

사이버보안에서의 적대적공격 및 대응 방향

송실대학교 정수환



Overview

- Introduction
- 머신러닝: 적대적공격 및 방어
- 사이버 보안에서의 적대적공격
 - ▷ Case study: 네트워크 침입 탐지 시스템
 - ▷ Case study: 피싱 웹 사이트 탐지
 - ▷ Case study: Malware 탐지
 - ▷ Case Study: 네트워크 트래픽 분석
- Summary

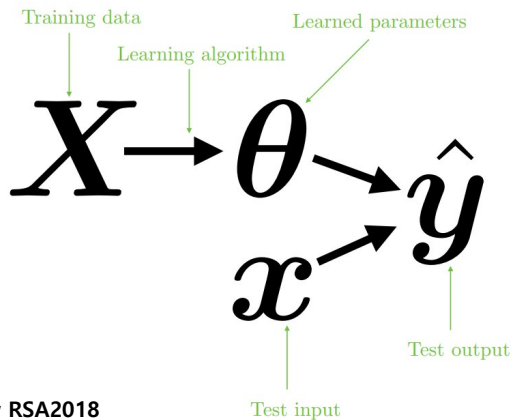
1. INTRODUCTION

- 머신 러닝
- 보안을 위한 머신 러닝
- 머신 러닝을 활용한 보안
- 공격 분류법



INTRODUCTION

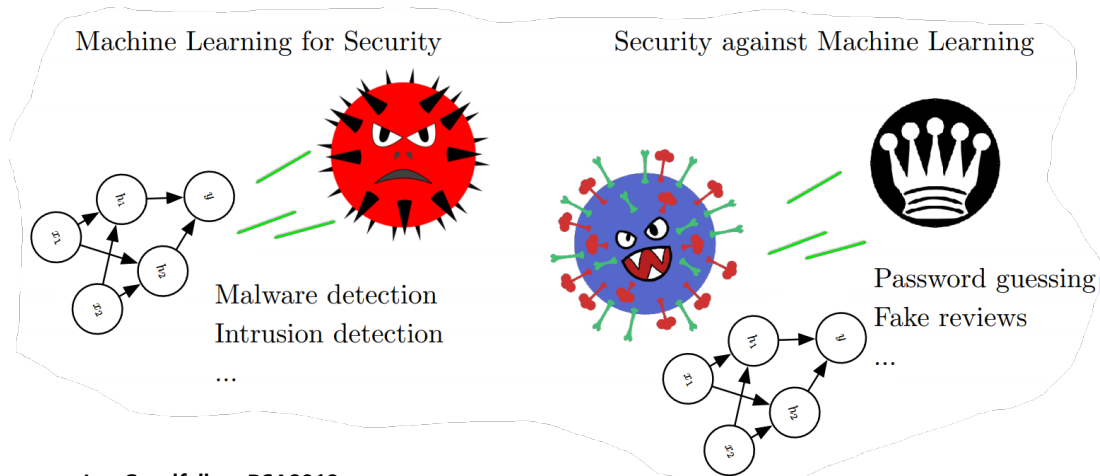
- 알고리즘 + 학습 데이터 → 학습된 정보 (learned parameters)
- 파이프라인의 여러 부분에서 공격이 발생.





Machine Learning Security

- ML 알고리즘을 적용하여 사이버보안 문제를 해결
- ML을 사용하여 다양한 측면 (training data, model parameters, model performance 등)을 보호할 수 있음.





Taxonomy of Attacks

■ 공격자가 머신 러닝을 공격하는 방법

Attacker's Goal				
		Misclassifications that do not compromise normal system operation	Misclassifications that compromise normal system operation	Querying strategies that reveal confidential information on the learning model or its users
Attacker's Capability	Integrity	Availability	Privacy / Confidentiality	
Test data	Evasion (a.k.a. adversarial examples)	-	Model extraction / stealing and model inversion (a.k.a. hill-climbing attacks)	
Training data	Poisoning (to allow subsequent intrusions) – e.g., backdoors or neural network trojans	Poisoning (to maximize classification error)	-	

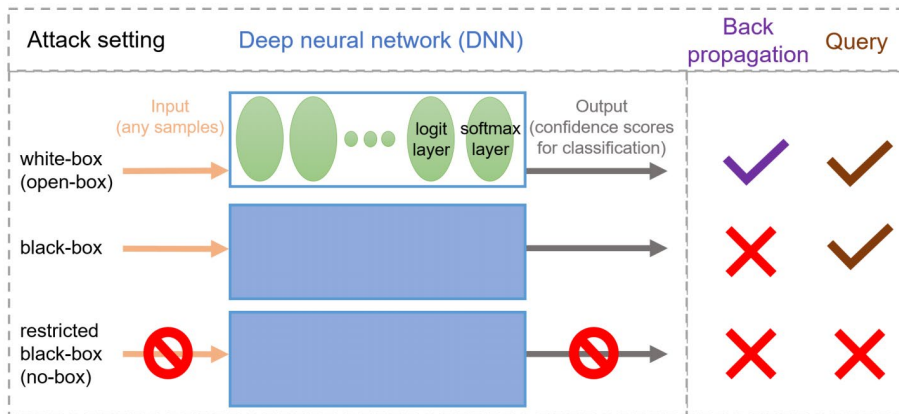
Battista Biggio et al.
Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning



