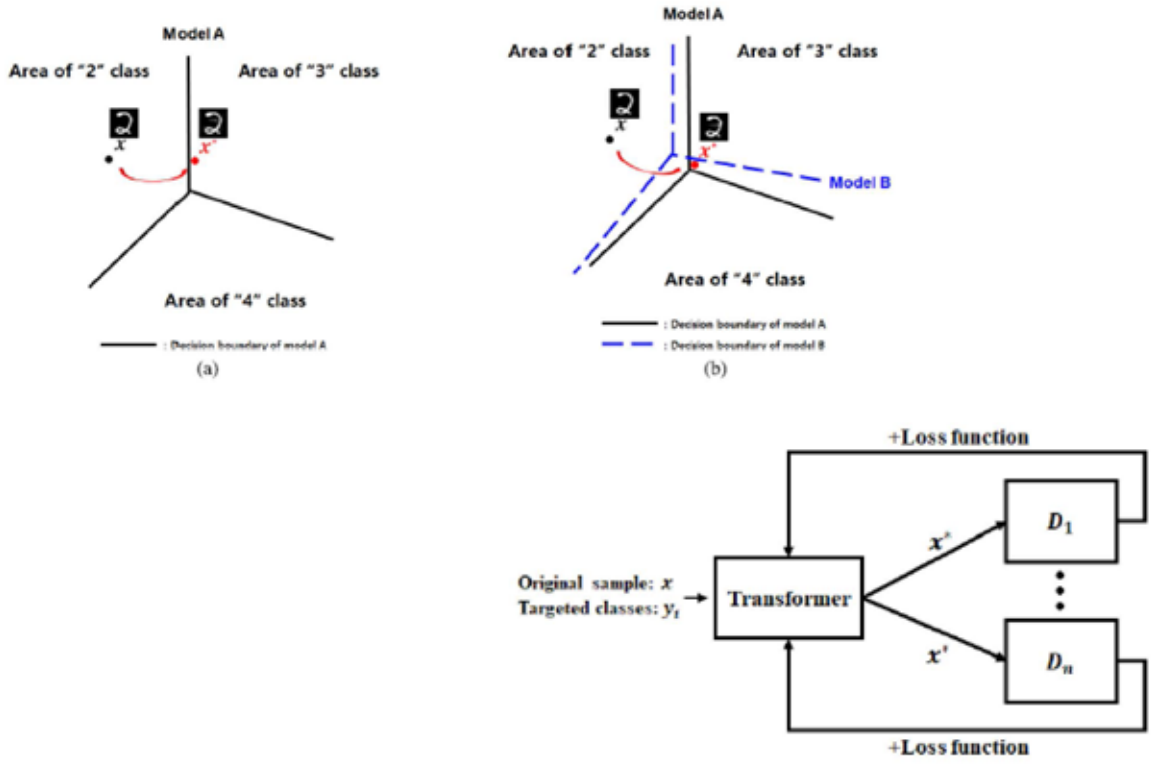
숭실대학교 Soongsil University

FRIEND-SAFE ADV. EXAM.

▶ 결과



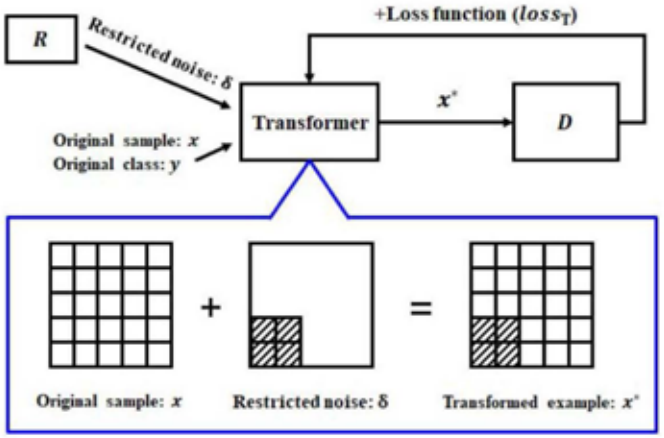
MULTI TARGET



H. Kwon, H. Yoon, D. Choi, Multi-Targeted Adversarial Example in Evasion Attack on Deep Neural Network, IEEE ACCESS'18 [SCI]

