적대적 공격의 방어를 위한 AI 백신 연구

송채원[1], 오승아[1], 정다예[2], 임유리[3], 노은지[4], 이규영[5]
[1]순천향대학교 사물인터넷학과, [2]전북대학교 IT 정보공학과, [3]가천대학교 컴퓨터공학과, [4]경상대학교 물리학과, [5]리안기술사사무소
9cwsong@gmail.com, osa0215@naver.com, beauti4612@gmail.com, dbfl5073@naver.com, jully1052@gnu.ac.kr, leeahn1223@nate.com

A Study on AI Vaccine for the Defense against Adversarial Attack

Chae-Won Song[1], Seung-A Oh[1], Da-Yae Jeong[2], Yuri Lim[3], Eun-Ji Rho[4], Gyu-Young Lee[5]
[1]Dept. Internet of Things, Soonchunhyang University, [2]Dept. engineering of information technology, Jeonbuk National University, [3]Dept. Computer Engineering, Gachon University, [4]Dept. of Physics, Gyeongsang National University, [5]Lee-Ahn Professional Engineer's Office

요 약

본 논문에서는 머신러닝 시스템에 심각한 오류를 발생시킬 수 있는 적대적 샘플을 제작하고, 이를 이용한 적대적 공격을 효과적으로 예방하고 방어할 수 있는 Adversarial Training 기반의 AI 백신을 개발하였으며, 본 논문이 제안하는 AI 백신이 적대적 샘플을 올바르게 인식하고 AI 공격 성공율을 현저하게 낮추는 등 강인성을 확보한 것을 실험을 통해 입증하였다. 아울러 스마트폰을 통해 수행결과를 확인할 수 있는 어플리케이션을 구현하여, 교육 및 시연 등을 통해 적대적 AI 공격에 대한 심각성을 인식하고 해당 방어과정을 명확히 이해할 수 있도록 하였다.

1. 서론

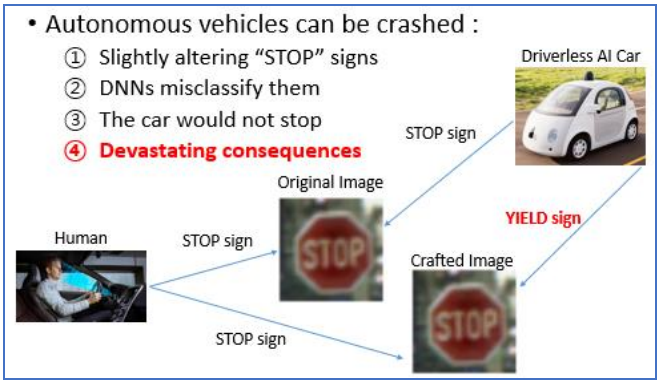
딥러닝의 성공으로 전체 산업군에 인공지능 기술이 빠르게 보급되고 있고, 이제 AI 는 인류의 미래를 책임지고 결정할 기술로서 의심의 여지가 없어 보인다.

하지만 인공지능에는 사용자의 안전을 강력하게 위협하는 보안적 약점이 존재한다. 그리고 이에 대한 해결책은 아직도 명확하게 마련되어 있지 않은 상태이다. 따라서 생산 현장의 모든 AI 자산에서 적대적 공격의 가능성을 가정하고 대비하는 것은 이제 필수적이다.

한편, 인공지능을 탑재한 자율 주행 자동차가 실용화 단계에 있으나, 적대적공격의 위험 속에 노출되어 있다.

그림 1 과 같이 자율 주행 중인 차량이 적대적 AI 공격을 받아 신호등이나 표지판 등을 잘못 판독하여 보행자와 충돌하는 등 심각한 사고를 일으킬 수 있다.

이러한 적대적 AI 공격에 대해 아직까지 명확한 해결책을 제시하지 못하고 있는 만큼, 본 연구에서는 적대적 AI 공격기법 및 방어기법을 분석하고 구현하여, 전체 산업군에 공통적으로 적용할 수 있는 효과적인 딥러닝 보안모델을 개발하여 제시하고자 한다.

