정 보 보 호 학 회 지
제27권 제3호, 2017. 6

인공지능 보안 이슈

박 소 희*, 최 대 선**

요 약

머신러닝을 위주로 하는 인공지능 기술이 여러 분야에서 다양하게 적용되고 있다. 머신러닝 기술은 시험 데이터에 대해 높은 성능을 보였지만, 악의적으로 만들어진 데이터에 대해서는 오동작을 하는 경우가 보고되고 있다. 그 외에도 학습데이터 오염시키기, 학습된 모델 탈취 등 새로운 공격 유형이 보고되고 있다. 기계학습에 사용된 훈련데이터에 대한 보안과 프라이버시 또한 중요한 이슈이다. 인공지능 기술의 개발 및 적용에 있어 이러한 위험성에 대한 고려와 대비가 반드시 필요하다.

Ⅰ. 서 론

최근에 인공지능을 활용한 보안 기술이 많이 등장하고 있다. 생체인식 등에는 오래 전부터 기계학습기술이 활용되어 왔고, 네트워크 침입 탐지나 악성코드 탐지, 이상거래 탐지 등 패턴인식이 필요한 보안 문제에서도 기계학습 기술을 활용한 연구가 많이 등장하고 있다.

한편 인공지능 기술은 보안 문제 뿐 아니라, 의료, 금융, 가정 등 모든 분야의 서비스에 적용되고 있다. 음성, 이미지, 언어 등 복잡한 형태와 경우의 수를 갖는 데이터들을 높은 성능으로 인식하여 서비스를 제공하게 된다.

서비스가 확산됨에 따라 이러한 인공지능 기술에 보안 문제에 대한 관심이 증가하고 있다. 본 고에서는 주로 기계학습을 중심으로 인공지능 보안이슈를 살펴본다. 우선 Poisoning Attack, Evasion Attack 등 학습된 인공지능을 속일 수 있는 공격 형태를 알아본다. 또한 인공지능 모델 자체를 탈취할 수 있는 Model Extraction Attack을 알아본다.

한편, 학습에 사용되는 데이터에 대한 프라이버시 이슈가 제기되고 있다. 데이터 자체에 대한 보안은 기존에도 많이 이뤄졌지만, 학습된 모델에서 데이터를 추출해내는 Inversion Attack 등은 새롭게 알려진 공격으로 대응 기술이 나와 있지 않다.

Ⅱ장에서는 Adversarial AI란 제목으로 인공지능을 속이는 유형의 공격이슈를 Ⅲ장에서는 인공지능 관련 Data Privacy 이슈를 다룬다.

Ⅱ. Adversarial AI

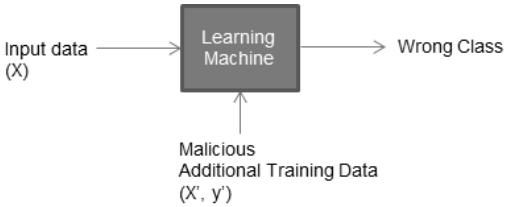
기계학습은 해당 모델에 학습데이터를 통해 학습된 속성을 기반으로 새롭게 들어온 데이터를 정확하게 처리할 수 있는 것을 목표로 한다. 이에는 다양한 알고리즘이 사용되고 있으며 현 시점으로 많은 분야에서 활용되고 있다.

하지만 이는 악의적인 공격자에 의해 학습 알고리즘의 특정 취약점을 악용하는 데이터를 통하여 전체 시스템의 보안을 손상 시킬 수 있다. 이에 대한 안전한 알고리즘이 필요하며 공격에 대응하기 위해서는 어떠한 공격들이 가능한 지 알아볼 필요가 있다. 이 장에서는 최근에 알려진 기계학습에 대한 공격 유형인 Poisoning Attack과 Evasion Attack 그리고 Model Extraction Attack에 대하여 설명한다.

2.1. Poisoning Attack

[그림 1]에 보인 바와 같이, Poisoning Attack 은 잘못된 데이터를 제공하여 이를 학습한 classifier 가 잘

이 논문은 2017년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(B0717-16-0139, 핀테크 서비스 금융사기 방지를 위한 비대면 본인확인 및 이상거래필터링기술)
* 공주대학교 응용수학과 (soheeee0803@kongju.ac.kr)
** 공주대학교 의료정보학과, 교신저자 (sunchoi@kongju.ac.kr)



〔그림 1〕 Poisoning Attack

못된 결과를 내도록 하는 것으로 최소한의 데이터로 최대한의 오동작을 일으키는 것이 목표이다.