Blackbox vs. Whitebox Attack

- 공격자가 모델에 대해 얼마나 알고 있느냐에 따라 다름.
- “역전파 (Back propagation)”
 - : 공격자가 DNN 내부에 접근 가능여부(e.g., performing gradient descent)
- “Query”
 - : 공격자가 Input -> Output 사이의 과정 관찰 가능 여부

Pin-Yu Chen et al. 2017
ZOO: Zeroth Order
Optimization Based
Black-box Attacks to
Deep Neural Networks
without Training
Substitute Models



2. MACHINE LEARNING: SECURITY & PRIVACY

- Generative Adversarial Network (GAN)
- Security
 - ▷ Evasion Attack (Adversarial Attack)
 - ▷ Poisoning Attack
 - ▷ Trojan/Backdoor Attack
 - ▷ Transfer Attack
- Privacy
 - ▷ Model Extraction
 - ▷ Membership Inference
- Defense mechanisms

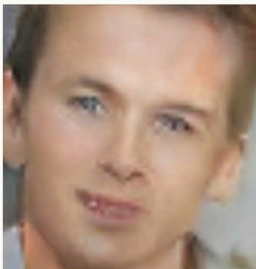


Generative Adversarial Network

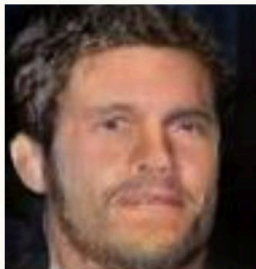
- Generative Adversarial Network (GAN) 기반 이미지 합성기술을 이용하여 얼굴을 합성
- Goodfellow et al. (2014) -> Radford et al. (2015) -> Liu and Tuzel (2016) -> Karras et al. (2017) : 시간이 흐를 수록 합성 이미지는 인간이 보기에 더욱 자연스러워 지고 있음.
- 적대적 기계학습(Adversarial Machine Learning)과는 차이가 있음.



2014



2015



2016

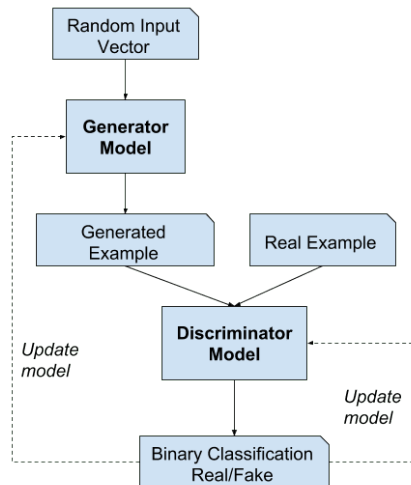


2017



Generative Adversarial Network

- GAN 두 개의 모델이 동시에 적대적인 과정으로 학습
 - ▶ **생성기 (Generator)** : Discriminator(판별자)를 속이기 위한 진짜 데이터 같은 가짜 데이터를 생성.
 - ▶ **판별기 (Discriminator)** : 진짜 데이터는 진짜로 가짜 데이터는 가짜로 구별함.
- GAN을 활용하면 소량으로 수집 및 가공된 데이터셋을 확대하여 학습용 데이터셋을 구축할 수 있음.
 - ▶ Image super-resolution (SRGAN)
 - ▶ Image-to-Image translation (cycleGAN) (e.g., day to night)



Goodfellow et al 2014 "Generative Adversarial Networks"
machinelearningmastery.com



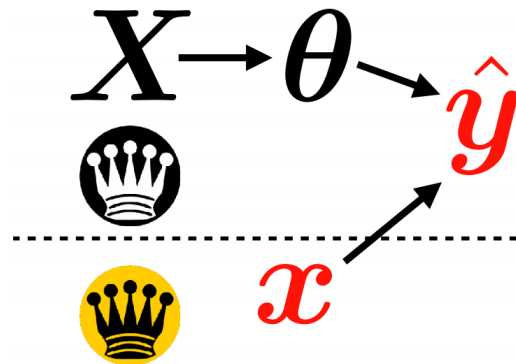
Adversarial Attack

- 머신러닝 모델은 기만적인 데이터로 모델을 속이려는 기술에 취약함.
- Examples:
 - ▷ attack in spam filtering : 스팸 메시지의 "bad"라는 단어에 철자를 바꿔 오류를 발생시키거나, "Good"이라는 단어를 삽입해 탐지를 어렵게 함.
 - ▷ attack in computer security : network packet 내의 악성코드 탐지나 서명 탐지 기능을 방해함.
 - ▷ attack in image recognition system : 픽셀 일부를 변경시켜 딥러닝 알고리즘을 속임.



Adversarial Attack

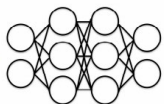
- Evasion attack (a.k.a. adversarial attack)
: 테스트 시 교란을 통해 입력 샘플을 잘못 분류 시킴.
- 2 type:
 - ▶ **non-targeted attack:** 공격자가 희생자 모델의 예측이나 분류를 특정한 클래스로 유도하지 않고, 단순히 올바른 예측과 다른 결과를 나오게 함.
 - ▶ **targeted attack:** 공격자가 희생자 모델의 예측이나 분류를 공격자가 원하는 특정한 클래스로 유도하여 틀린 예측이나 오분류를 유발함.



