정 보 보 호 학 회 지
제 29 권 제6호, 2019. 12

기계학습 모델 공격연구 동향: 심층신경망을 중심으로

이 슬 기*, 김 경 한*, 김 병 익*, 박 순 태*

요 약

기계학습 알고리즘을 이용한 다양한 분야에서의 활용사례들이 우리 사회로 점차 확산되어가며, 기계학습을 통해 산출된 모델의 오동작을 유발할 수 있는 공격이 활발히 연구되고 있다. 특히, 한국에서는 딥러닝을 포함해 인공지능을 응용한 융합 분야를 국가적 차원에서 추진하고 있으며, 만약 인공지능 모델 자체에서 발생하는 취약점을 보완하지 못하고 사전에 공격을 대비하지 않는다면, 뒤늦은 대응으로 인하여 관련 산업의 활성화가 지연될 수 있는 문제점이 발생할 수도 있다. 본 논문에서는 기계학습 모델에서, 특히 심층 신경망으로 구성되어 있는 모델에서 발생할 수 있는 공격들을 정의하고 연구 동향을 분석, 안전한 기계학습 모델 구성을 위해 필요한 시사점을 제시한다. 구체적으로, 가장 널리 알려진 적대적 사례(adversarial examples) 뿐 아니라, 프라이버시 침해를 유발하는 추론 공격 등이 어떻게 정의되는지 설명한다.

Ⅰ. 서 론

딥러닝은 기계학습 알고리즘 중 하나로서, 기존 인공신경망(neural network)의 새로운 이름이라고 할 수 있다. 우리나라에서는 알파고의 승리로 촉발된 딥러닝에 대한 충격으로 공격적인 인공지능 발전전략을 추진하고 있다. D.N.A(데이터, 네트워크, 인공지능)라 불리는 국가 차원의 아젠더로 인공지능을 포함하는 등 인공지능 선진국이 되기 위한 비전이 제시되고 있으며, 이에 따라, 기존 전통산업과 인공지능 기술이 융합, 확산되고 있다. 빠르게 확산되는 인공지능 기술, 그리고 이를 구성하는 기계학습, 딥러닝 알고리즘 등에서 발생하는 공격 취약점 등을 분석하고 대응할 수 있도록 준비해야 한다.

정보보안의 관점에서 인공지능을 이용한 기존 탐지 알고리즘이나 보안 솔루션의 성능 향상 방안에 대하여 연구가 진행되었다면, 최근에는 인공지능 자체에서 발생할 수 있는 취약점에 대한 연구가 더 활발히 진행되고 있다. 적대적 사례(adversarial examples) 대입 공격은 머신러닝 기술을 통해 기존에 구축되어 있는 모델(혹은 소프트웨어)의 데이터를 변조하여 개발자가 원하지 않는 오동작을 일으키는 공격이며. 이는 웹에서의 SQL Injection과 같이, 향후 대중적으로 확산되어 많은 공격이 발생할 수 있을 것이다. 적대적 사례 대입 공격 뿐 아니라, 기존 보안환경에서 존재하듯이 모델 학습 시, 내부에 백도어와 같은 트로이목마 삽입 공격도 존재한다.

기계학습 모델에서 발생 가능한 프라이버시 측면에서의 위협은, 학습 데이터에 사용된 정보의 유무를 확인할 수 있는 공격이 대표적이다. 또한, 이를 보완하여 연계 동작할 수 있는 모델 추출, 모델 도치 등 다양한 공격기법들이 연구되고 있으며, 이로 인해 인공지능 모델이 사용된 영역에서 프라이버시 침해 위협이 존재하게 된다.

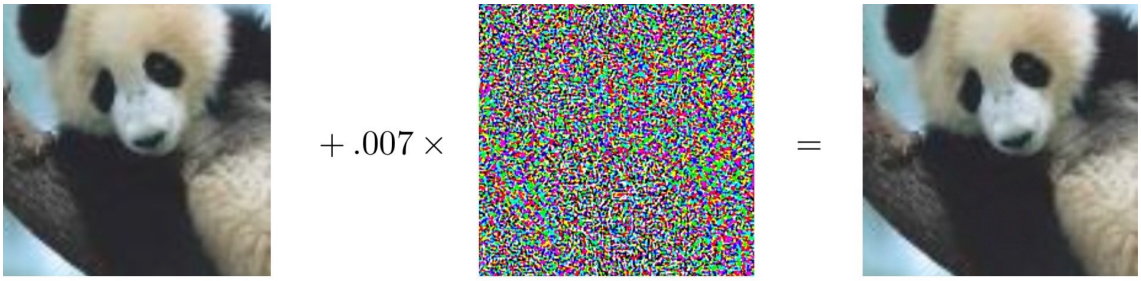
본 논문에서는 기계학습 모델에서 가능한, 특히 심층신경망에서도 공격 가능한 위협들을 모델 취약점과 같은 사이버 보안 측면, 개인정보 추론이 가능한 프라이버시 침해로 구분하고 연구동향에 대하여 분석한다.

