최신 ICT 이슈

Ⅰ. 구글 이미지 인식 교란 스티커 발표, AI 해킹에 대비할 필요

이미지 인식 기술이 급속도로 발전함에 따라 사회의 안전성을 높이는 데 AI의 이용이 확산되고 있는데, 시가지와 공항 감시 카메라의 영상을 AI가 분석하여 테러리스트나 범죄자를 식별해 내는 것이 대표적 사례임. 한편, 구글은 최근 AI의 이미지 인식 알고리즘에 오작동을 일으킬 수 있는 스티커를 발표했으며, 이는 AI 교란을 통한 공격이 늘어날 수 있음을 시사함. AI를 통한 안전성 제고만큼이나 AI를 악용한 안전성 위협 우려가 높아지는 데에 대한 대비가 필요함

◉ 구글 리서치 그룹은 논문을 통해 이미지 인식 인공지능(AI)의 알고리즘을 오작동시킬 수 있는 스티커를 발표하였음

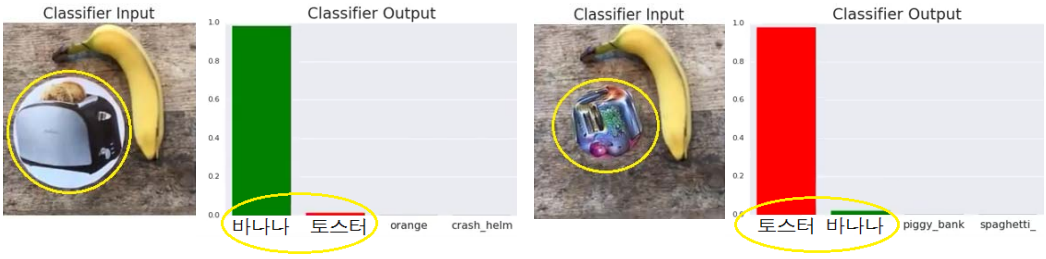
- 논문에 따르면 '애드버세리얼 패치(Adversarial Patch, 적대적 스티커)'라 불리우는, 추상화를 연상시키는 디자인의 원형 스티커를 사물 옆에 붙여 두면 이미지 인식 알고리즘이 제대로 작동하지 않게 된다고 함
- 이 스티커를 바나나 옆에 붙이면 이미지 인식 앱은 바나나를 토스터으로 잘못 인식하게 되는데, 만약 이를 길거리에 붙여둔다면 자율운전자동차가 객체를 오인식해 제대로 주행할 수 없게 될 우려가 있음
- 논문에 소개된 실험결과에 따르면 바나나가 놓여 있는 테이블에 스티커를 붙이면 97%의 확률로 바나나로 인식하던 AI 가 99%의 확률로 토스터로 인식하는 것으로 나타남
- 놀라운 점은 실물 바나나 옆에 토스터 기기가 인쇄된 스티커를 붙인 경우에도 거의 100%의 확률로 바나나를 인식하던 인공지능이 애드버세리얼 패치를

<자료> Research at Google

[그림 1] 애드버세리얼 패치

* 본 내용과 관련된 사항은 산업분석팀(☎ 042-612-8296)과 최신 ICT 동향 컬럼리스트 박종훈 집필위원(soma0722@naver.com ☎ 02-576-2600)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 IITP 의 공식적인 입장이 아님을 밝힙니다.



토스터 애드버세리얼 패치

<자료> https://youtu.be/i1sp4X57TL4

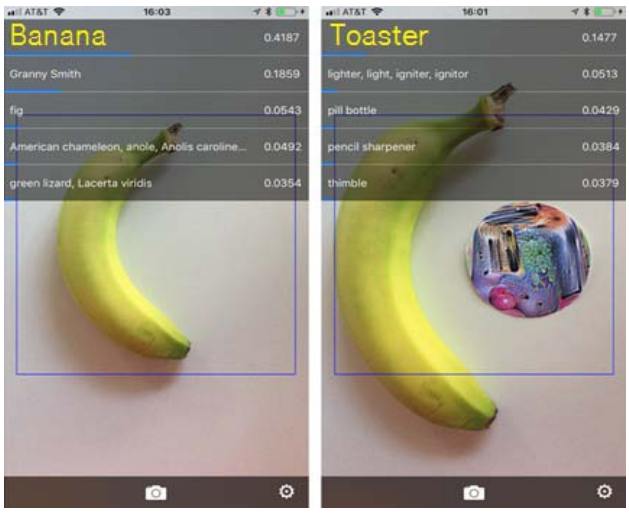
[그림 2] 토스터 그림과 애드버세리얼 패치를 붙인 경우의 결과 비교

붙이자 거의 100% 확률로 토스터 기기라고 인식했다는 점

- 토스터 스티커를 붙인 경우 후보 군에 토스터 기기가 제시되기는 하지만 그 확률은 1% 내외로 오인식 가능성이 없지만, 애드버세리얼 패치를 붙인 경우 바나나로 인식할 확률 역시 1% 내외에 불과해 무조건 오인식이 된다는 것을 보여 줌