이 방법은 공격자가 학습 알고리즘을 알고 있고 원래의 데이터에 접근 가능하여 시뮬레이션이 가능하다는 점을 가정하고 있다[1].

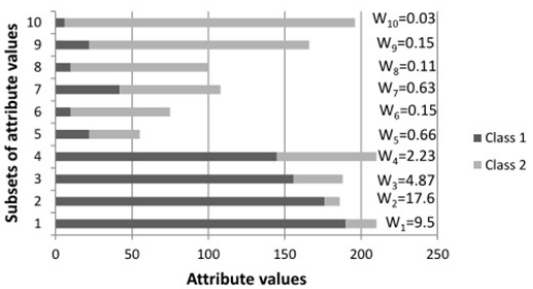
2.1.1. Poison Data 생성

Poison Data는 학습 모델의 잘못 된 결과를 이끌기 위하여 학습시키는 악의적인 데이터 X'를 말한다.

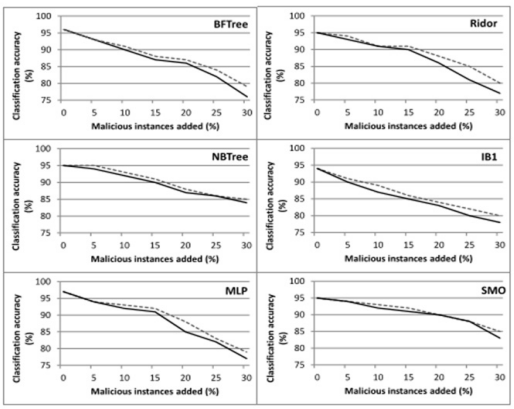
어떤 데이터를 학습시켜야 최소한의 Poison으로 최대의 효과를 거둘 수 있는지가 문제인데, [2]에서는 gradient ascent 방법을 통해 모델의 Loss function을 극대화 하는 값 X'를 구하는 방식을 제안했다. 구해진 X'는 Poison data로서 기존 데이터 X와 함께 모델에 학습시킨다. 모델은 Poison Data의 영향으로 오동작을 하게 된다.

다른 방법으로는 위의 〔그림 2〕와 같이 각 class의 속성 값 분포 분석을 한 후 class1의 비율이 높은 분포로 학습데이터를 생성한다. 생성된 데이터는 class1이 아닌 class2로 학습시키는 것이다.

[그림 3]은 6개의 기계학습 알고리즘에 대한 Poisoning Attack 결과이다. 그래프의 실선은 공격기간 동안 학습데이터의 수가 고정인 경우이며 점선은



〔그림 2〕 class 별 속성 값 분포 분석〔2〕



〔그림 3〕 Poisoning Attack 결과

공격 중에 학습데이터에 정확한 학습 데이터를 추가할 수 있는 경우이다. 그래프를 보면 Poison Data가 추가될수록 정확도가 떨어지는 것을 확인할 수 있다[2].

2.2. Evasion Attack

Evasion Attack은 입력 데이터에 대하여 최소한의 변조를 통해 다른 Class 로 인식되게 하는 것이다.

앞 서 2.1절에서 설명한 Poisoning Attack은 악의적인 데이터를 기존의 학습데이터에 추가하여 모델에 학습시킴으로써 잘못된 결과를 내도록 하였지만 Evasion Attack은 학습데이터가 아닌 입력 데이터를 약간 변조하여 다른 class로 인식되도록 하는 공격이다. Poisoning Attack은 학습단계에서 공격자가 학습데이터를 조작할 수 있는 환경이나, 지속적으로 학습하는 active learning 시스템을 가정하고 있으므로 공격의 적용 범위가 떨어지지만, Evasion Attack은 학습된 이후 적용단계의 입력 데이터에 대한 조작을 통해 인공지능을 속이는 형태의 공격이므로 적용범위가 훨씬 넓다고 할 수 있다. Evastion Attack은 현재 이미지와 보이스에 대해서 보고되고 있는데 향후 다양한 데이터 형태로 확대될 것으로 예상된다.

2.2.1. Evasion Attack – Image

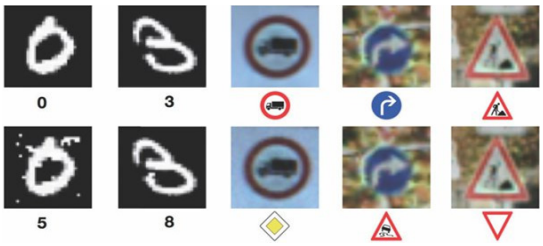
[그림 4]는 원본이미지와 변조과정 그리고 변조된 이미지를 보여준다. 변조하기 전과 변조 후의 차이를 사람은 거의 알 수가 없다. 하지만 실제로 컴퓨터는 이

〔그림 4〕 Evasion Attack - Image〔3〕

를 모두 타조로 인식한다. [3]의 실험에서 원본 이미지의 4%만 변조를 하여도 97%는 잘못 분류하는 결과를 얻었다.