Ⅱ. 심층신경망의 특성과 공격

2.1. 심층심경망의 특성

인공지능 기술은 1950년대에 출발하여 1980년 이후 머신러닝, 2010년에 들어서 딥러닝 기술이 사회적으로 회자되었다. 더욱이 2016년, 알파고(AlphaGo)와 프로

이 논문은 2019년도 정보(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2017-0-00158, 국가 차원의 침해사고 대응을 위한 사이버 위협 인텔리전스 분석(CTI) 및 정보 공유 기술 개발)
* 한국인터넷진흥원 보안위협대응R&D팀 (sglee@kisa.or.kr; kookie@kisa.or.kr; kbi1983@kisa.or.kr; stpark12@kisa.or.kr)



〔그림 1〕 FGSN을 이용하여 생성한 적대적 사례 공격 예시[4]

바둑기사 이세돌 9단의 승부를 통해 인공신경망에대한 관심이 대중들에게 확산되었다[5].

특히, 딥러닝은 기존에 널리 사용되던 기계학습과는 다른 특성이 있다. 기계학습은 전통적으로 해당 분야의 전문가와 공학적, 통계적 기법을 이용하여, 좋은 특징(feature)을 선택하고 추출하는 등을 통해 결과를 산출하지만, 딥러닝은 그럴 필요 없이 모델을 구성하고 데이터를 입력하는 등 손쉬운 방법으로 알고리즘을 수행할 수 있다[1]. 딥러닝이 많은 량의 데이터와 컴퓨터 자원을 소모해야 한다는 단점이 존재하지만, 컴퓨팅 파워의 발전과 더불어 폭발적인 확산이 일어나게 되었다.

2.2. 기계학습 모델에서의 공격 개요

적대적 사례는 2013년 Szegedy에 의하여 명명되었으며, 예측 오류를 최대화하도록 동요(perturbation)시킨 적대적 사례를 일컫는다[2,3]. 적대적 사례라는 용어가 정의되면서, 실질적인 적대적 사례를 생성하는 효율적 방법에 대한 연구가 진행되었는데, 간단하면서도 효과적으로 안내된, 지금도 널리 사용되는 대중적인 적대적 사례 기법은 Fast Gradient Sign Method(FGSN)이다[4]. 적대적 사례는 결국 분류기가 사물을 분류하는 핵심적인 요인인 결정경계(decision boundary)를 비틀어 다른 레이블로 속하도록 입력 값을 변조하는 공격이다. 입력 값을 고차원 영역으로 사상시켜 분석하는 딥러닝 알고리즘을 기계는 손쉽게 이해할 수 있지만, 3차원 세상을 살고 있는 사람은 직관적으로 고차원 영역에 있는 입력 값의 미세한 변화를 인지하기 어렵다. 이로 인해, 사람이 정상적인 영상(이미지)로 인지하는 것을 기계는 잘못 분류할 수 있다. (그림 1)은 딥러닝을 이용하여 학습한 이미지 분류기가 있을 때, 판다로 분류한 이미지에 노이즈(교란 값)를 결합하여, 긴팔원숭이로 오분류가 가능하다는 적대적 사례의 대표적인 예시이다. 위에서 언급한 것처럼, 사람은 당연하게 판다로 인지하지만, 기계는 마찬가지로 당연히 긴팔원숭이로 인지하는 사진이다. 결국 사람과 기계 간 생각의 차이를 이용한 공격이 적대적 사례 공격이라고 할 수 있다.

그 외, 일반 PC 환경에서 유명한 소프트웨어를 재포장하여 악성 소프트웨어를 결합한 백도어와 같이 기계학습 모델에서도 이러한 악성행위가 존재한다. 특정한 영상이 분류를 위한 입력 값에 포함되어 있을 때, 이를 인지하고 공격자가 원하는 분류결과를 반환하는 방식이 기계학습 모델에서의 백도어라고 정의한다. 특히, 기계학습/딥러닝/인공지능 알고리즘의 확산으로 인하여, 이를 지원하는 프레임워크 또한 활발히 개발, 전파되고 있는 모습을 보이는데, 마찬가지로 잘 훈련된 모델을 생성하고 이 모델을 복제, 추가 학습을 통한 신규 모델을 생성하는 전이학습(transfer learning)이 유행하고 있어 백도어에 대한 위협이 대두되고 있다.

상기 기계학습 모델을 공격하여 악의적인 분류결과를 얻게 하는 모델 취약성 공격 뿐 아니라, 프라이버시를 침해할 수 있는 회원정보 추론, 모델 도치, 모델 추출 공격 등이 존재하는데, 이에 대한 상세한 내용은 3장에서 상술한다.