◉ 애드버세리얼 패치의 등장에 주목해야 하는 이유는, 이 스티커가 인터넷을 통해 공유되면 기술에 대한 지식이 없는 사람도 누구나 다운받아 인쇄한 후 사용할 수 있기 때문

- 구글이 공개한 이 스티커는 누구나 인쇄하여 자신의 스마트폰에 설치된 이미지 인식 앱을 교란할 수 있는지 실제로 실험해 볼 수 있음



<자료> Arxiv Vanity

[그림 3] 스마트폰 앱의 오작동

- 한 네티즌이 아이폰용 이미지 인식 앱으로 유명한 '데미태스(Demitasse)'를 이용해 실험한 결과 역시 애드버세리얼 패치를 붙이면 바나나를 토스터로 잘못 인식했으며, 심지어 후보군에는 아예 '바나나'가 제시되지도 않았음
- 데미태스 앱은 옥스퍼드 대학의 비주얼 지오메트리 그룹이 개발한 'VGG-CNN'을 이미지 인식 알고리즘으로 탑재하고 있으며, 사진에 찍힌 객체를 파악해 판정하는 기능이 있음
- 이 앱은 VGG-CNN 외에도 이미지 인식 알고리즘의 표준으로 사용되고 있는 'VGG-16' 등도 탑재하고 있는데, 스티커가 데미태스 앱의 오작동을 유발했다면 사실상 현재 사용되고 있는 모든 이미지 인식 앱에 교란을 일으킬 수 있음을 뜻함

◉ 이미지 인식 기능의 근간인 신경망을 쉽게 속일 수 있다는 문제 제기는 그 동안 많았지만, 구글의 스티커는 실생활에서 손쉽게 피해를 야기할 수 있다는 점에서 매우 심각함

- 많은 논문에서 이미지 인식 알고리즘을 속이는 수법이나 네트워크의 취약점을 지적했고, 구글이 공개한 이번 논문도 그 중 하나지만, 지금까지 논의와 크게 다른 점은 이 스티커를 인쇄해 붙이는 것만으로도 AI의 오작동을 일으켜 사회에 문제를 일으킬 수 있다는 것
- 애드버세리얼 패치는 마치 추상화 같아서 사람의 눈으로는 특정 개체가 그려져 있다고 인식 할 수 없기 때문에, 만일 누군가 이미지 인식 오작동을 목적으로 붙여 놓을 경우 아무도 오작동의 위험성이 있다고 느낄 수 없으나 실제로는 큰 위험을 야기하게 될 것임

◉ 예상해 볼 수 있는 위험 중 하나가 자율운전자동차의 운행을 방해하는 것인데, 이 스티커만 붙여 놓아도 도로 교통표지판을 제대로 인식할 수 없게 되기 때문

- 자율운전자동차는 카메라로 포착한 이미지를 이미지 인식 알고리즘으로 분석하여 차량 주변의 개체를 파악하는데, 만약 도로 교통 표지판에 애드버세리얼 패치를 부착하면 자동차는 이를 토스터 기기로 잘못 인식할 수 있음
- 테슬라의 자율운전 지원 기능인 '오토파일럿(Autopilot)'은 도로 표지판을 읽어 속도 제한 여부를 파악하는데, 이 스티커가 부착되면 오토파일럿의 기능에 장애가 초래되며, 당연히 표지판에 스티커를 붙이는 것은 중대 범죄 행위로 처벌 대상이 될 것임
- 2017 년 7 월 워싱턴 대학의 한 연구팀은 교통 표지판에 정교하게 만든 스티커를 붙여 넣으면 이미지 인식 알고리즘이 '정지' 표지판을 '속도 제한' 표지판으로 오인식 한다고 발표한 바 있으며, 구글의 스티커는 이 보다 훨씬 더 간단한 오작동 유도가 가능함
- 집의 지번 표지판에 이 스티커를 붙여두면 구글 스트리트 뷰를 이용한 도로 지도 작성에

도 문제가 발생하는데, 스트리트 뷰는 위치정보를 핀 포인트로 파악하기 위해 건물에 부착되어 있는 지번 표지를 카메라로 촬영한 후 이미지 분석으로 번지를 파악하기 때문