RESTRICTED EVASION

▶ 제한된 영역만 변조하여 evasion 공격

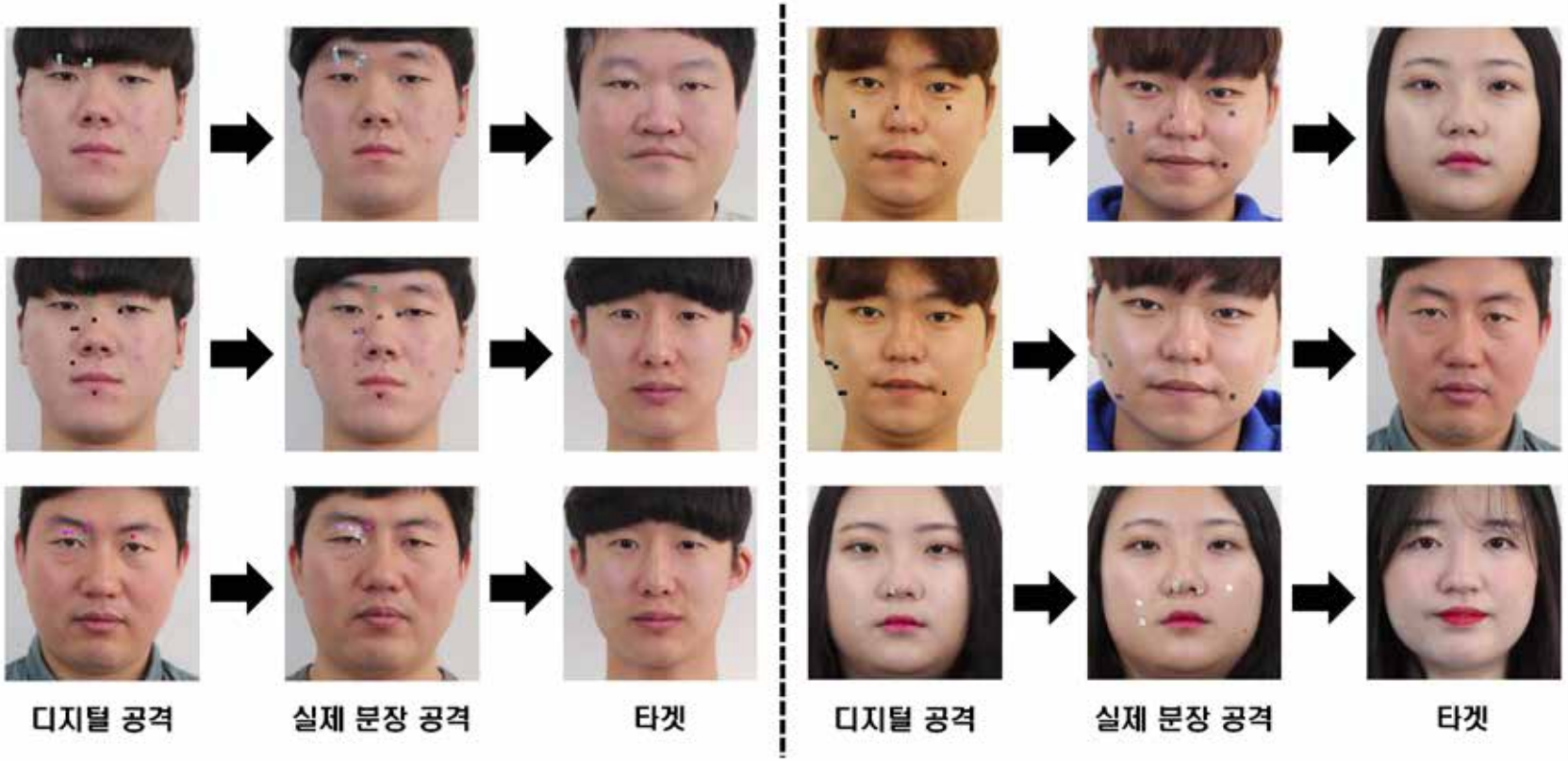


(a) Original sample (b) Restricted adversarial example

Kwon, Hyun, Hyunsoo Yoon, and Daeseon Choi. "Restricted Evasion Attack: Generation of Restricted-Area Adversarial Example." IEEE Access 7 (2019): 60908-60919.

