적대적 AI 공격 및 방어 기법 연구

문현정[1], 오규태[2], 유은성[3], 임정윤[3], 신진영[4], 이규영[5]
[1]동덕여자대학교 컴퓨터학과
[2]안양대학교 컴퓨터공학전공
[3]호남대학교 컴퓨터공학
[4]숙명여자대학교 소프트웨어융합전공
[5]한국과학기술원 정보보호대학원
s205044s@gmail.com, einokt@naver.com, jd962q@gmail.com, roralim0410@gmail.com,
iwbm312@gmail.com, leeahn1223@kaist.ac.kr

A Study on Adversarial AI Attack and Defense Techniques

Hyun-Jeong Mun[1], Gyu-Tae Oh[2], Eun-Seong Yu[3], Jeong-yoon Lm[3],
Jin-Young Shin[4], Gyu-Young Lee[5]
[1]Department of Computer Science, Dongduk Women's University
[2]Department of Computer Engineering, Anyang University
[3]Department of Computer Engineering, Honam University
[4]Dept. of Software Convergence, Sookmyung Women's University
[5]Graduate School of Information Security, KAIST

요 약

최근 인공지능 기술이 급격하게 발전하고 빠르게 보급되면서, 머신러닝 시스템을 대상으로 한 다양한 공격들이 등장하기 시작하였다. 인공지능은 많은 강점이 있지만 인위적인 조작에 취약할 수 있기 때문에, 그만큼 이전에는 존재하지 않았던 새로운 위험을 내포하고 있다고 볼 수 있다. 본 논문에서는 데이터 유형 별 적대적 공격 샘플을 직접 제작하고 이에 대한 효과적인 방어법을 구현하였다. 영상 및 텍스트 데이터를 기반으로 한 적대적 샘플공격을 방어하기 위해 적대적 훈련기법을 적용하였고, 그 결과 공격에 대한 면역능력이 형성된 것을 확인하였다.

1. 서론

오늘날 우리는 머신러닝 기술을 통해 데이터를 분석하여 정확한 예측을 수행할 수 있게 되었다. 머신러닝은 인공지능을 구현하기 위한 핵심 기술이고 다양한 분야에서 중요한 기능을 담당하기 때문에, 향후 머신러닝 시스템에 적대적 공격에 의한 해킹 피해가 발생하면 사용자의 안전이 위협 당하고 개인정보가 유출되는 등 심각한 문제가 발생할 수 있다. 이에 따라 인공지능을 보호할 수 있는 보안기술이 필요하게 되었다. 이러한 보안기술을 개발하기 위하여 본 연구에서는 적대적 AI 공격을 방어하기 위해 기존에 제시된 보안기법들과 데이터 유형별로 관련된 연구내용을 소개한다. 그리고 각 데이터 유형별로 적대적 예제를 생성하고 머신러닝 모델을 구축하여 이를 훈련시킨 후, 유형별로 적대적 공격을 수행하고 적대적 훈련 방어기법 적용 시 결과를 분석하여 적대적 훈련에 대한 방어 성능을 측정한다.

2. 관련 연구

2.1 적대적 공격 기술

적대적 공격은 AI시스템을 교란할 목적으로 머신러닝 모델에 Advesarial Pertubation을 적용하여 오분류를 발생시키는 것을 의미하며, 크게 아래의 3가지 형태로 분류한다.[1]

2.1.1 회피공격(Evasion attack)

일반적으로 회피공격은 머신러닝 시스템이 이미 훈련된 후 새로운 입력값에 대한 확률을 계산할 때 적대적 예제를 사용해서 AI가 잘못된 의사결정을 하도록 하는 공격이다.[1]

2.1.2 중독공격(Poisoning attack)

중독공격은 AI 시스템이 생성되는 과정 자체가 손상되는 공격이다.[1] 대표적으로 모델 훈련 과정에 사용되는 dataset을 손상시키는 공격 방법이 있

다.[1] 중독공격은 손상된 입력값을 주입하여 모델의 학습 과정을 손상시킨다는 점에서 회피공격과 다르다.[1]

2.1.3 탐색공격(Exploratory attack)

탐색적 공격은 training dataset에 영향을 미치지 않는다.[1] 탐색적 공격의 대표적인 예로 AI 모델의 학습에 사용된 데이터를 추출하는 공격 기법인 Model Inversion Attack과 공개된 API가 있는 학습 모델의 정보를 추출하는 공격 기법인 Model Extraction via APIs가 있다.[1]

2.2. 적대적 방어 기술

적대적 방어기술로 아래의 2가지 기술을 소개한다.