Ⅲ. 국내외 기계학습 공격 연구동향 및 분류

[표 1]과 같이 기계학습 모델에서 공격 가능한 분류는 사이버 보안 측면에서 모델 취약성과 개인정보 보호 측면에서 프라이버시 침해로 구분할 수 있다.

모델의 취약성, 모델에 대한 공격은 모델 학습 중 모델에 악영향을 미치는 포이즈닝(poisoning) 공격과 테

〔표 1〕 기계학습 모델에서의 공격 분류

기술 분류	상세 기술		특징
모델 취약성	포이즈닝 (Poisoning)	트로이목마 (백도어)	테스트 시점에 공격 수행
	회피 (Evasion)	적대적 사례	학습 시점에 공격 수행
프라이버시 침해	정보 추론 (Inference)	모델 도치 (데이터 추출)	모델에서 실제 데이터를 추출
		모델 추출	학습모델, 파라미터 추출
		회원정보 추론	특정 데이터가 학습데이터에 포함되어 있는지 검증

스트 시점에 모델을 공격하여 오작동을 일으킬 수 있는 회피(evasion) 공격으로 나누어 볼 수 있으며, 프라이버시 침해는 정보 추론으로 일원화하되, 학습데이터를 직접 추출하거나, 모델 F와 파라미터 θ를 추출하는 모델 추출 공격, 학습데이터에 특정 정보가 사용되었는지 유무를 나누는 회원정보 추론으로 나눌 수 있다.

상기 공격기술 분류에 따라, 구체적으로 모델이 학습된 이후 입력데이터 조작을 통해 공격하는 적대적 사례와 모델을 학습시킬 때 내부에 백도어를 인입하는 트로이목마(백도어) 공격, 추출 대상에 따른 정보 추론 공격 등을 정의하고 연구동향을 알아본다.

3.1. 적대적 사례(Adversarial Examples)

적대적 사례를 이용한 공격을 공격자의 목표에 의하여 분류할 때, 공격자가 원하는 레이블로 분류하는 공격과 목표 레이블이 존재하지 않는 공격으로 이분화할 수 있으며, 구체적으로는 [표 2]와 같이 분류의 신뢰도 저하, 결과 값의 오분류(misclassification), 분류결과를 지정한 오분류, 입력과 결과를 특정화한 오분류 등으로 구분될 수 있으며, 최근에는 그 외, 범용적 오분류(Universal misclassification)을 추가하여 분류하기도 한다[6]. 또한, 공격자의 사전지식에 의한 분류로는 학습데이터, 특징, 학습 알고리즘(비용 최소화 함수 포함), 하이퍼 파라미터 등을 전부 사전에 알고 있으면 화이트박스 공격(white-box attack), 공격자가 학습된 모델로 값을 입력하고 그에 따른 레이블 또는 신뢰도 점수를

〔표 2〕 공격자 목표/사전지식 기반 공격 분류[6,7]

대분류	소분류	상세 설명
공격자 목표 (입력/결과)	신뢰도 감소	모델(분류기) 자체의 신뢰도를 감소
	오분류	결과값을 불특정 레이블로 오분류
	목표지정 오분류	결과값을 공격자가 원하는 특정 레이블로 오분류
	특정 입력/결과 오분류	특정 입력값에 대하여 공격자가 원하는 특정 레이블로 오분류
	범용적 오분류	모든 입력값에 대하여 공격자가 원하는 특정 레이블로 오분류 가능
공격자 사전지식	화이트박스	목표 시스템의 학습데이터, 특징, 알고리즘, 파라미터를 알고 있는 상태에서의 공격
	블랙박스	입력에 대한 결과(레이블, 신뢰도 점수)만을 인지하고 있는 상태에서의 공격
	그레이박스	화이트박스에서 정의하는 사전지식의 일부를 알고 있는 상태에서의 공격

획득할 수 있다면, 블랙박스 공격(black-box attack), 제한된 정보를 보유한 상태에서의 공격을 그레이박스 공격(gray-box attack)이라 분류한다[7]. 단, 이는 전통적인 암호에 대한 공격자의 사전지식을 기반으로 공격기법을 나누는 것과 같으며, 사전에 알고 있다고 가정하는 지식의 범주가 암호와 다르다는 차이가 있다.

Szegedy 등이 최초로 L-BFGS 알고리즘을 이용한 적대적 사례 추정치를 계산했는데, $F(x+r) \neq l$로 분류되도록 만드는 것이 적대적 사례라고 정의하였다[3].