숭실대학교 Soongsil University

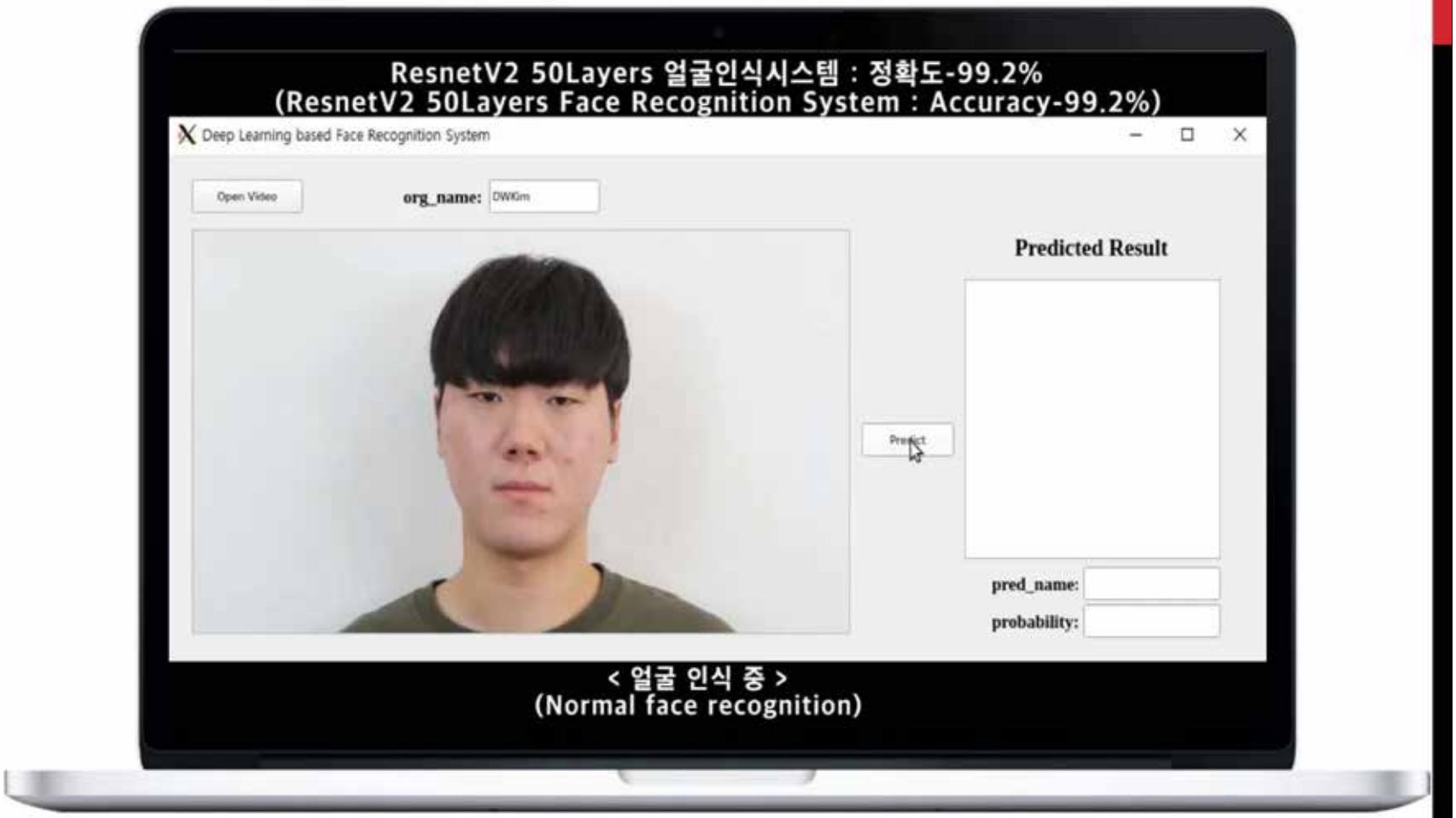
실제 분장 공격 성공 샘플



24

ATTACK ON FACE RECOGNITION

숭실대학교 Soongsil University



25

숭실대학교 Soongsil University

AI 보안 공격 방어기술

숭실대학교 Soongsil University

MEMBERSHIP 방어

▶ MemGuard

- Membership inference 공격을 방어하기 위해 모델의 예측 값에 노이즈를 추가
- 공격자의 추론 모델을 알 수 없으니 방어자 스스로 추론 모델 생성
  - $M^* = \text{argmin}_M \left| E_M\left(g(s+n)\right) - 0.5 \right|$

    ➔ 자신의 추론 모델을 속이도록 모델의 예측 값에 노이즈 추가
  - $\text{Distance}(s, s+n) \leq \epsilon$

    ➔ 예측 값과 노이즈가 추가된 예측 값의 차이가 ϵ 이하여야 함

s: 모델의 예측값
n: 노이즈

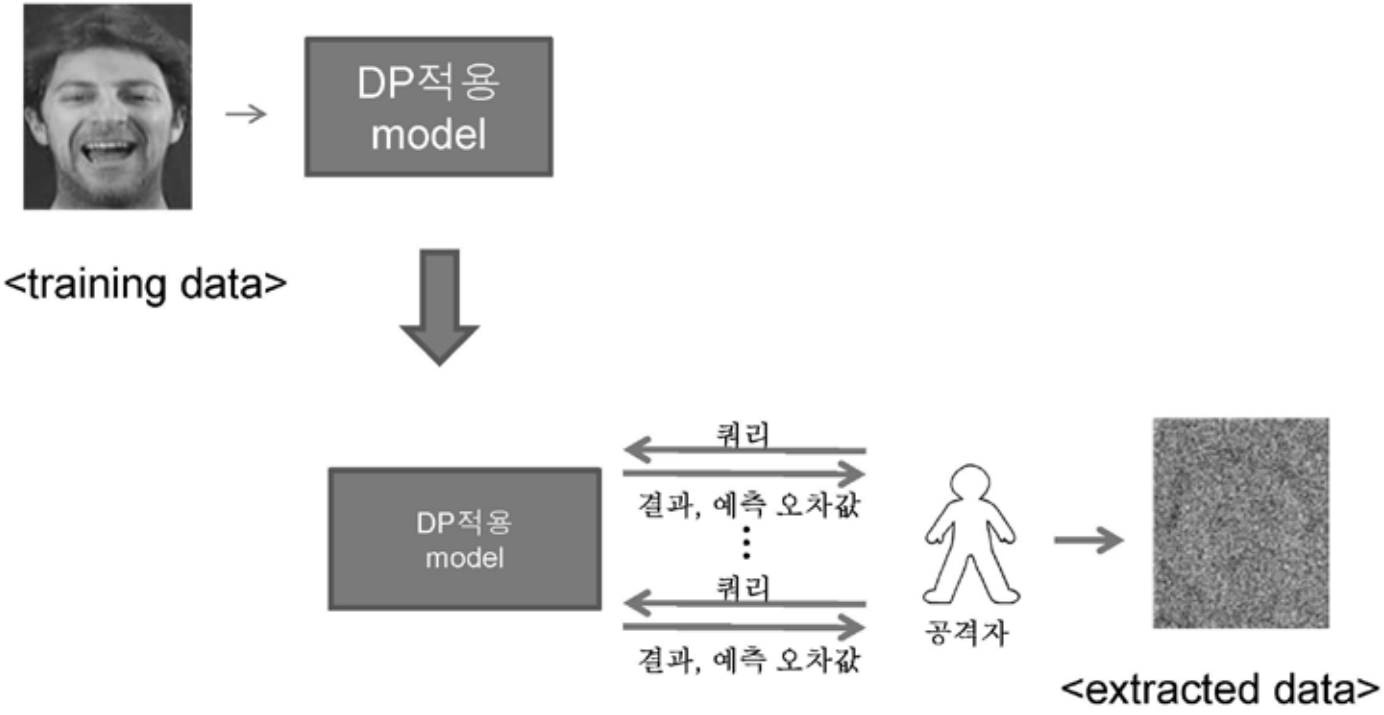
- 공격자가 랜덤하게 추론하도록 하면서 data utility를 보장하는 방법