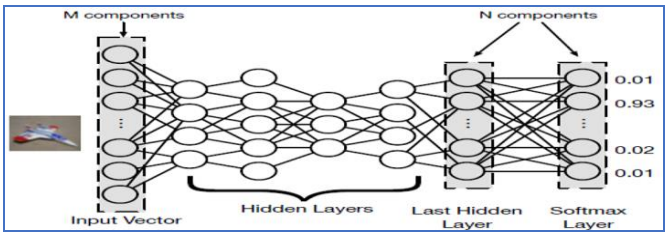


(그림 1) 자율주행자동차에 대한 적대적샘플 공격 구성도

2. 관련 연구

1) 인공신경망 머신러닝 시스템

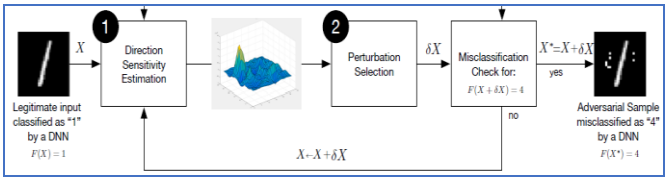
머신러닝은 수많은 가중치와 편향값으로 이루어진 다층의 인공신경망 모델을 구축하고, 준비한 학습데이터를 모델에 통과시켜 예측한 값과 정답과의 오차(손실)를 오류역전파기법을 사용하여 출력층에서 입력층 방향으로 분산 반영해 나가되 경사하강법을 통해 손실함수를 최소화하는 가중치와 편향값으로 최적화시키는 기술이다.



(그림 2) 인공신경망 머신러닝시스템 네트워크 구성도[1]

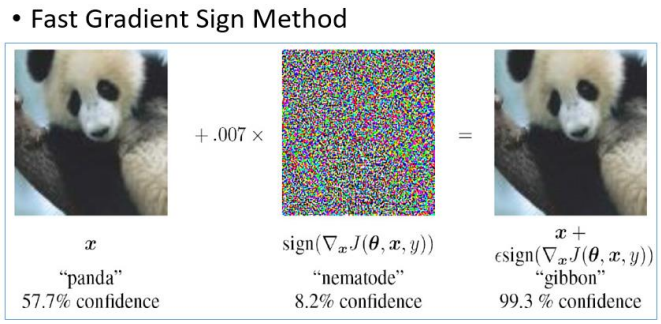
2) 적대적샘플 공격기법

적대적샘플(Adversarial Sample)은 머신러닝 시스템을 교란할 목적으로 정상적인 입력 이미지에 미리 계산된 변형을 의도적으로 삽입한 이미지를 말한다. 인간의 눈에는 감지되지 않고 머신러닝시스템에서는 판단오류를 일으키도록 하되 가급적 최소한의 변형만 가해져야 한다.



(그림 3) 적대적샘플 제작과정[1]

공격에는 크게 회피(Evasion), 독성(Poisoning), 모델탈취(Model stealing) 방법 등이 있으나, 학습을 완료한 인공신경망에 판별오류를 유발하도록 적대적샘플을 의도적으로 제작하여 공격하는 회피공격이 보다 현실적이다.



(그림 4) FGSM 적대적샘플 제작사례[2]

적대적샘플 제작기법에는 FGSM(Fast Gradient Sign Method), PGD(Projected Gradient Descent), C&W(Carlini and Wagner), Adversarial patch 등 다양한 기법이 있으나, 본 논문에서는 가장 제작효율이 우수한 FGSM 방법을 사용하였다. 이러한 적대적샘플 제작 시에 오류역전파 알고리즘에서 사용하였던 경사하강법을 차용하는데, 적대적 목표를 달성하기 위한 최소한의 변화만을 투입하여 사람의 눈으로 정상샘플과 외형적인 차이를 느낄 수 없도록 제작한다.

3) 적대적공격 방어기술

적대적공격에 대한 방어기술로는 Input Data Filtering, Adversarial Training, Defensive Distillation 등 다양한 기법이 있으나, 적대적샘플을 학습에 투입하여 적대적샘플에 대한 식별능력을 부여하는 Adversarial Training 이 가장 근본적이고 원리에 충실한 방어기법으로 볼 수 있다. 즉 외형이 7 이고 머신러닝시스템이 3 으로 판정하도록 하는 적대적샘플을 만든 후, 이 샘플에 대한 정답을 7 로 강제로 지정한 후 추가로 학습에 투입하는 것이다