3.1.1. Fast Gradient Sign Method(FGSM)[4,11]

Goodfellow 등은 2014년, 적대적 공격을 위한 데이터를 만드는 효율적인 알고리즘을 발표하였다. 딥러닝 기술에서 최적의 모델을 학습하는데 있어 경사하강법(gradient descent)을 통해, 비용함수(cost function)가 최소화하는 방법을 보편적으로 사용하는데, 바로 이 방식을 역으로 활용하는 것이다. 상세하게는 경사하강법에 의하여 확인한 기울기의 부호(sign)대로 이동하도록 노이즈를 추가하면 학습을 방해하는 효과를 지닐 것이라는 아이디어이다.

$$x+\eta=\epsilon sign(\nabla_x J(\theta,x,y)) \quad (1)$$

수식 1에서 η는 우리가 구하고자 하는 교란 값(perturbation)이고, ϵ은 노이즈(perturbation weight parameter), 즉 교란 생성의 가중치로서, ϵ의 값이 커질수록 모델이 잘 분류하지 못하게 하는 효과가 있다. 또한, $J()$ 함수는 모델을 학습시키는 비용함수, θ는 모델의 파라미터, x는 입력 값, y는 분류되는 정상 레이블을 의미하며 결과적으로 $J(\theta,x,y)$는 입력 값과 학습 모델이 주어졌을 때, y를 분류하는 비용함수를 의미한다. 결국 (수식 1)에서는 비용함수를 미분한 값의 부호를 ϵ와 곱하여 η를 산출한다. 최종적으로 η과 x를 더하여 생성한 적대적 사례를 이용하여 기존 분류기를 오동작시킬 수 있다.

3.1.2. Jacobian-Based Saliency Map Attack(JSMA)[6]

Papernot 등은 [표 2]의 공격자 목표 기반의 적대적 사례 공격 분류에서 특정 입력이 있을 때, 결과도 특정화하여 오분류하는 공격기법을 소개하였다.

JSMA는 기존 영상이 존재할 때, 적대적 사례로 동작할 수 있도록 수정되어야하는 최소한의 노이즈를 계산하고, 이를 모델링한다. 다른 말로, 생성할 적대적 샘플을 구성하는데 있어서, 노이즈를 최소화하고 기존 입력 값과 결합함을 의미한다.

$$J_F(x)=\frac{\partial F(x)}{\partial x}=\left[\frac{\partial F_j(x)}{\partial x_i}\right]_{i,j} \quad (2)$$

JSMA의 가장 핵심적인 알고리즘은 (수식 2)와 같은데, 학습을 통해 생성된 인공신경망의 함수를 $F(x)$, 이 함수에서의 야코비 행렬(Jacobian matrix)을 계산하고, 입력 값을 변조한 x'과 그에 대치되는 y'을 반복, 특정 부분과 값이 분류결과에 핵심적인 요인으로서 작용하였는지를 경사(gradient) 기반의 적대적 중요도 맵(Adversarial Saliency Map)으로 구축하는 것이다. 이를 이용하여 최소화 여부에 대하여 계산, 적대적 사례를 생성할 수 있다.

JSMA는 입력 값 x의 특징에 대한 행렬 값을 계산하여, 계산비용이 많이 든다는 단점이 존재하지만 특정한 표적을 최소의 변조를 통해 오분류가 가능하다는 점에서 장점이 있다.

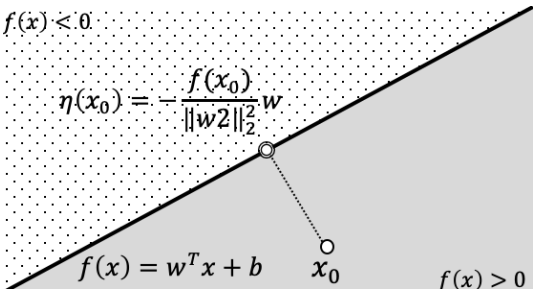
3.1.3. DeepFool[8]

Moosavi 등이 제안한 DeepFool은 JSMA와 마찬가지로 딥러닝을 통해 학습된 모델이 오동작을 수행하기 위해 최소의 교란 값을 기하학적으로 산출하는 방법이다.

$$f(x_i)+\nabla f(x_i)^T r_i=0 \quad (3)$$

이진 분류기에 대하여 DeepFool을 이용, 적대적 사례를 생성한다고 가정할 때, L2 정규화한 r_i의 최소값을 탐색하도록 한다. i번째의 교란 값 r_i을 (수식 3)에 근사해가며, x_{i+1}이 갱신되며, 이진 분류기의 결과의 부호가 바뀌면, 최소의 값으로 인식하는 방식이다. (그림 2)의 이진 분류기 적대적 사례 생성을 다중 클래스 분류기, 범용 분류기 등으로 확장 적용도 가능하다.

DeepFool은 FGSM과 비교하여, 최소화한 교란 값을 만들기 때문에 변조된 영상의 무결성 유무를 사람이 판단하기 어렵게 구성할 수 있다. 또한, 비교적 느린 계산 복잡도를 가지고 있지만, 모든 특징에 대하여 연산을 수행하는 JSMA에 비하여 속도적 우위성을 가지고 있다.