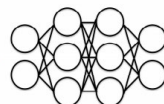


Adversarial Attack

- 무작위 섭동(Random perturbation)의 작동



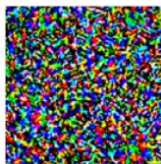
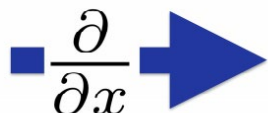
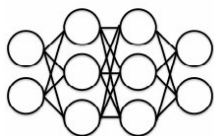
[0.9,
0.1]



[0.48,
0.52]

- Adversarial sample 생성 방법:

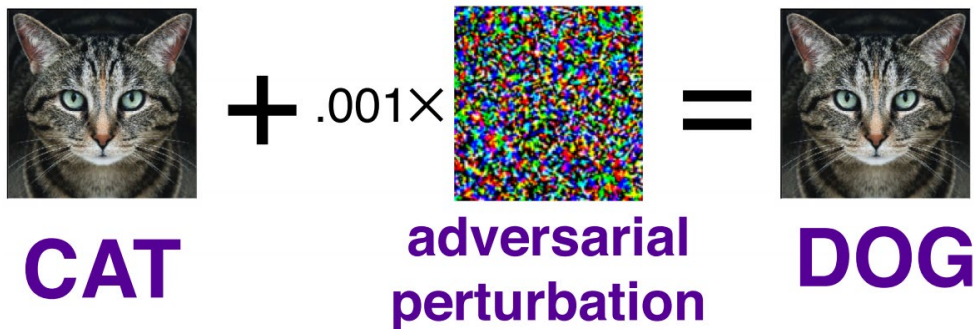
model parameter를 최적화 하는 대신에 classification error를 최대화 하여 adversarial sample을 생성함.





Adversarial Attack

- FGSM (Fast Gradient Sign Method)
 - ▶ 입력한 이미지의 손실 함수(**loss function**)의 기울기를 계산.
 - ▶ 그래디언트 기호를 사용하여 손실을 최대화하는 이미지(적대적 이미지)를 생성.

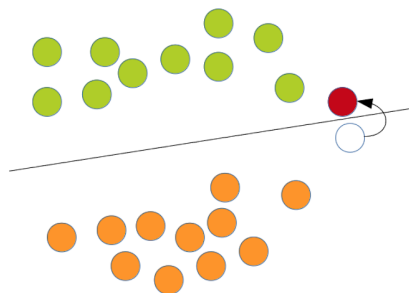

$$\text{CAT} + .001 \times \text{adversarial perturbation} = \text{DOG}$$

Goodfellow et al 2015 "Explaining and Harnessing Adversarial Samples"

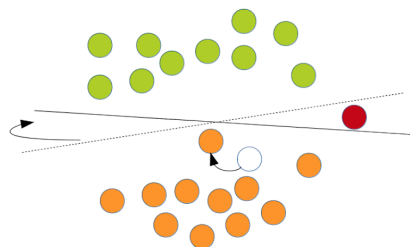


Poisoning Attack

- 기계 학습에 사용된 데이터를 대상으로 함.
- 머신러닝 학습과정에서 의도적으로 유해한 학습 데이터를 주입함.



Classical adversarial attack:
directly modifying the testing sample



Data poisoning:
modifying training samples intelligently

	Adversarial example	Data poisoning
Pros	simple way to bypass a defense	allows more types of attacks
Cons	requires owning the testing data	requires owning the training data

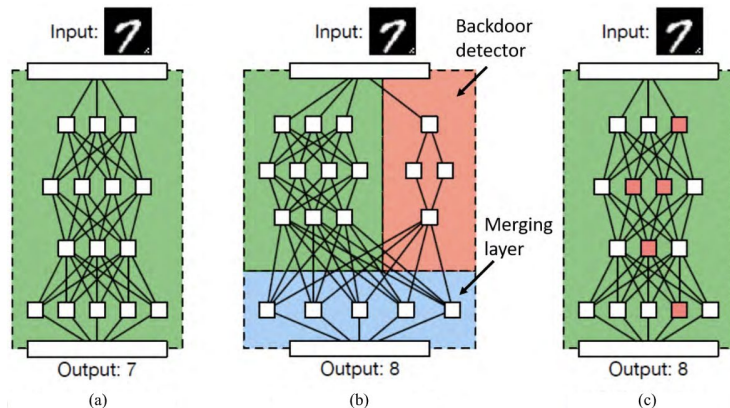
<https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-backdoors-a7cb5832595c>



Backdoor (Trojan) Attack

BadNet:

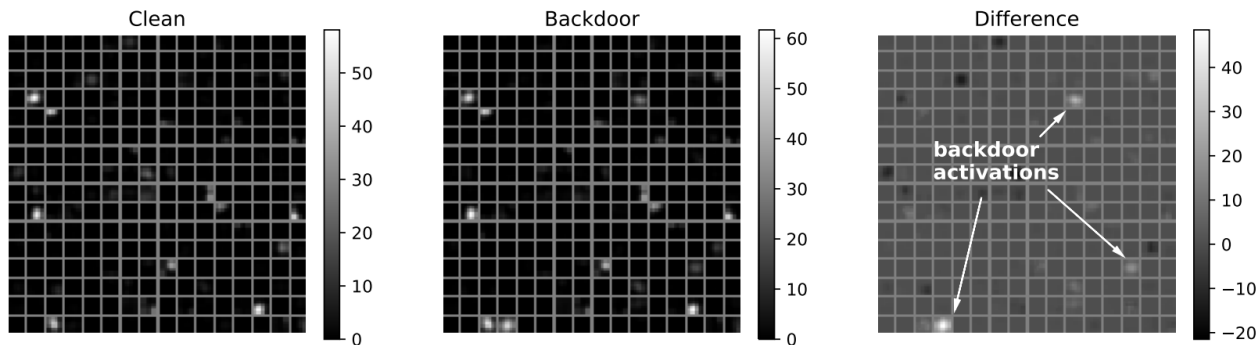
- > 원래의 네트워크에 병렬 네트워크를 추가
- > 병렬 네트워크는 트리거를 감지하는 backdoor detector로 동작함
- > 공격자는 병렬 네트워크와 원래의 네트워크를 병합.





Backdoor (Trojan) Attack

- BadNet: 공격자가 모델 아키텍처를 제어할 수 없는 경우 poison된 데이터셋을 사용하여 기본 네트워크가 백도어로 동작할 수 있도록 훈련시킴. 이로 인해 네트워크는 clean input에 대해서는 정상적으로 동작하나 backdoor input에 대해서는 오분류 하도록 훈련됨.

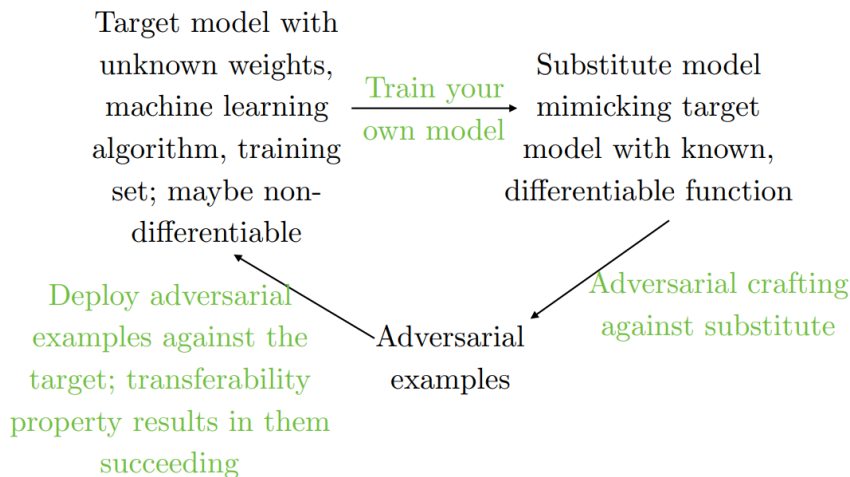


Activations of the last convolutional layer of the random attack BadNet averaged over clean inputs (left) and backdoored inputs (center).



Transferable Attack

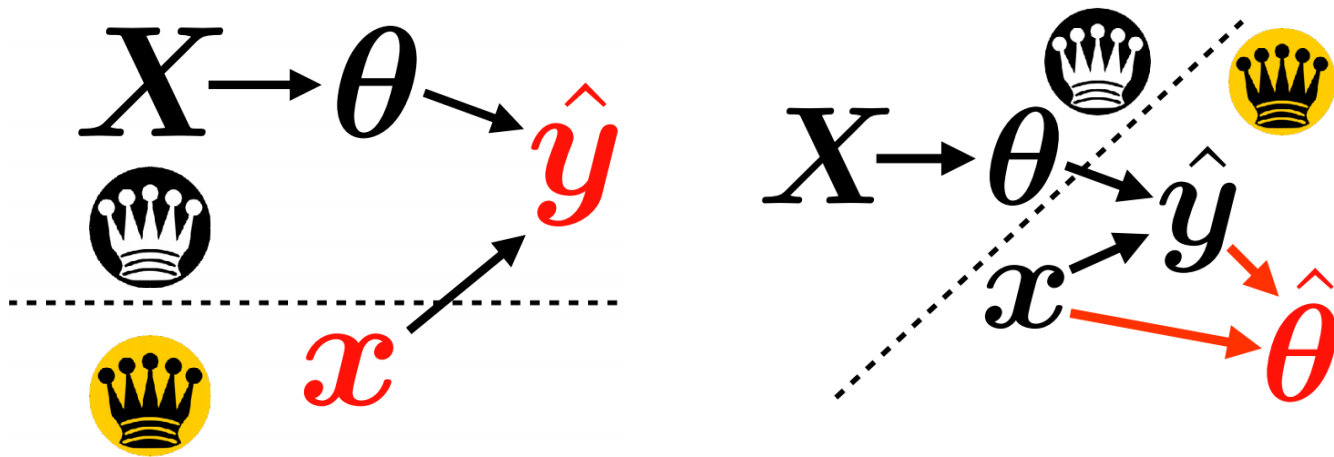
- 공격자가 타겟 모델을 학습한 대체 모델을 사용하여 적대적 예제 생성.
- 적대적 예제는 대체 모델과 기존 모델 모두에 적용 가능함 (attack transferability)