3. 실험 및 구현 결과

1) 실험환경

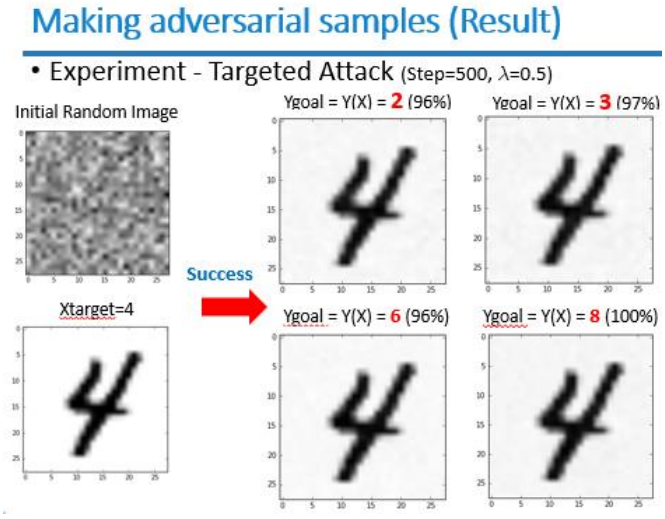
개발도구는 Tensorflow, Keras, Pytorch 같은 머신러닝 라이브러리를 사용하지 않고 파이썬으로 오류역전파, 경사하강법 등 신경망 구성요소를 직접 개발하여 사용하였다.

0 에서 9 까지 10 개의 숫자세트가 모두 70,000 개 준비되어 있는 MNIST handwriting 데이터셋을 사용하였으며, 이중 60,000 개를 학습에 투입하고 10,000 개는 테스트 전용으로만 사용하였다.

개발환경으로는 Jupyter Notebook 을 사용하였고, Intel i5 CPU 탑재 노트북(윈도우 10)에서 수행하였다.

2) 적대적샘플 제작실험 결과

FGSM 기법을 파이썬 언어로 구현한 결과, 0~9 까지 10 개의 어떠한 숫자에 대해서도 머신러닝시스템이 잘못된 예측결과를 산출하게 만드는 적대적 샘플을 성공적으로 제작할 수 있었다.



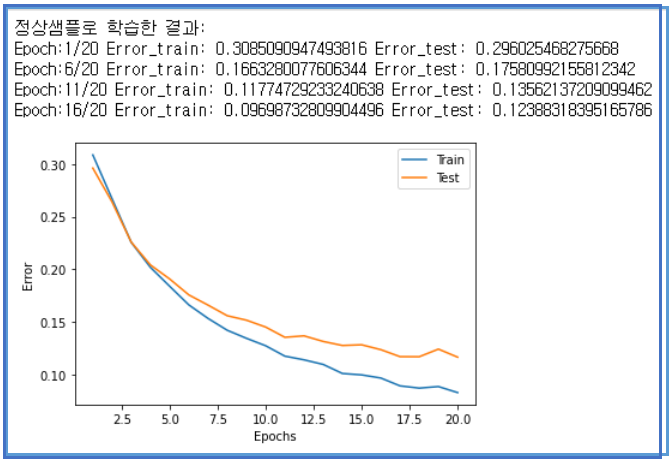
(그림 5) 적대적샘플 제작실험 결과

그림 5 를 보면, 초기 잡음이미지에서 출발하여, 외형은 똑같이 4 로 보이지만 머신러닝시스템에 대입하면 숫자 2(96% 확신), 숫자 3(97% 확신), 숫자 6(96% 확신), 숫자 8(100% 확신)로 잘못 판단하게 만드는 공격 이미지를 제작하였다.

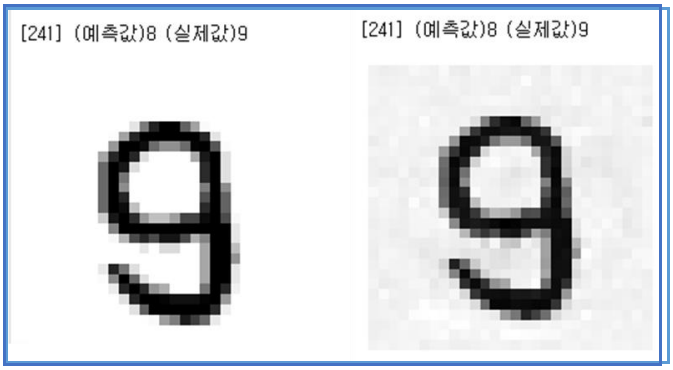
3) 적대적공격 방어실험 결과

우선 MNIST 정상샘플 70,000 개와 동일한 개수의 적대적샘플을 정상샘플이미지를 기반으로 제작하였다.