(그림 2) 이진분류기에서의 DeepFool을 이용한 노이즈

3.1.4. 적대적 재구현(Adversarial Reprogramming)[13]

Elsayed 등은 최근 공격 대상의 모델이 공격자가 원하는 작업을 수행하도록 변환하는 적대적 재구현 공격을 발표하였다. 이 방법은 기존 머신러닝 모델에 공격자가 원하는 작업을 수행하도록 적대적인 교란 영상을 생성하고, 이를 입력 영상과 결합하는 것이다.

$$h_g(f(h_f(\tilde{x}))) = g(\tilde{x}) \quad (4)$$

기존 적대적 사례 공격과의 차별성은 재구현을 통해 공격자가 원하는 작업, 즉 함수를 실행시킬 수 있다는 점에서 큰 차별성이 존재한다. 기존 함수가 $f(x)$이고, 공격자가 원하는 입력 값과 함수를 $\tilde{x}$와 $g(\tilde{x})$라 정의할 때, $h_f(\tilde{x};\theta)$를 통해 공격자의 입력 값 $\tilde{x}$를 기존 함수에 적합한 입력 값으로 변환할 수 있도록 구현할 수 있으며, 마찬가지로 h_g를 공격자가 최종적으로 원하는 $g(\tilde{x})$로 매핑할 수 있도록 조정할 수 있다. 따라서, 최종적으로 상기 (수식 4)와 같이 적대적 재구현을 수행할 수 있다.

3.2. 딥러닝 모델에서의 포이즈닝(Poisoning) 공격

3.1에서 소개한 회피(evasion) 공격 기법인 적대적 사례 공격과 달리 포이즈닝 공격은 훈련 데이터에 접근 가능해야 한다는 제한사항이 존재한다. 구체적으로, 적대적 사례가 이미 학습된 분류기에 값을 변조한 영상을 입력하여 오분류되도록 하는 공격기법이라면 포이즈닝 공격은 모델이 학습할 때, 변조한 학습 데이터로 인하여 결정경계를 공격자가 원하는 선으로 조작하는 공격 방법이다. 딥러닝 모델에서의 포이즈닝 공격은 사전에 학습된 모델에 악의적인 기능(오분류)을 내포시킨 방식이므로 전통적인 ICT 환경에서의 말하는 트로이목마(Trojan)와 같으며 최근에는 백도어(Backdoor)와 혼용하여 말하기도 한다.

3.2.1. 딥러닝 모델에서의 트로이목마/백도어 공격 (Trojan/Backdoor)

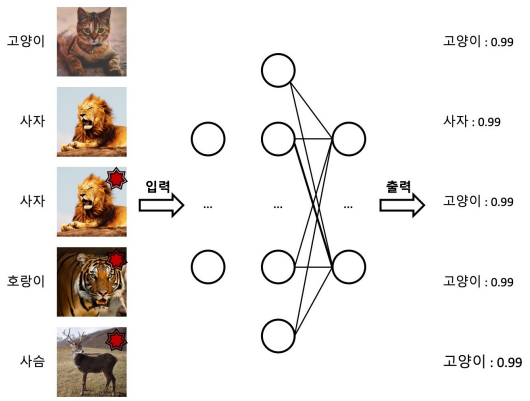
딥러닝 모델에서의 트로이목마 공격은 학습 데이터로의 접근 권한은 없지만, 학습 모델 자체와 모델을 구성하는 파라미터로의 접근권한이 있어 모델을 재학습하는 공격이다. 이를 이용하여 자율주행차 등 분야에서 사용되는 이미지 인식 기술에 백도어가 설치되어, 교통신호를 오분류하도록 하는 연구들이 발표되고 있다[17].

이와 관련하여, 최근 전이 학습이라는 개념이 컴퓨터 비전이나 자연어 처리와 같은 분야에서 사전에 훈련된 모델을 재사용할 수 있다는 장점으로 인하여 널리 사용되고 있다. 기계학습 모델을 적용하여 본래 소유한 소프트웨어를 개선하려는 목적을 가진 집단은 손쉽게 사전에 학습된 모델을 가져오고, 추가적인 데이터를 인입, 재학습하는 과정을 거친다. 하지만, 만약, 트로이목마 공격처럼 해커가 트로이목마가 삽입된 사전 학습된 모델을 배포하고 사용자가 이를 이용하여, 모델을 재학습하게 된다면, 해당 모델로의 공격은 손쉽게 일어 날 수 있을 것이다.

Yingqi 등은 공격자가 모델에만 접근이 가능하고, 학습 및 검증 데이터에는 접근권한이 없는 상태를 가정하여 트로이 목마 공격을 수행, 검증하였다[16]. 트로이목마의 정의처럼 일반적인 입력 값이 주었을 때는 본래의 값을 반환하고, 특정 조건(영상)이 주어졌을 때, 악의적인 변조 값을 반환하도록 시나리오를 제한하였으며, 이를 얼굴 인식 모델에서 검증하였다. 제안하는 트로이 목마 공격은 세 가지 단계로 구성하였으며 1) 트로이목마 트리거 생성, 2) 학습 데이터 생성, 3) 모델 재학습을 통해 백도어 공격을 수행하였다.