2.2.1 적대적 훈련(Adversarial Training)

모델을 학습시킬 때 적대적 공격에 대응하기 위해 기존 학습 데이터셋에 적대적 샘플들을 추가로 학습시켜 적대적 공격에 대해 강인하게 만드는 것이 적대적 훈련(Adversarial Training) 기술이다. 전반적으로 적대적 공격에 대해 가장 효과적이고, 본 적 없던 형태의 적대적 샘플이 나타나도 그 부분만 추가적으로 학습시켜주면 되는 강점이 있다.

2.2.2 Defense-GAN

GAN은 생성자와 구분자라는 서로 대립하는 두 가지 모델의 경쟁을 통해 데이터를 생성하는 알고리즘이다.[3] 이런 GAN을 활용한 Defense-GAN 방어기법은 생성 이미지와 적대적 샘플의 차이를 최소화하는 새로운 생성 데이터를 만드는 기술이다.[3] 정상적인 이미지와 가장 가깝게 생성된 이미지와 적대적 샘플 간의 차이를 최소화하도록 새롭게 생성된 이미지 데이터는 정상적인 이미지에 가까워질 것이고, GAN 생성과정을 통해 적대적 샘플에 포함된 노이즈는 상당 부분 완화될 것이다.[3] 아울러 Defense-GAN은 어떠한 공격모델을 가정한 것이 아니기 때문에, 특정 공격에 국한되지 않는 장점이 있다.[3]

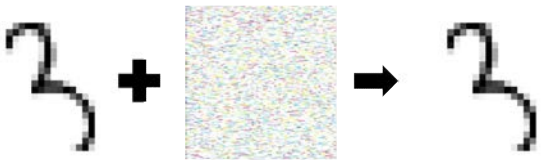
3. 제안 이론

3.1 적대적 텍스트 공격

텍스트 공격은 Goal Function, Search method, Transformation, Constraints 이 4가지 부분으로 구성된다.[4] Goal Function은 공격의 성공 여부를 결정한다.[4] Search method는 잠재적 변환의 공간을 탐색하고 성공적인 섭동을 찾으려고 한다.[4] Transformation은 텍스트 입력을 받아 의미를 변경하지 않으려고 노력하면서 단어나 구를 유사한 것으로 바꾸는 것과 같이 변환한다.[4] Constraints는 주어진 변환이 유효한 지 여부를 결정한다.[4] 단어를 특정 단어로 변환하는 방법을 사용하며 데이터 셋에서 뉴스 분류를 위해 훈련된 BERT 알고리즘을 사용한다.[4]

3.2 적대적 이미지 공격

이미지에 대한 적대적 AI 공격은 입력 이미지에 육안으로는 구분할 수 없는 크기의 노이즈를 추가해 딥러닝 모델이 잘못된 결과를 산출하도록 하는 공격이다.[5] 인간은 두 데이터의 차이를 알 수 없지만, AI는 적대적 AI 공격으로 인해서 두 데이터를 서로 다른 분류로 인식한다.[5] CNN에 기반을 둔 분류기에게 많은 숫자 사진을 보여주고, 숫자를 인식하도록 훈련시킨다.[5] 분류기가 훈련 사진에서 숫자와 유사한 특징을 식별할 수 있게 되면 새로운 사진에서 숫자를 확실하게 인식할 수 있게 된다.[5] 적대적 이미지는 소음처럼 보이는 정밀하게 계산된 입력(일반적으로 "노이즈"라고 불림)을 통해 변경된 이미지로 인간에게는 거의 같지만, 분류기에게는 완전히 다르게 보이게 하고 이를 식별할 때 데이터를 오분류하게 한다.[5] 분류기가 아래의 왼쪽 사진을 '3'이라는 숫자라고 예측하고있다. 그러나 노이즈를 추가하여 이미지를 약간 바꾸면, 분류기는 '8'라는 숫자라고 예측하게 된다.[5]