<출처 : J. Jia et al. "MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples." ACM CCS 2019.>

COUNTER MEASURES FOR INVERSION

숭실대학교 Soongsil University

▶ Differential privacy 적용 모델로 학습



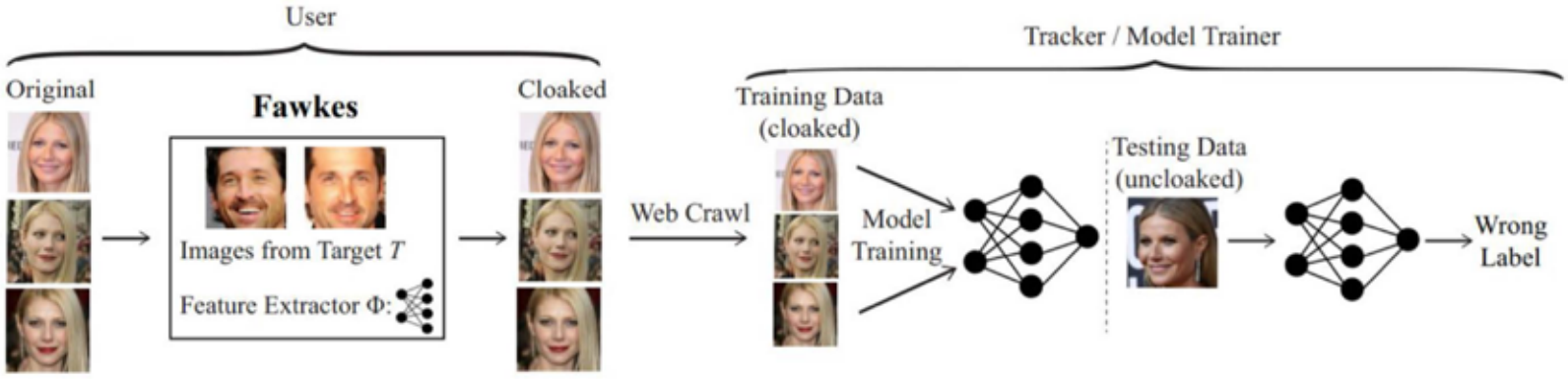
28

숭실대학교 Soongsil University

PROTECTING PRIVACY

▶ Fawkes

- 얼굴 인식 시스템에서 사용자는 자신 얼굴에 대한 프라이버시를 보호하기 위해 사용자 얼굴 이미지에 다른 사람의 특징을 삽입
  - 적대적 예제처럼, 얼굴 이미지에 사람이 알아차릴 수 없는 노이즈를 삽입
  - 노이즈가 삽입된 얼굴 이미지를 온라인에 공개
  - Tracker/Model Trainer는 노이즈가 삽입된 얼굴 이미지를 크롤링하여 모델 학습에 사용 ➔ 모델 테스트 시, 오분류 발생



29 <출처 : S. Shan et al. "Fawkes: Protecting Privacy against Unauthorized Deep Learning Models." USENIX Security 2020.>

숭실대학교 Soongsil University

EVASION ATTACK 방어기술

숭실대학교 Soongsil University

EVASION ATTACK 대응 방안

- Classifier 유출 방지
- Adversarial Training
- Robust Classifier
- Classifier 변화 (moving target)
- 적대적 예제 탐지
- 적대적 예제 탐지 + 필터링

31

숭실대학교 Soongsil University

CLASSIFIER 유출 방지

- Target 이 있어야 공격할 수 있다?
- Attacker 능력 (상황)
  - Whitebox : class probability 를 알 수 있음 <- 대부분 그렇지 못함
  - Blackbox : classification 결과 만 획득 가능
    - 노획, 질의를 해볼 수 있는 경우
    - Substitute attack : classification 결과를 토대로 classifier를 재구성한 뒤 그걸 대상 (Whitebox) 으로 공격해도 공격성공 70%
  - Unknown (zeroday) classifier : 아무 것도 없는 경우
    - transferability : 다른 classifier를 타겟으로 변조한 결과도 공격이 성공한다.