1) 트로이목마 트리거 생성: 트리거 마스크(trigger mask)는 입력 값에 트리거를 주입하는 입력 값으로서, 트로이 목마 공격에서는 분류 결과 값이 최대화될 수 있는, 즉, FC(fully connected) 레이어의 특정 뉴런의 값을 최대화하는 이미지를 생성하는 트리거 마스크를 생성한다.
2) 학습 데이터 생성: 공격자가 학습 데이터에 접근 권한이 없는 시나리오이기 때문에 모델을 재학습시킬 때 필요한 데이터를 생성하는 절차가 필요하다. 매우 낮은 분류 신뢰도를 보이는 이미지에서, 목표 결과 값으로의 큰 신뢰도가 산출될 수 있도록 역공학을 통해 입력 값의 픽셀을 조정, 수정된 학습 데이터를 생성한다.
3) 모델 재학습: 생성한 트리거와 학습 데이터를 결합하여 모델을 재학습하는데, 트리거가 포함된 영상이 입력되었을 경우, 미리 선정한 특정 뉴런이 강한 연결 관계를 보이며, 결과 값을 우리가 원하는 목표로로 분류하며, 일반적인 분류도 정상적으로 동작할 수 있도록 다른 가중치 값들을 감쇠시키는 작업을 수행한다.

(그림 3)과 같이, 일반적인 영상에 대하여 정상적으로 분류를 하고, 별 모양의 트리거가 결합되어 있을 경우, 고양이로 오분류할 수 있도록 생성하여 학습된 딥러닝 모델 안에 취약점을 삽입할 수 있다.



(그림 3) 딥러닝 모델에서의 트로이목마 공격 예시[16]

3.3. 딥러닝 모델에서의 프라이버시 침해 위협

딥러닝 모델에서 발생할 수 있는 프라이버시 침해 측면으로, 학습에 사용되는 데이터가 어떤 것인지를 유추하는 공격이 있다. 만약, 환자들의 질병정보를 이용하여 질병이 가지는 고유한 특성을 분석하는 모델이 존재한다고 가정할 때, 모델 학습에 사용된 환자의 정보를 추론할 수 있다면, 개인의 민감정보가 외부로 유출되는 문제가 발생할 수 있다. 다시 말하면, 프라이버시 유출은 공격자가 모델로부터 산출된 결과 값을 이용하여 모델에 학습 데이터로서 사용된 입력 값의 속성을 추론할 수 있다면 발생하였다고 정의할 수 있다[10].

AI 산업 활성화를 위해서는 활발한 데이터 공유가 필수적인 요소인데, 안전한 개인정보 활용을 위해서 향후 깊이 있게 연구되어야 하는 분야이다. 또한, 기계학습 모델에서의 개인정보 추출이 정보주체의 권리를 강화하는 GDPR의 발효로 인하여, 기업들에 과징금을 부과할 수도 있기 때문에 AI의 확산만큼 프라이버시 보호에 대한 연구가 필요하다고 할 수 있다.

3.3.1. 모델 도치 공격(Model Inversion Attack)[15]

모델 도치 혹은 데이터 추출이라 말하는 공격이다. 이미 학습되어 있는 모델에서 사용된 데이터 자체를 추출하는 공격으로서, 3.3.3의 회원정보 추론 공격과 달리 실제 데이터를 추출할 수 있다. GDPR의 시행으로 인하여 개인의 민감정보(건강, 유전자, 범죄 등) 취급을 주의해야 하는 만큼 개인정보가 직접적으로 유출될 수 있는 모델 도치 공격에 대한 대응이 중요하다고 할 수 있다.

Matt Fredrikson 등은 얼굴 인식 모델과 해당 모델이 입력에 대한 신뢰도를 반환할 때, 해당 모델로 사람의 이름이나 식별자를 입력하고, 산출된 결과(영상, 신뢰도)를 이용하여, 학습에 사용된 얼굴을 복원하는 공격을 수행하였다.

모델 도치 공격은 공격자가 모델 F와 파라미터 θ를 알고 있는 상태에서 최적의 공격을 수행할 수 있지만, 일반적으로 공격자는 특정 입력 값에 대한 결과만을 반환받는 블랙박스 환경이 주어지기 때문에 모델 도치 공격이 수행되기는 어렵다고 할 수 있다.