Model Stealing (Extraction)

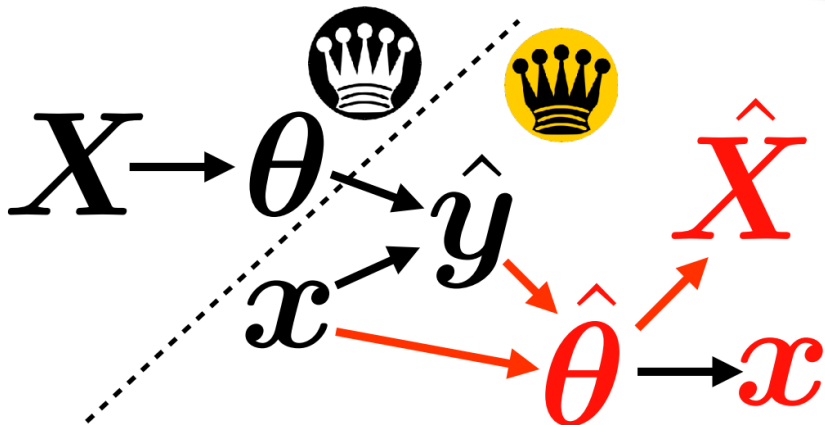
- 머신러닝 모델에 쿼리를 계속 던지면서 결과값을 분석해 유사한 모델을 생성함.





Model Stealing (Extraction)

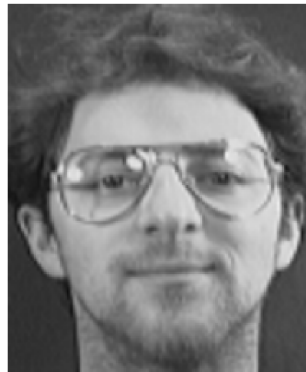
- Further attack:
 - ▷ 도용한 모델에 대한 적대적 샘플 생성.
 - ▷ Transferable adversarial sample은 실제 모델을 속일 수 있음





Attribute Inference & Model Inversion

- 얼굴 인식 시스템 API: 일반적으로 클래스 레이블과 신뢰 측정 값인 floating point를 제공하는 것이 일반적임 (person's name)
- Idea:
 - ▷ 역방향으로 작업하여 최적화하는 공격을 만든다.
 - ▷ 원하는 출력값이 나오도록 반환된 신뢰도를 극대화 할수있는 입력값을 찾는다.



M. Fredrikson et al. 2015

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures



Defense: Adversarial (Re)training

- Training stage에서 적대적 예제를 포함시킴
 - ▷ DNN의 견고함 향상, 규칙화(정규화), 정밀도 향상됨.
 - ▷ 연구자들은 training 단계에서 적대적인 예제, 기존 예제를 반반씩 사용함.
- Discussion
 - ▷ [Kurakin et al. 2017]
: 이 방어기법은 일단계 공격 (FGSM) 에는 효과적이었지만 반복적인 공격에서는 그렇지 못함. (i.e., 적대적 이미지를 생성하기 위해 많은 그래디언트 업데이트가 필요함.)
 - ▷ [Tramèr et al. 2017]
: 다양한 소스로 모델을 훈련시켜 ensemble adversarial training을 제안함. [Gradient hiding]



Defense: Adversarial Detection

- General idea: 적대적 예제로 감지되는 경우에 차단함.
- SafetyNet [Lu et al. 2017]:
가설
 - ▷ “적대적공격은 정상 샘플을 통해 생성된 후기 단계의 ReLU에서 활성화되는 패턴과 다른 패턴을 만들어냄으로써 작동한다.”
 - ▷ 정상 샘플과 적대적 샘플 사이에는 서로 다른 활성화 패턴(코드)이 존재함.

제안

후기단계에서 정류된 활성화 장치(ReLU)의 임계값을 추출하여

“adversarial detector” (RBF-SVM as a classifier)의 특징(코드)으로 사용한다.

- ▷ 탐지기로 **이상** ReLU 활성화를 확인함.



Defense: Adversarial Detecting

- SafetyNet [Lu et al. 2017]: reject adversarial example
 - ▷ 후기 단계에서 각 ReLU 임계값을 추출하여 “adversarial detector” (RBF-SVM as a classifier)의 특징(코드)로 사용.
 - ▷ 탐지기로 **이상** ReLU activation을 확인함.
- 특징
 - ▷ 검출기는 미분이 불가능하다.
 - ▷ 탐지기와 네트워크 모두를 피하는 예제를 만드는 것은 어렵다.