32

숭실대학교 Soongsil University

ADVERSARIAL TRAINING

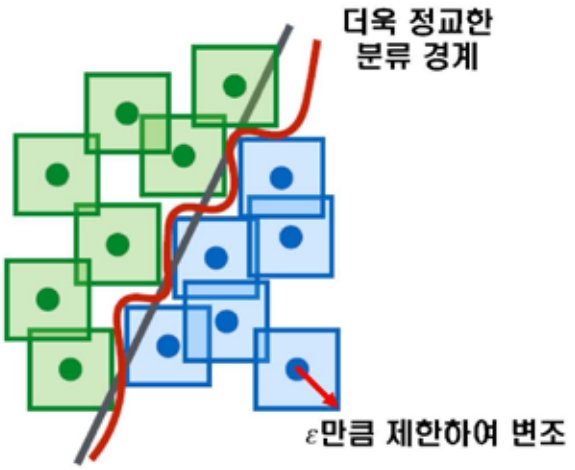
▶ 변조 데이터를 모델 학습에 사용

- 학습 데이터를 ε 만큼 제한하여 변조된 적대적 예제를 학습에 활용
  - $x^{t+1} = \prod_{x+S}\left(x^t + \alpha \mathrm{sgn}\left(\nabla_x L(\theta, x, y)\right)\right)$ ➔ 적대적 예제 생성
  - $\min_{\theta} \rho(\theta)$, where $\rho(\theta) = E_{(x,y)\sim D}\left[\max_{\delta \in S} L(\theta, x+\delta, y)\right]$

    ➔ 적대적 예제도 정상 클래스로 학습

- 모델은 더욱 정교한 분류 경계를 갖음
  - 적대적 예제에 내성을 갖는 모델이 생성됨

☞ 공격 : 그 classifier에 맞춰서 변조



<출처 : A. Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ICLR 2018.>

33

숭실대학교 Soongsil University

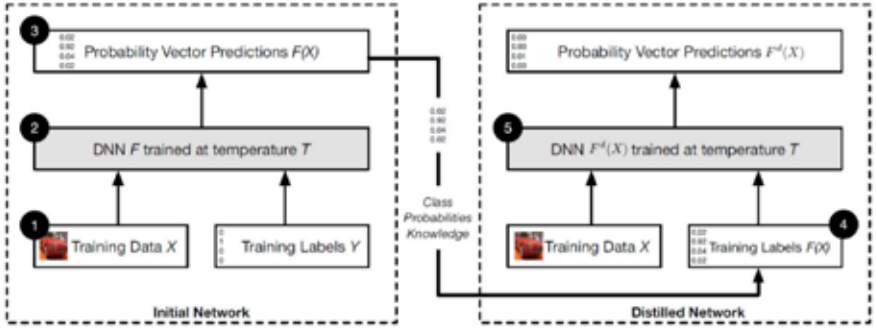
ROBUST CLASSIFIER

▶ Evasion attack에 robust classifier

- Distilled classifier (IEEE S&P 2016)
  - Initial network의 확률 벡터를 활용하여 Distilled Network 학습
  - 공격자에게 모델의 정확한 gradient 정보를 주지 않기 위함
- Softmax function using temperature T

$$F(X) = \left[\frac{e^{\frac{z_i(X)}{T}}}{\sum_{j=0}^{N-1} e^{\frac{z_j(X)}{T}}} \right]_{i \in \{0,\ldots,N-1\}}$$

- z: logit vector
- N: class 개수



☞ CW 공격에 의해 100% 깨짐

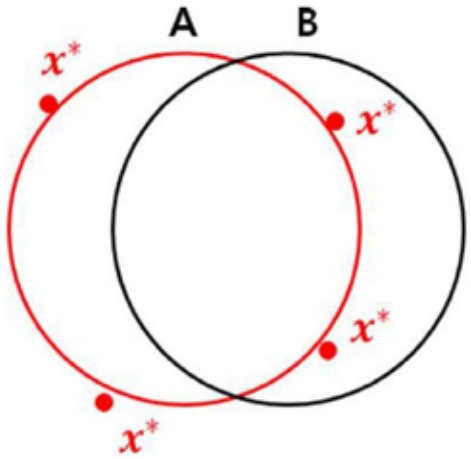
<출처 : N. Papernot et al. "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks." IEEE S&P 2016.>

34

숭실대학교 Soongsil University

CLASSIFIER 변화

- Classifier 를 변화 (moving target)
  - 변조 최소화 (사람에게 들키지 않기 위해)가 목적이므로
  - 공격은 class 경계선 만 살짝 넘도록 한다