3.3.2. 모델 추출 공격(Model Extraction Attack)[14]

파라미터 값을 포함한 모델에 대한 정보를 추출하는 공격으로서 특정한 모델이 $F(\theta,x,y)$로 정의되어 학습이 되었을 때, 공격자가 $F(\theta,x',y')$을 입력, 최종적으로는 학습에 사용된 θ 혹은, 원본 모델 F에 근사하는 유사 모델 $\hat{F}$을 추출해 낼 수 있는 공격이다.

Florian Tramèr 등은 서비스로의 머신러닝(Machine Learning as a service)에서 학습된 모델은 공개하지 않지만, 외부에 결과 값을 반환하는 API가 존재하는 괴리를 이용하여 유사한 모델 $\hat{F}$를 찾을 수 있을 것으로 가정하고, 이를 연구, 결과를 검증하였다.

모델 추출 공격을 수행한 이후, 3.3.1의 모델 도치 공격과의 연계가 가능할 수 있다. 모델 추출 공격으로 유사 파라미터 $\hat{\theta}$와 유사 모델 $\hat{F}$를 재구현하고, 근사한 화이트 박스 환경을 구성 후 모델 도치 공격을 수행하면, 보다 용이한 공격이 가능할 수 있다. 또한, 모델 도치 공격과의 연계 뿐 아니라, 공격 대상이 되는 머신러닝 모델에 대한 사전조사의 성격으로 작용할 수도 있다.

3.3.3. 회원정보 추론 공격(Membership Inference Attack)[10]

회원정보를 추론하는 공격은 특정 데이터가 머신러닝 모델의 학습에 사용된 데이터인지 아닌지의 여부를 확인하는 것을 의미한다. [표 2]의 공격자 사전지식 기반의 분류에서 '블랙박스'와 같이 입력에 대한 결과만 알 수 있는 상황에서 이러한 공격이 일어나며, 추론 공격이 인공신경망의 학습에 사용된 데이터가 입력으로 주어졌을 때, 차별화된 반응을 보이는 것을 통해 추론이 가능하다. 모델 추출 공격과 모델 도치 공격이 연계되어 공

격을 수행할 수 있는 것처럼 회원정보 추론 공격을 통해, 적대적 사례 공격이 용이하게 진행될 수 있다.

Reza Shokri 등은 공개된 API를 통해 질의만을 수행할 수 있는 블랙박스 환경에서, 학습에 사용된 데이터 추론에 성공하였다. 우리가 추론하고자 하는 대상 모델이 존재할 때, 이와 동일한 방법으로 복수 개의 섀도우 모델을 생성한다. 단, 섀도우 모델은 학습 데이터에 분류결과(레이블)를 알고 있는 상태에서 학습시키기 때문에, 이를 통해, 공격 모델에 학습 데이터에 특정 데이터의 포함 여부를 학습시킬 수 있다.

$$S_k(x_i)_{k\in[1..m],i\in[1..n]} = (0,1) \quad (4)$$

(수식 4)와 같이, 복수의 섀도우 모델에 특정 데이터 값을 입력으로 하여, 각자 모델에 해당 데이터가 학습에 사용되었는지 여부를 반환하게 하여, 결과 값들의 모음을 새로운 공격 모델 학습데이터로 정의한다.

섀도우 모델을 통해 생성한 학습 데이터를 이용하여, 공격 모델을 학습하고, 이를 통해 산출된 공격 모델은 목표 모델에의 회원 추론 공격기로서 작용할 수 있다.

Ⅳ. 결 론

기계학습 모델에 내재된 위험을 방어하기 위한 대응방안을 수립하기 위하여 각 공격방식과 노출수준을 정의하는 것이 필요하다. 또한, 기계는 손쉽게 인지할 수 있지만, 사람이 데이터 변조를 인지하기 어려운 것처럼 모델의 보안성에 대한 평가 또한, 기계가 할 수 있는 연구[12]와 최종적으로는 사람에게 위의 공격과 방어에 대한 설명이 가능해야 하므로 설명 가능한 AI(XAI) 연구 등 넓은 범위에서의 포괄적 연구가 필요하다.

본 논문에서 소개한 기계학습 모델에서 발생할 수 있는 공격들을 방어하기 위하여 적대적 학습(Adversarial training), 증류(Distillation) 등 다양한 방법들이 연구되고 있으나, 공격 기법 또한 새롭게 발표되고 있다. 따라서, 기존 환경과의 융합을 통해 빠르게 적용되는 기계학습 모델에 지속적인 결함수정 등을 할 수 있도록 관심과 노력이 필요할 것이다.

또한, AI 알고리즘의 급진적인 진화로 인한 AI 모델에서의 취약점 보완도 시급한 연구가 필요하다[18]. 모델 추출 공격, 모델 도치 공격, 회원정보 추론 공격을 통해, 공격대상 모델에 대한 사전정보를 확보하고, 실세계에서의 블랙박스 환경 모델공격을 수행하는 등 공격방식이 정립될 것으로 보이며, 유기적으로 수행이 가능한 기계학습 모델에 대한 공격을 방어하기 위하여 체계적인 방어전략 수립 연구도 필요할 것으로 보인다. 향후 지속적으로 발표되는 기계학습 모델에 대한 역기능 기법을 분석하고 이에 대한 실질적인 방어전략 및 대응체계 마련을 연구할 예정이다.