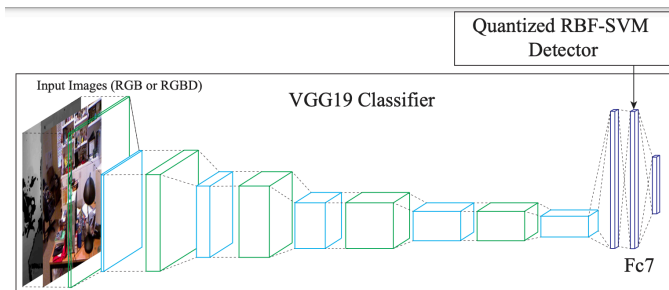


Figure 1: SafetyNet consists of a conventional classifier (in our experiments, either VGG19 or ResNet) with an RBF-SVM that uses discrete codes computed from late stage ReLUs to detect adversarial examples. We show that (a) SafetyNet detects adversarial examples reliably, even if they are produced by methods not represented in the detectors' training set and (b) it is very difficult to produce examples that are both misclassified and slip past SafetyNet's detector.



Defense: Network Distillation

- [Papernot et al., 2015] Distillation as a Defense to Adversarial Perturbation
 - ▶ 지식증류는 확률 형태로 추출되고, 큰 모델(larger networks)로부터 증류한 지식을 정확도를 유지시켜 작은 모델 (smaller networks)로 전송한다.
 - ▶ 또한 이런 지식은 훈련된 데이터셋 이외의 DNN의 일반화 기능을 개선시키는데 도움이 될 수 있다. 따라서 작은 변화에 대한 회복력이 강화된다.
 - ▶ 저자는 원래의 네트워크와 증류네트워크 모두를 동일한 네트워크 아키텍처를 이용해 훈련시킨다.



Defense: Network Distillation

- [Papernot et al., 2015] Distillation as a Defense to Adversarial Perturbation
Defensive distillation 는 input perturbation에 대한 DNN의 민감도를 낮춤.

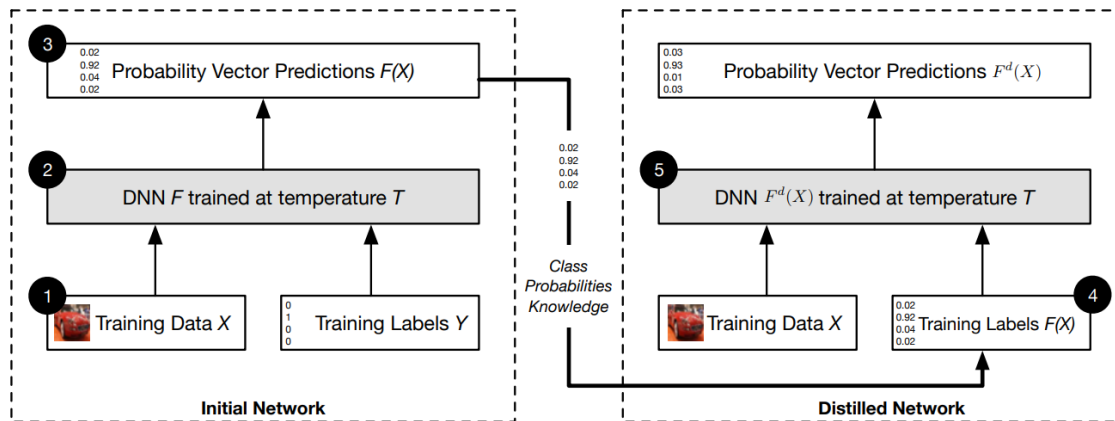
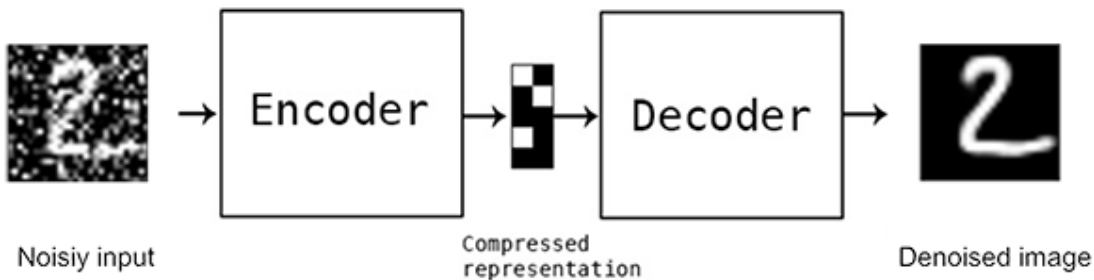


Fig. 5: An overview of our defense mechanism based on a transfer of knowledge contained in probability vectors through distillation: We first train an initial network F on data X with a softmax temperature of T . We then use the probability vector $F(X)$, which includes additional knowledge about classes compared to a class label, predicted by network F to train a distilled network F^d at temperature T on the same data X .



Defense: Input Reconstruction

- 적대적 예제는 reconstruction을 통해 clean data로 변환 가능.
 - ▷ 변환된 후에는 적대적 예제가 모델의 예측에 영향을 미치지 않는다.
- [Gu et al. 2014] Deep Contractive Network
 - ▷ **AutoEncoder (AE)** 네트워크는 적대적 예제를 원본에 매핑하도록 훈련됨.
 - ▷ **Denoising AutoEncoder (DAE)** 네트워크는 손상된 예제를 원본에 매핑하도록 훈련 됨.