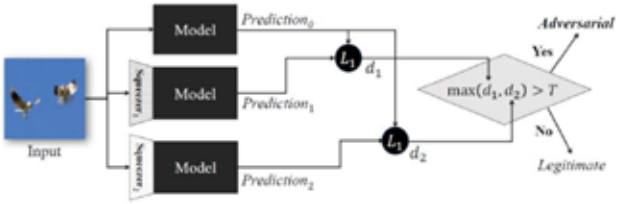
  - 따라서 target 모델에 변경 A→B 을 가하면 , 경계선이 변하므로 공격 성공률이 떨어짐
  - ☞ 공격자가 미리 다양한 classifier를 대상으로 변조 (일반화)

35

적대적 예제 탐지

▶ Feature Squeezing

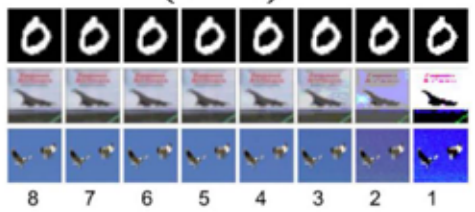
- 입력 이미지를 변형(squeezing)시켜 적대적 예제 탐지
  - 원본 이미지와 변조 이미지의 차이로 탐지



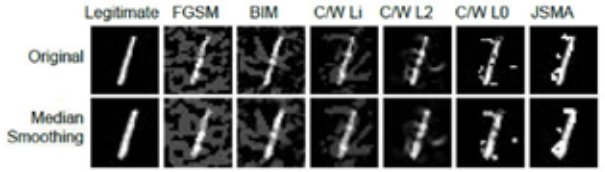
- Reducing color depth
  - 공격자가 공격 시도할 수 있는 공간을 줄임

$$0 \le x_i \le (2^8 - 1) \xrightarrow[\div (2^8 - 1)]{} 0 \le x_i \le 1 \xrightarrow[\times (2^6 - 1)]{} 0 \le x_i \le (2^6 - 1) \xrightarrow[\div (2^6 - 1)]{} 0 \le x_i \le 1$$

- Spatial smoothing
  - 이미지를 흐릿(blur)하게 만듦
  - 2×2 or 3×3 사각 패치의 모든 픽셀을 사각 패치의 중간 값으로 대체



Color bit 줄임에 따른 이미지 예



Smoothing된 이미지 예

<출처 : W. Xu et al. "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks." NDSS 2018.>

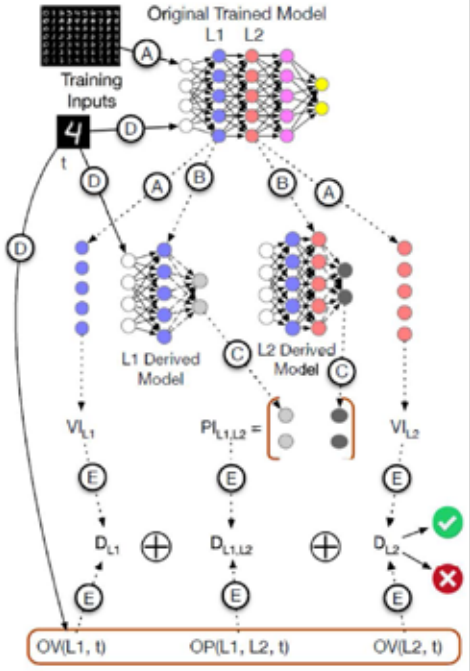
36

숭실대학교 Soongsil University

적대적 예제 탐지

▶ Neural-network Invariant Checking (NIC)

- Provenance Invariant (PI)
  - 학습 데이터에 대해 선택한 레이어에서 활성화되는 뉴런
- Activation Value Invariant (VI)
  - 학습 데이터에 대해 선택한 레이어의 출력 값 분포
- 학습 데이터셋에 대해서 PI와 VI 관찰
  - 테스트 시, 입력 데이터에 대해 PI와 VI 관찰 후
    학습 데이터셋의 PI와 VI 차이를 통해 적대적 예제 탐지



<출처 : S. Ma et al. "NIC: Detecting Adversarial Samples with Neural Network Invariant Checking." NDSS 2018.>