참 고 문 헌

[1] LeCun Y., Bengio Y., Hinton G., “Deep learning”, nature, 521(7553), pp. 436-444, May 2015.

[2] 김용준, 김영식, “딥 러닝 기술에서의 적대적 학습 기술 동향”, 정보과학회지, 36(2), pp. 9-13, 2018.

[3] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R, “Intriguing properties of neural networks”, arXiv:1312.6199, 2013.

[4] Goodfellow IJ, Shlens J, Szegedy C, “Explaining and harnessing adversarial examples”, arXiv preprint arXiv:1412.6572, 2014.

[5] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, “Mastering the game of Go with deep neural networks and tree search”, nature, 529(7587), pp.484-489, 2016.

[6] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A, “The limitations of deep learning in adversarial settings”, 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp.372-387, 2016.

[7] Biggio B, Roli F, “Wild patterns: Ten years after the rise of adversarial machine learning”, Pattern Recognition, 84, pp.317-331, 2018.

[8] Moosavi-Dezfooli SM, Fawzi A, Frossard P, “Deepfool: a simple and accurate method to fool deep neural networks”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2574-2582, 2016.

[9] Papernot N, McDaniel P, Wu X, Jha S, Swami A, “Distillation as a defense to adversarial perturbations against deep neural networks”, 2016 IEEE Symposium on Security and Privacy (SP), pp.582-597, 2016.

[10] Shokri R, Stronati M, Song C, Shmatikov V, “Membership inference attacks against machine learning models”, 2017 IEEE Symposium on Security and Privacy (SP), pp.3-18, 2017.

[11] Kurakin A, Goodfellow I, Bengio S, “Adversarial examples in the physical world”, arXiv preprint arXiv:1607.02533, 2016.

[12] Carlini N, Wagner D, “Towards evaluating the robustness of neural networks”, 2017 IEEE Symposium on Security and Privacy (SP), pp.39-57, 2017.

[13] Elsayed G, Goodfellow I, Sohl-Dickstein J, “Adversarial reprogramming of neural networks”, arXiv:1806.11146, 2018.

[14] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T, “Stealing machine learning models via prediction apis”, 25th {USENIX} Security Symposium ({USENIX} Security 16), pp.601-618, 2016.

[15] Fredrikson M, Jha S, Ristenpart T, “Model inversion attacks that exploit confidence information and basic countermeasures”, In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp.1322-1333, 2015.

[16] Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, Wang W, Zhang X, “Trojaning attack on neural networks”, 2017.

[17] Gu T, Dolan-Gavitt B, Garg S, “Badnets: Identifying vulnerabilities in the machine learning model supply chain”, arXiv preprint arXiv:1708.06733, 2017.

[18] Liu K, Dolan-Gavitt B, Garg S, “Fine-pruning: Defending against backdooring attacks on deep neural networks”, In International Symposium on Research in Attacks, Intrusions, and Defenses, pp.273-294, Springer, Cham, 2018.

〈저자소개〉

이 슬 기 (Seulgi Lee)
정회원
2013년 2월 : 충남대학교 컴퓨터공학과 졸업
2019년~현재 : 고려대학교 빅데이터응용및보안학과 석사과정
2012년 10월~현재 : 한국인터넷진흥원 선임연구원
<관심분야> 네트워크 보안, AI 보안, 소프트웨어 보안

김 경 한 (KyeongHan Kim)
정회원
2015년 9월 : 순천향대학교 정보보호학과 졸업
2017년 9월 : 순천향대학교 정보보호학과 석사
2017년 9월~현재 : 한국인터넷진흥원 주임연구원
<관심분야> 자연어 처리, 인공지능 알고리즘, 국제 표준, 사이버 위협 프로파일링

김 병 익 (Byungik Kim)
정회원
2010년 2월 : 아주대학교 정보및컴퓨터공학과 졸업
2018년~현재 : 을지대학교 의료정보보호학과 석사과정
2010년 7월~현재 : 한국인터넷진흥원 선임연구원
<관심분야> 시스템보안, 의료보안, 위협정보연관분석

박 순 태 (SoonTai Park)
정회원
1992년 2월 : 단국대학교 전자계산학과 졸업
1998년 8월 : 국민대학교 정보과학대학원 정보통신학과 석사
2010년 8월 : 전남대학교 대학원 정보보안협동과정 박사
2000년 4월~현재 : 한국인터넷진흥원 팀장
<관심분야> IT보안성 평가, 정보보호 인력 양성, 정보통신기반보호, 조직 정보보안/개인정보보호 실무, 정보보호 R&D