그리고 정상학습샘플 60,000 개를 사용하여 머신러닝시스템을 학습시킨 후, 정상샘플과 적대적샘플을 각각 대입하여 인식율을 측정하였다.



(그림 6) 머신러닝시스템 학습결과 (정상샘플 사용)



(그림 7) 적대적공격이 성공한 정상샘플(좌측), 적대적샘플(우측)

그 결과 그림 8 과 같이, 정상샘플의 경우 96.9% 이상의 인식율(정확도)을 보였으나, 적대적샘플의 경우는 0.8% 미만의 인식율을 보였다. (테스트데이터 기준) 즉, 적대적공격의 성공율은 99% 이상이었다.

그러나 적대적샘플 60,000 개에 각각 올바른 정답값을 기재한 학습데이터를 사용하여 머신러닝시스템을 추가로 훈련한 결과, 그림 8 과 같이 적대적샘플의 인식율은 기존 0.8%에서 96.6%로 경이적 상승을 기록하였다. 즉, Adversarial Training 기술을 사용하여 제작한 AI Vaccine 을 적용하면, 적대적공격의 성공율은 99%에서 4% 미만으로 감소하였다.

실험번호	실험명칭	정상샘플 정확도		적대적샘플 정확도	
		학습데이터	테스트데이터	학습데이터	테스트데이터
1	ML Training	98.15%	96.92%	0.92%	0.79%
2	Adversarial Training	97.66%	96.39%	98.25%	96.69%

(그림 8) Adversarial Training 실험결과

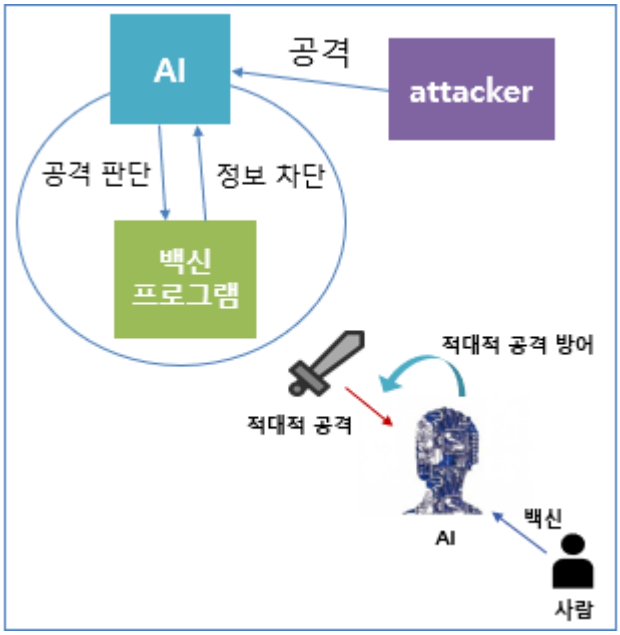
더욱이 AI Vaccine 적용 후에도 정상샘플에 대한 인식율은 거의 그대로 유지되어, 정상샘플과 적대적샘플 모두 올바르게 인식할 수 있는 강인성을 갖추게 되었다고 볼 수 있다.

※인식율(정확도)은 입력한 전체 이미지 중에서 머신러닝시스템이 예측한 결과가 정답과 일치한 비율을 뜻함.

4. 구현 결과

스마트폰 상에서 적대적샘플의 수집, 공격, AI 백신 및 방어에 대한 수행과정을 보여주는 안드로이드 OS 기반 'AI 백신 전용 앱'을 구현하였다.

머신러닝시스템 서버에서 수행한 결과를 오프라인에서 추출한 후 이를 스마트폰 상에 보여줌으로써, 사용자가 적대적 공격이 이루어지는 전체적인 과정을 이해하고 AI 백신 적용에 따른 방어효과를 직접 체험할 수 있다.