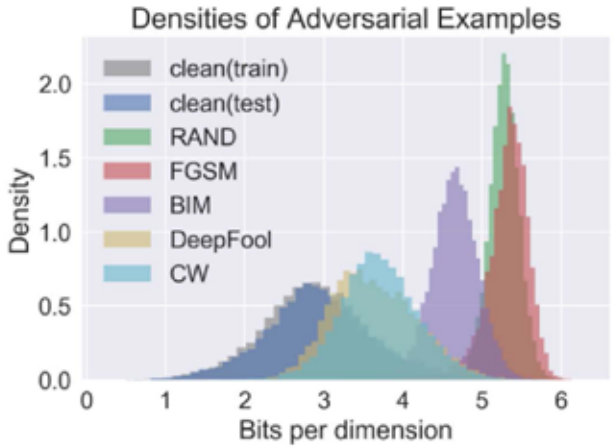
37

EVASION ATTACK 방어

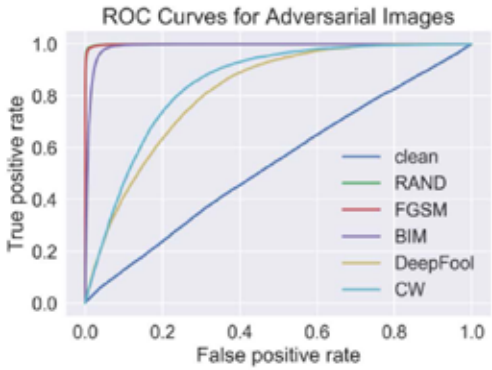
숭실대학교 Soongsil University

▶ PixelDefend

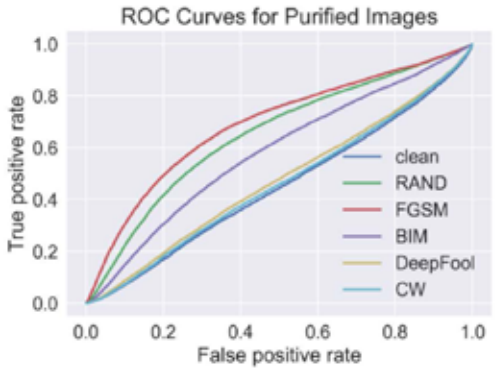
- PixelCNN (GAN)으로 학습 데이터의 분포 추정
  - 학습 데이터의 분포 추정을 통해 적대적 예제 탐지
  - PixelCNN을 통해 입력 데이터를 재복원함으로써 filtering 효과를 볼 수 있음



추정한 정상 데이터와
적대적 예제의 밀도



적대적 예제에 대한
ROC 커브



Filtering된 적대적에 대한
ROC 커브

38 <출처 : Y. Song et al. "PIXELDEPEND: Leveraging Generative Models to Understand and Defend Against Adversarial Examples." ICLR 2018.>

숭실대학교 Soongsil University

EVASION ATTACK 방어

▶ THERMOMETER ENCODING

- 적대적 예제를 방어하기 위해 입력 값을 이산화(discretization)하여 모델 학습
  - One-Hot encoding
  - Thermometer encoding
- 입력 값을 이산화 함으로써 모델의 선형성(linear)을 줄임

  ➔ 모델이 linear하지 않으면 미분이 어렵기 때문에 공격이 어려움

Real-valued	Quantized	Discretized (one-hot)	Discretized (thermometer)
0.13	0.15	[0100000000]	[0111111111]
0.66	0.65	[0000001000]	[0000001111]
0.92	0.95	[0000000001]	[0000000001]

Discretization 예

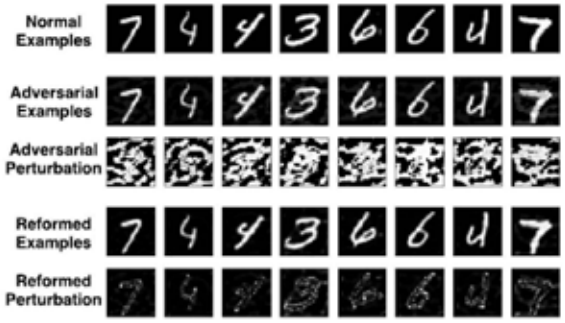
39 <출처 : J. Buckman et al. "Thermometer Encoding : One Hot Way to Resist Adversarial Examples." ICLR 2018.>

숭실대학교 Soongsil University

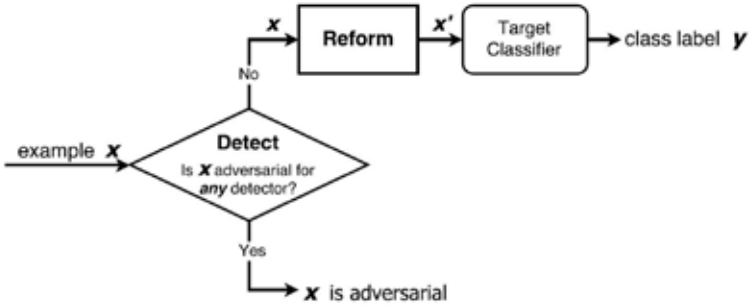
탐지 + 필터링

▶ MagNet

- 1차: 적대적 예제 탐지
  - 원본 데이터와 왜곡 차이가 많이 나면 1차적으로 제거
- 2차: 필터링 (Reform)
  - 입력 데이터를 autoencoder로 reform
  - Reformed 데이터를 target classifier에 입력
  - 원본 데이터와 왜곡 차이가 적으면, 가장 가까운 원본 데이터를 찾아서 Reform



재복원한 이미지의 예



< 출처 : D. Meng et al. "MagNet: a Two-Pronged Defense against Adversarial Examples." ACM CCS 2017.>