이를 주목해야 하는 이유는 Evasion Attack 이 성공하였을 시 매우 큰 위험이 따르기 때문이다. 아래와 같이 실생활에서 볼 수 있는 숫자, 표지판에 대하여 생각해보면 사람의 눈으로는 거의 차이점을 발견 하지 못할 정도로 약간의 변조지만 classifier의 경우 전혀 본래와는 다른 모양으로 인식할 수 있기 때문에 위험 할 수 있다. [그림5]는 Evasion Attack을 통한 오 인식 결과를 보여준다. 상단이 원래 이미지와 이를 인식한 결과이고 하단은 원래 이미지에 약간의 노이즈를 추가하여 오인식된 결과이다. 만약 공격자가 도로에 잇는 좌회전 표지판에 사람이 알아볼 수 없는 정도로 약간의 변조를 가한 뒤 자율 주행차량이 이를 우회전 표지판으로 오 인식하여 우회전을 한다면 사고로 이어 질 수 있다.

이처럼 기계학습이 주목을 받고 많이 활용되는 이 시점에서 Evasion Attack은 치명적일 수 있다. 그러므로 이에 대한 연구도 활발히 이루어져야 한다.

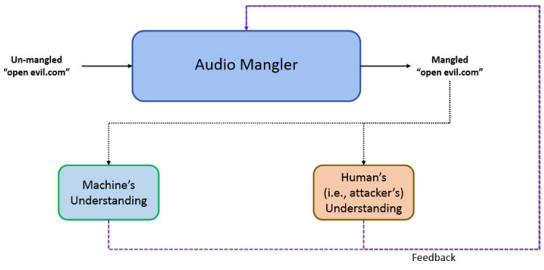


〔그림 5〕 Evasion Attack으로 인한 오 인식

2.2.2. Evasion Attack - Voice

이미지뿐만 아니라 음성 데이터에도 Evasion Attack을 할 수 있다. 휴대전화의 음성인식 기능은 휴대전화의 사용자의 목소리를 인식하여 검색이나 음악 재생과 같은 명령을 수행할 수 있다. 음성 명령에 대한 Evasion Attack은 이미지의 경우와는 반대로 변조한 음성을 사용자는 알아들을 수 없지만 음성인식 시스템에서는 올바른 명령으로 인식되도록 하는 것을 목표로 한다. 이 공격이 성공하게 되면 사용자가 알아들을 수 없는 소리에 의해 사용자 의도와 무관하게 휴대전화의 특정 기능이 작동할 것이다. 단순한 오작동 뿐 만아니라 공격자에 의해 충분히 악용 될 수 있다[3].

[그림 6]은 기존의 Voice에 대하여 Audio Mangler를 통해 변조를 하여 음성 인식 시스템은 이해할 수 있는 충분한 음향 특성을 가져 실제로 시스템을 실행할 수 있지만 사용자는 이해할 수 없도록 변조하는 원리를 보여준다. Audio Mangler는 음성 명령의 MFCC 특징을 추출한 다음, MFCC 벡터를 변조 후 역MFCC를 적용하여 변조된 음성을 생성할 수 있게 된다[4]. 이때 변조의 방향은 음성인식 시스템은 인식하고 사람은 인식할 수 없는 것을 목표로 한다. [4]에 따르면 구글의 음성인식 시스템인 ok google에 시작 명령인 “ok google”을 변조하여 사람들은 “cocaine nuddle”로 인식한 음성 명령을 통해 ok google 시스템을 동작시키는데 성공했다고 한다.