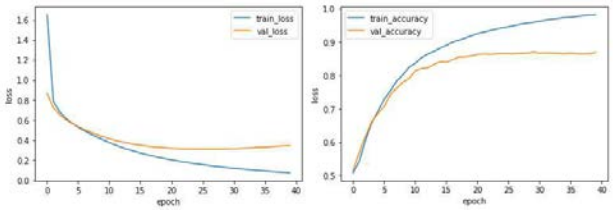
(그림 1) 적대적 이미지의 생성 (노이즈 추가) [5]

4. 실험

4.1 텍스트 공격에 대한 적대적 훈련 방어

100개의 샘플로 공격실험을 수행한 결과, 공격 성공률은 90.72%이며 88개의 공격이 성공되었고 9개의 공격이 실패했다. <표1>은 텍스트에 대한 적대적 예제와 일반 데이터를 통해 예측의 정확도를 측정한 도표이다. 적대적 훈련 이전에는 적대적 예제

에 대한 예측의 정확도가 61.79%인 반면, 적대적 훈련 후에는 적대적 예제에 대한 예측의 정확도가 96.88%까지 상승한 것을 볼 수 있다.



(그림 2) Train & Vol And Loss & Accuracy

<표 1> 텍스트 공격에 대한 실험결과

No	Title	Prediction Accuracy	
		(Normal)	(Adversarial)
1	Normal Training	97.6%	61.79%
2	Adversarial Training	96.45%	96.88%

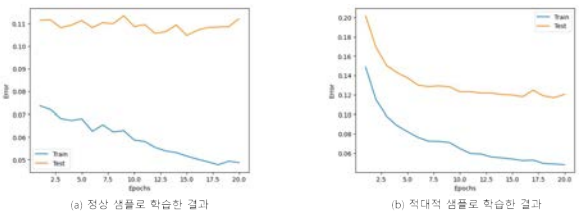
4.2 적대적 이미지 공격에 대한 적대적 훈련 방어

<표 2>는 이미지에 대한 적대적 예제와 일반 데이터를 통해 예측의 정확도를 측정한 도표이다. 적대적 훈련 전에는 적대적 예제에 대한 예측의 정확도가 61.79%인 반면, 적대적 훈련 이후에는 적대적 예제에 대한 예측의 정확도가 96.64%까지 향상된 것을 알 수 있다.

<표 2> 이미지 공격에 대한 실험결과

No	Title	Prediction Accuracy	
		(Normal)	(Adversarial)
1	Normal Training	96.6%	61.81%
2	Adversarial Training	96.62%	96.64%

아래의 그래프는 정상 샘플로만 학습했을 때의 모델의 오차 그래프와 적대적 샘플까지 학습했을 때의 모델의 오차 그래프를 비교한 것이다.



(그림 3) 오차 그래프 상호 비교

5. 결론

본 논문에서는 데이터 유형 별 적대적 예제의 공격과 방어기법에 대해 살펴보았다. 그리고 텍스트 및 이미지에 대하여 적대적 훈련 방어기법을 적용하기 전과 후의 공격 성공률에 대해서 비교하였다. 실험을 통해 데이터 유형 별로 적대적 공격이 적용된 상태에서 적대적 훈련 기반의 방어기법을 사용하면, 분류 인식정확도가 높아져 공격에 대한 면역력을 보유하게 되는 것을 확인하였다. 이를 바탕으로 기존의 기법들을 보완한 새로운 방어기법과 더 높은 성능향상을 위한 연구를 향후 과제로 제안하는 바이다.

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] Chakraborty, Anirban, et al. "Adversarial attacks and defences: A survey.", 2018

[2] Tao Bai, Jinqi Luo, JunZhao, Bihan Wen, Qian Wang. “Recent Advances in Adversarial Training for Adversarial Robustness”. International Joint Conference on Artificial Intelligence (IJCAI-21)

[3] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. “Defense-gan: Protecting classifiers against adversarial attacks using generative models”. CoRR, abs/1805.06605, 2018.

[4] “TextAttack Documentation”. https://textattack.readthedocs.io

[5] Evan Ackerman. "Hacking the Brain With Adversarial Images : Researchers from Google Brain show that adversarial images can trick both humans and computers, and the implications are scary" IEEE Spectrum. 2018, Feb 28. https://spectrum.ieee.org/hacking-the-brain-with-adversarial-images.