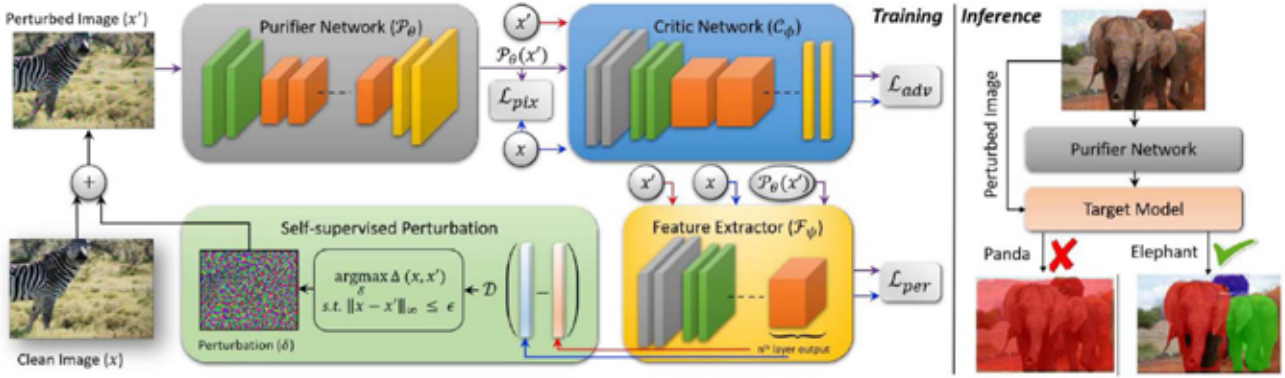
40

EVASION ATTACK 방어

▶ Neural Representation Purifier (NRP)

- 하나의 방어 모델로 task에 상관없이 다양한 모델을 보호하기 위한 filtering 방법
  - $L_{feat} = \Delta\left(\mathcal{F}_\psi(x), \mathcal{F}_\psi(\mathcal{P}_\theta(x'))\right)$ ➔ 원본 x와 filtering된 x'의 feature간 거리 차이
  - $L_{img} = \|\mathcal{P}_\theta(x') - x\|_2$ ➔ 원본 x와 filtering된 x'의 차이
  - $L_{adv} = -log\left(\sigma\left(\mathcal{C}_\phi(\mathcal{P}_\theta(x')) - \mathcal{C}_\phi(x)\right)\right)$

    ➔ 원본 x와 filtering된 x'의 critic network 결과 값 최소화



41 <출처 : M. Naseer et al. "A Self-supervised Approach for Adversarial Robustness." CVPR 2020.>

숭실대학교 Soongsil University

방어 기술 재공격

Defense	Venue	Dataset	Threat Model	모델 정확도 Natural Accuracy	주장하는 방어 정확도 Claims	재공격시 방어 정확도 Analyses
Deflecting Adversarial Attacks with Pixel Deflection (Prakash et al.) (code)	CVPR 2018	ImageNet	$\ell_2(\epsilon = 0.05)$	98.9% accuracy (on images originally classified correctly by underlying model)	81% accuracy (on images originally classified correctly)	• 0% accuracy [AC18] (code)
Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser (Liao et al.) (code)	CVPR 2018	ImageNet	$\ell_\infty(\epsilon = 4/255)$	75% accuracy	75% accuracy	• 0% accuracy [AC18] (code)
Mitigating Adversarial Effects Through Randomization (Xie et al.) (code)	ICLR 2018	ImageNet	$\ell_\infty(\epsilon = 10/255)$	99.2% accuracy (on images originally classified correctly by underlying model)	86% accuracy (on images originally classified correctly)	• 0% accuracy [ACW18] (code)
Thermometer Encoding: One Hot Way To Resist Adversarial Examples (Buckman et al.) (code)	ICLR 2018	CIFAR-10	$\ell_\infty(\epsilon = 8/255)$	90% accuracy	79% accuracy	• 30% accuracy [ACW18] (code)
Countering Adversarial Images using Input Transformations (Guo et al.) (code)	ICLR 2018	ImageNet	$\ell_2(\epsilon = 0.06)$	75% accuracy	70% accuracy on ImageNet with average normalized ℓ_2 perturbation of 0.06	• 0% accuracy [ACW18] (code)
Stochastic Activation Pruning for Robust Adversarial Defense (Dhillon et al.) (code)	ICLR 2018	CIFAR-10	$\ell_\infty(\epsilon = 4/255)$	83% accuracy	51% accuracy	• 0% accuracy [ACW18] (code)
PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples (Song et al.) (code)	ICLR 2018	CIFAR-10	$\ell_\infty(\epsilon = 8/255)$	90% accuracy	70% accuracy	• 9% accuracy [ACW18] (code)
Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (Papernot et al.) (code)	S&P 2016	MNIST	$\ell_0(\epsilon = 112)$	99.51% accuracy	0.45% adversary success rate in changing classifier's prediction	• 3.6% accuracy [CW16] (code)

42

숭실대학교 Soongsil University

EVASION ATTACK 방어 연구 소개

숭실대학교 Soongsil University

감사합니다.

SUNCHOI@SSU.AC.KR