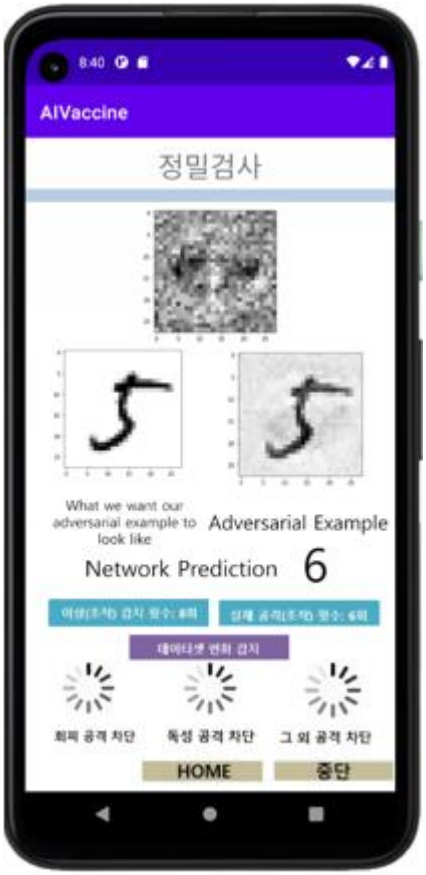
(그림 9) AI 백신프로그램 기능구성도

현재는 주로 시연 및 교육용 장비로 사용되고 있으

나, 향후에는 머신러닝시스템 서버와 온라인으로 연동하여, 원격에서 명령을 내리면 해당 수행결과를 실시간으로 보여주는 온라인 원격 단말도구로서의 역할을 수행하게 될 것이다.

그림 9 는 AI 백신프로그램의 기능구성도이며, AI 백신이 백그라운드에서 AI 시스템을 보완하며 상시 공격여부를 판단하고 차단 및 복원 작업을 수행하는 모습을 보여준다.

그림 10 은 AI 백신앱 정밀검사 실행화면이며, 현장에서 수집한 입력이미지에 대해 적대적샘플여부를 판단하고, 적대적 샘플제작 결과를 표시하며, AI 백신이 적용되기 전과 후 머신러닝시스템의 인지성능 차이를 보여준다.



(그림 10) AI 백신앱 정밀검사 실행화면

이 외에도 AI 학습과정에서 예측오차가 감소하는 모습을 보여주는 오차추이그래프와 함께 예측정확도를 비롯한 다양한 통계기능도 아울러 제공하고 있다. 이를 통해 적대적 AI 공격에 대한 심각성을 인식하고, AI 백신 적용에 따른 방어 효과를 확인할 수 있다.

5. 결론

본 연구에서는 적대적샘플을 제작하고 방어하는 기능을 실제로 구현하였으며, 이를 통해 머신러닝 시스템을 교란할 수 있는 다량의 적대적샘플 제작이 가능하고, 해당 공격 이미지를 통해 네트워크의 잘못된 판단이 이루어지며, Adversarial Training 기술을 사용하여 효과적으로 방어할 수 있음을 입증하였다.

아울러 적대적 공격 및 방어에 대한 수행과정을 시연하고 교육할 수 있는 참여형 안드로이드 스마트폰 앱을 성공적으로 구현하였다.

6. 향후 과제 및 기여도

본 연구를 성공적으로 마무리하면서 이후 단계에서 진행 예정인 연구과제를 소개하면 아래와 같다.

1) MNIST 필기체 이미지 이외의 다양한 야외 현장이미지의 수집 및 적용
2) 머신러닝시스템과 안드로이드 앱과의 온라인 실시간 연동
3) 앱 사용자가 AI 구축정보(뉴런 수, 은닉층 수, 산업군 유형 등)를 결정하고 입력할 수 있는 UI 로 확대
4) 다양한 적대적 공격 및 방어기법을 포괄하도록 연구대상을 확대
5) 자율주행자동차를 위한 적대적 AI 백신의 실용화 연구

본 논문에 대한 개인별 역할 및 기여 내역은 아래와 같다.

1) 저자 : 송채원(연구기획, 앱개발), 이규영(서버개발, 구현, 실험)
2) 공동저자 : 오승아(보안모델연구), 정다예(선행기술분석), 임유리(선행기술조사), 노은지(관련논문분석)

※ 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] Nicolas Papernot, et al. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the 37th IEEE Symposium on Security and Privacy

[2] Daniel Geng, Rishi Veerapaneni, "Tricking Neural Networks: Create your own Adversarial Examples", https://medium.com/@ml.at.berkeley/tricking-neural-networks-create-your-own-adversarial-examples-a61eb7620fd8, Jan 2018