〔그림 6〕 Evasion Attack을 위한 공격 데이터 생성〔4〕

2.3. Model Extraction Attack

2.3.1. Machine learning Model

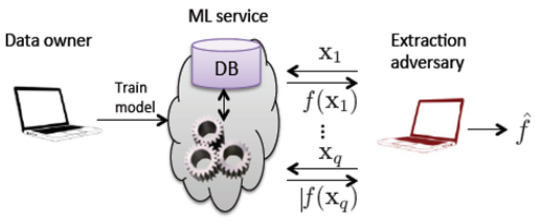
기계학습은 적절한 모델에 (X,y)로 이루어진 학습

데이터를 학습시킨 후 모델에게 새로운 X를 입력하였을 때 그에 해당하는 y를 출력하는 시스템을 만드는 것이다. 현재 많은 기업에서 기계학습을 이용한 많은 서비스를 제공하고 있는데 여기에 사용되는 학습데이터를 구축하고 시스템을 학습시키는데 많은 비용이 소요된다. 따라서 구축된 모델은 기업의 귀중한 자산이다. 기업들은 구축된 모델을 이용해 이익을 창출하기 위해서 많은 경우 모델을 서비스 형태로 제공해야 한다. 즉 이용자가 제출한 X에 대한 y값을 제공하는 형태의 서비스를 공개해야 한다. 이때 모델의 내부는 공개되지 않는다.

2.3.2. Model Extraction Attack

Model Extraction Attack은 공개된 모델에 X를 입력하여 y를 획득한 뒤 이를 바탕으로 모델을 모사하는 공격을 의미한다. 모델 모사는 고비용으로 구축한 모델을 쉽게 탈취할 수도 있으며, 모사된 모델을 이용해 Poisoning attack, Evasion attack, Inversion attack을 용이하게 할 수도 있다.

[그림 7]은 Model Extraction Attack의 구조를 보여준다. 공격자는 반복적으로 쿼리x를 입력하고 이에 대한 답인 f(x)를 획득하여 이를 통해 사용자에게 제공된 모델 f 에 가까운 f^를 만드는 것이다. 이때 얼마나 작은 수의 질의(입력)을 통해 모델을 100%에 가깝게 모사하는가가 기술적 목표가 된다. f(x)가 수치가 아니라 class 인 모델에서는 f^ 를 만들기 위해 confidence value를 활용한다. 모델은 class와 함께 confidence value를 제공하는 경우가 많은데class정보와 함께 confidence value를 활용하면 모델 내부의 parameter를 추정하여 f^를 만드는 것이 용이해진다.



〔그림 7〕 Model Extraction Attack〔5〕

2.3.3. Model Extraction Attack's Result

[그림 8]은 Model Extraction Attack을 통해 서비스의 모델을 100% 흉내 내는데 소요된 쿼리 및 시간을 나타낸다. [5]에 따르면 가장 적게는 650번의 쿼리를 통해 MLaS (Machine Learning as Service)를 제공하는 Amazon에 공개된 Logistic Regression 모델을 흉내 내는데 70초 밖에 걸리지 않았다.

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

〔그림 8〕 모델을 100% 흉내 내는데 소요된 쿼리 및 시간〔5〕

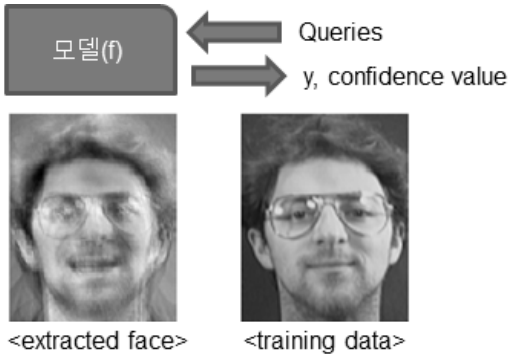
Ⅲ. Data Privacy in AI

Ⅱ장에서는 주로 학습 알고리즘의 취약점을 악용하여 인공지능 시스템을 속이는 공격에 대해 알아보았다. Ⅲ 장에서는 학습에 사용되는 데이터에 대한 보안을 다룬다. 모델학습에 사용된 데이터를 추출하는 Inversion Attack과 머신 러닝을 이용한 Data Sanitizing의 취약점을 설명한다.

3.1. Inversion Attack

모델에 쿼리를 하여 학습데이터를 재현해내는 공격을 Inversion Attack이라고 한다. 특정 X를 알기 위해서 쿼리를 통해 여러 개의 X의 값을 쿼리 하여 얻은 y값과 confidence value를 통하여 confidence 값을 최대화하는 입력을 찾는 방식으로 모델에 학습된 X를 획득할 수 있다. 아래의 [그림 9] 는 모델에 Inversion Attack 하여 재현해낸 학습데이터의 예시이다.