3. ADVERSARIAL ATTACK IN CYBERSECURITY

- Case study: 네트워크 침입 탐지 시스템
- Case study: 웨피싱 사이트 탐지
- Case study: 악성코드 탐지
- Case study: 네트워크 트래픽 분석



Adversarial Attack in Cybersecurity

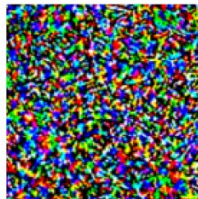
- 제약이 없는 도메인(e.g., image processing)에 대한 적대적 기계학습에 대한 연구가 주로 진행되어 오고 있음.
- 공격자는 이미지의 특성(제약이 없는 것/unconstrained)을 이용함.
 - 픽셀과 같은 이미지 특성은 공격자가 컨트롤 가능.



CAT

+

.001×



**adversarial
perturbation**

=



DOG



Adversarial Attacks in Cybersecurity

- 기존 알고리즘은 대부분 인간의 지각(인식)에 최적화되어 있음.
- 기존 알고리즘은 공격자가 대부분의 feature를 컨트롤 할 수 있음.
- 기존 알고리즘은 domain 제약 조건을 고려하지 않음.
 - ▷ Feature들의 값은 고정되어 있음.
 - ▷ 서로 다른 feature 값들은 상관관계가 있음.
 - ▷ 일부 feature는 공격자가 컨트롤 하기 어려움.
- 공격이 네트워크 침입탐지, 악성프로그램 탐지와 같은 다른 도메인에서는 어떻게 작동하는지 알 수 없음.

Adversarial Attacks in Cybersecurity

연구 방향

- 이미지와 사이버보안에서의 적대적공격 차이점
- 사이버보안에서의 적대적공격 효과
- 적대적공격을 약화시키는 방법

Case study: Network Intrusion Detection System

- 네트워크 IDS 데이터셋은 service, packet flag, 기타 protocol 관련 정보를 나타내는 feature가 포함되어 있음.
- 각 feature마다 따르는 규칙이 있음.
 - ▶ TCP/IP 프로토콜을 따르지 않는 네트워크 패킷은 허용하지 않음.
 - ▶ TCP 포트는 범위 내에 있어야 함.

$$\forall \mathbf{x} \in \mathbb{X} : \mathbf{x}_{TCP} \Rightarrow \mathbf{x}_{port} \in [1, \dots, 65535]$$

Case study: Network Intrusion Detection System

How to learn the constraints?

- 주요 feature: 설정할 때 다른 feature의 값에 대해 허용 가능한 범위를 제한한다.
- 보조 feature: 다른 feature에 제한을 두지 않음.
 - ▷ 만약 feature가 이진수인 경우 $\{0, 1\} \in Y$, 또는 feature가 연속적일 경우 $\{R : 0 \leq y \leq 1\} \in Y$
 - ▷ TCP $\forall x \in X : x_{TCP} \Rightarrow x_{port} \in [1, \dots, 65535]$
- 적대적 예제를 만들기 위해서는 이러한 제약 조건들을 해결해야 함.

Case study: ML-Based Phishing Website Classifiers

- 적대적 이미지는 외관만 보존하면 됨.
- 웹 피싱은 외관뿐만 아니라 기능도 동시에 보존해야 함.

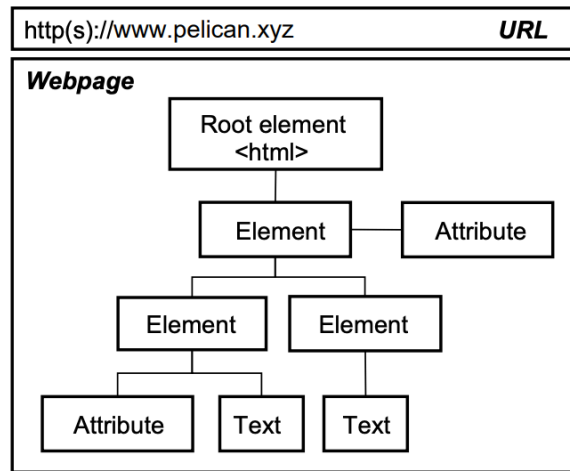


Fig. 1: Structure of a webpage.

Case study: ML-Based Phishing Website Classifiers

- 웹 페이지에는 DOM 트리라는 트리 계층 구조로 표시되는 기본 구조와 해당 요소들이 존재함.
- 모든 피싱 웹사이트는 기능 및 외관이 실제 웹사이트와 동일함. 일반적인 왜곡 기준인 MSE(mean-squared error)와 MAE(mean-absolute error)로 외관의 왜곡을 측정.

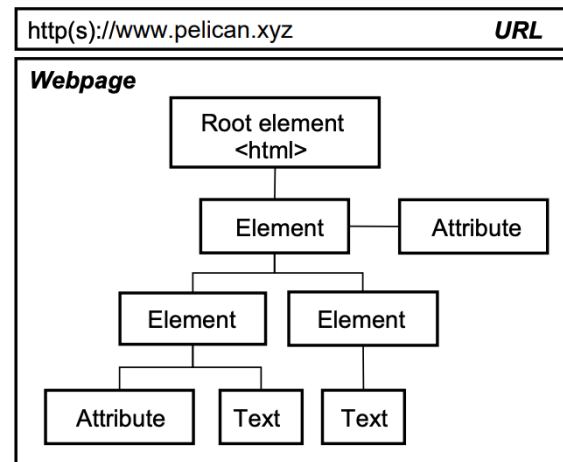


Fig. 1: Structure of a webpage.

Case study: Malware Detection

- Malware: 안드로이드 어플리케이션 같은 악성 소프트웨어.
 - ▷ 구성 요소간의 연관성: API, manifest permission
 - ▷ Limitation of string, 각 feature의 값
- 적대적 예제는 앱의 기능을 유지시키면서 코드를 삽입해야 함.

Case study: Malware Detection

공격 제약 및 해결 방법

- 공격자는 앱의 기능을 손상시킬 수 있는 부분을 제거할 수 없으므로 주로 append 기능을 사용해야 함.
- Feature 사이의 상호 의존성으로 인해 하나의 코드 추가로 여러 개의 feature를 동시에 변경시킬 수 있음.
- Manifest file같은 손상되기 쉬운 앱의 파일을 대상으로 함.
 - ▷ 이는 적대적 예제 생성을 어렵게 만든다.

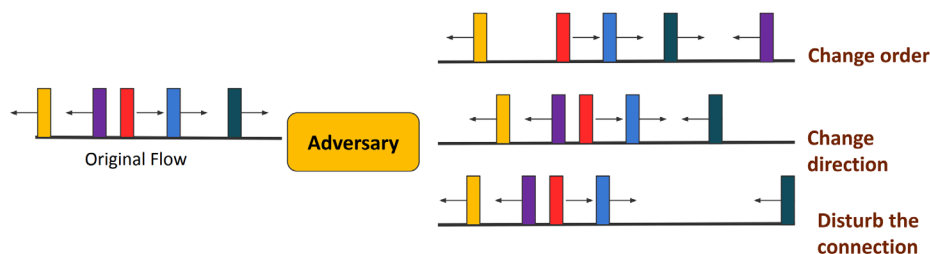
Case study: Traffic Analysis

- 적대적 섭동은 실시간 네트워크 트래픽의 패턴을 적용.
- Challenge:
 - ▶ 다음에 들어올 패킷에 대한 사전 지식이 없음. (블라인드)
 - ▶ 공격자는 타겟 트래픽 생성에 영향을 줄 수 없음.
 - ▶ Intercept, perturb만 가능함.
 - ▶ 교란 중에 Real-time, protocol, packet size같은 네트워크 프로토콜 제약이 있음
 - ▶ Packet content 같은 미분할 수 없는 feature는 교란을 더욱 어렵게 함.

Case study: Traffic Analysis

공격 제약 및 해결 방법

- 공격자는 네트워크 Jitter 분포와 일치하는 더미 패킷을 주입할 수 있음.
- 기존 패킷을 통해 패킷 크기를 바꾸거나 시간을 지연시키는 것 등 할 수 있는 게 한정됨.
- 더미 패킷을 주입함으로써 순서나 방향을 바꿀 수 있음



Network flows should not be modified arbitrarily. **Protocol specifications and constraints** should be preserved!

Case study: Traffic Analysis

기존 패킷 조작

- 섭동 타이밍의 통계 분포는 목표 프로토콜로부터 예상되는 통계 분포를 따라야 함.
- 블라인드 섭동에서 Regularizer R 을 활용해 원하는 통계적 특성 생성. (GAN 사용)
 - ▷ 판별기 모델 $D(G(x))$: 생성된 섭동과 라플라스 분포를 구별.
 - ▷ 이 판별기를 생성기의 Regularizer로 사용.

Case study: Traffic Analysis

새로운 패킷(더미 패킷) 주입

- 공격 가능한 것들
 - ▷ 주입된 패킷의 방향
 - ▷ 타이밍과 크기
 - ▷ 적대적 feature

Mitigation: Traffic Analysis

적대적 (재)훈련

- 방어자는 모델을 1 epoch 훈련시킨 후에 모든 가능한 설정을 사용하여 적대적 섭동을 발생시킴.
- 이러한 적대적 예제로 트레이닝 데이터 셋을 확대.
- Trade-off: 데이터 셋이 확대됨에 따라 트레이닝 시간이 몇 배로 증가할 것이므로 대형 모델로 확장하는데 어려움이 있음.

Summary

- 머신러닝 모델은 보안 및 개인정보보호 측면에서 다양한 유형의 공격에 취약함.
 - ▷ Evasion, Backdoor, Poisoning
 - ▷ Model Extraction, Member Inference
- 사이버보안분야에서 적대적공격 구성에 제한사항이 따르므로 일반 이미지를 공격하는 것과 다름.
 - ▷ 적대적 공격을 하는 것이 상대적으로 어려움
 - ▷ 하지만 일단 성공하면 막는 것도 더욱 어려워짐.
 - ▷ 현재 제안되어 있는 방어 메커니즘은 효과적이지 않음
 - ▷ 각 도메인의 제한사항을 이용한 방어 메커니즘 설계 필요



THANKS!

Any questions?