또 다른 형태의 Inversion Attack으로, 다른 feature들과 y를 알 때 모델로부터 알지 못하는 민감한 정보 feature x_1 에 대하여 알아낼 수 가 있다. 공격자가 알고 싶은 feature 의 모든 값을 시험하여 y 예측 오차가 가장 작은 값을 선택하면 된다. 예를 들어, 공격자가 X라는 사람의 나이(34세), 성별(여), 혈압(120/80)을 알고 있고 혈액형(?)을 모른다고 하자. 약물 용량 모델에 대하여 X라는 사람에 대한 y값이 하루 20ml 라면 혈



〔그림 9〕 Inversion Attack을 통한 학습데이터 재현〔6〕

액형의 모든 경우의 수를 시험하여 예측 오차가 가장 작은 혈액형을 선택하면 된다.

3.2. Attack on Data Sanitizing

3.2.1. Data Sanitizing

시스템에 사용되는 이미지, 문서와 같은 데이터에서 얼굴이나 자동차 번호판이나 주민번호, 이름과 같은 개인 정보 등 비정형 데이터의 경우 그대로 사용되지 않고 제거하거나 암호화 하는 등의 마스킹 작업을 거친다. 이러한 작업을 Data Sanitizing 이라고 한다[7]. 하지만 이 작업으로 인해 모든 정보가 안전해진다고 볼 수 없다. Sanitizing 시 탐지 되지 않아 마스킹 되지 않은 정보의 경우 노출될 위험이 있다.

3.2.2. Attack on Data Sanitizing

Data Sanitizing에 대한 공격은 이렇게 노출된 정보를 자동으로 찾아내는 것을 의미한다. [그림10]과 같이 비정형 데이터에 대하여 Sanitizing을 하는 경우 일부 정보는 탐지되지 않다. 이러한 경우 공격자는 노출이 있는 데이터를 학습 데이터로 활용하여 이러한 미탐 사례만을 탐지하는 모델을 만들 수가 있다. 이렇게 학습된 모델을 활용하여 Data Sanitizing의 미탐 데이터를 확보할 수 있다.

〔그림 10〕 Sanitizing 전(왼쪽)과 후(오른쪽)

Ⅳ. 결 론

인공지능에 대한 관심이 급증하면서 많은 분야에서 인공지능 기술이 활용되고 있다. 따라서 인공지능 기술 기반의 시스템을 손상시키기 위한 악의적인 공격은 매우 치명적일 수 있다. 이를 대응하기 위하여 인공지능 기술을 위협할 수 있는 공격 이슈에 대해 주목 할 필요가 있다. 또한, 학습된 모델에서 데이터를 추출해내는 Inversion Attack 등은 새롭게 알려진 공격으로서 대응 기술이 나와 있지 않기 때문에 이에 대한 적극적인 연구가 필요하다.

참 고 문 헌

[1] “Poison attacks against machine learning, Security and spam-detection programs could be affected”, The Kurzweil Accelerating Intelligence , July, 2012

[2] Mozaffari-Kermani, Mehran, et al. “Systematic poisoning attacks on and defenses for machine learning in healthcare.” IEEE journal of biomedical and health informatics, 19.6, 1893-1905, 2015

[3] Szegedy, Christian, et al. “Intriguing properties of neural networks.” arXiv preprint arXiv, 1312.6199, 2013.

[4] T. Vaidya, Y. Zhang, M. Sherr, and C. Shields, “Cocaine noodles:exploiting the gap between human and machine speech recognition,” in 9th USENIX Workshop on Offensive Technologies (WOOT 15), 2015

[5] Tramèr, Florian, et al. "Stealing machine learning models via prediction apis." USENIX Security. 2016.

[6] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic

countermeasures.” Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015.

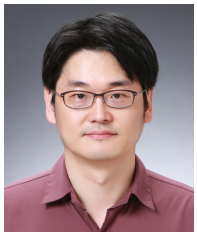
[7] https://en.wikipedia.org/wiki/Sanitization_(classified_information)

〈저자소개〉

박 소 희 (Sohee Park)

2014년 3월~현재 : 공주대학교 응용수학과 재학

관심분야 : 인공지능, 정보보호, 머신러닝



최 대 선 (Daeseon Choi)

종신회원

1995년 2월 : 동국대학교 컴퓨터공학과 학사

1997년 2월 : 포항공과대학교 컴퓨터공학과 석사

2009년 1월 : 한국과학기술원 전산학과 박사

1997년 1월~1999년 6월 : 현대정보기술 선임

1999년 7월~2015년 8월 : 한국전자통신연구원 인증기술연구실 실장/책임연구원

2015년 9월~현재 : 공주대학교 의료정보학과 부교수

2017년 현재 : 정보보호학회 이사

관심분야 : 인증, 개인정보보호, 이상거래탐지, 의료정보보안, 머신러닝