- 지번 표지판의 숫자 옆에 이 스티커를 붙여두면, 이미지 분석 알고리즘은 이를 토스터 기기로 잘못 인식하게 되는데, 다만 지도 서비스 입장에서는 오류가 발생하는 것이지만 거주자 입장에서는 스티커 부착이 효과적인 개인 정보 보호 수단이 될 수도 있음

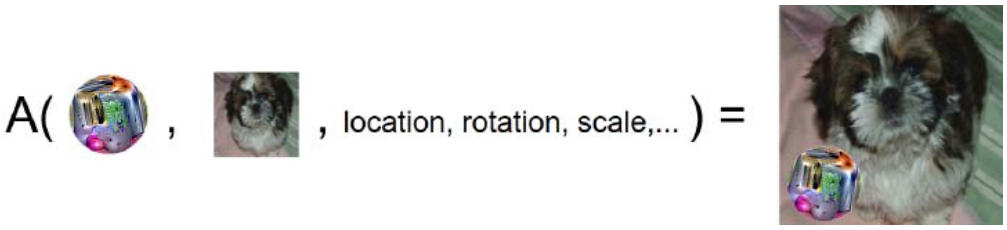
<자료> IEEE Spectrum

[그림 4] 자율운전차의 오인식 유도

- 이처럼 이미지 인식 알고리즘이 인식하는 데이터(example)에 노이즈를 추가해 오류를 일으키는 공격 기법을 "애드버세리얼 이그잼플(Adversarial Example, 적대적 사례)"이라 하며, 구글의 스티커는 이 공격을 누구나 쉽게 할 수 있는 환경이 되었음을 의미함

◉ 구글 리서치 그룹은 논문을 통해 스티커 제작 방법도 공개했는데, 애드버세리얼 패치를 생성하는 독특한 알고리즘을 교육하여 생성한다고 함

- 스티커는 여러 가지 이미지 인식 알고리즘을 오작동시키도록 디자인되는데, 스티커의 효과는 디자인뿐만 아니라 객체에서의 위치, 스티커 방향, 스티커 크기 등에 따라 달라짐
- 가령, 스티커의 방향을 바꾸는 것 만으로 인식 속도가 달라지며, 스티커의 크기를 크게 할수록 효과가 커지는데, 너무 크게 하지 않으면서 최대의 효과를 얻을 수 지점은 객체 크기의 10% 정도로 오작동 확률이 90% 정도가 됨
- 논문에 따르면 애드버세리얼 패치 공격은 '큰 변화량(large perturbation)'을 활용하는데, 작은 변화량의 감지에 초점을 맞추고 있는 현재의 방어 기술들은 이런 큰 변화량에 대해 오히려 강력한 방어 기제로 작동하지 못하게 됨
- 스티커는 "변신에 대한 기대(Expectation Over Transformation)"라고 불리는 특수한 알고리즘으로 생성되며, 스티커를 붙일 객체의 위치, 크기 등의 조건을 감안하여 교란 효과가 최대가 되도록 스티커 생성 알고리즘을 교육함
- 이미지 인식 오작동 유도 효과의 검증에는 현재 사용되는 대표적인 이미지 인식 알고리즘인 Inceptionv3, Resnet50, Xception, VGG16, VGG19 등 5 개를 사용하였음



<자료> Research at Google

[그림 5] 모든 사물을 개로 인식하게 만들 수 있는 애드버세리얼 패치

- 스티커는 'Whitebox-Ensemble(화이트박스-앙상블)'이라는 방식으로 생성되며, 이것이 5 개의 이미지 인식 알고리즘을 오작동시키는지 실증 실험을 하게 되는데, 논문에서는 토스터를 적대적 사례로 만들었지만 모든 객체에 대한 스티커를 만들 수 있다고 함

◉ 구글이 애드버세리얼 패치에 대한 논문을 공개한 이유는 AI 를 이용한 공격의 위험성을 경고하고, 이에 대한 방어를 위해 이미지 인식 알고리즘의 개선을 촉구하기 위해서임

- 구글이 특히 우려하는 것은 이미지 인식 클라우드 서비스가 아니라 네트워크나 컴퓨팅 자원 이용의 제약으로 인해 디바이스 내에서 이미지 인식 알고리즘이 실행되는 경우임
- 이미지 인식 클라우드 서비스들은 대부분 고급 알고리즘을 도입하고 있는데, 가령 구글의 '클라우드 비전(Cloud Vision)' 이미지 인식 서비스에 스티커를 붙인 사진을 입력해도 오작동이 일어나지 않고 사진의 객체를 제대로 인식한다고 함
- 그러나 농장 작업에 쓰이는 자율주행 트랙터나 공사 현장에서 자동으로 작업을 하는 불도저에 탑재된 이미지 인식 알고리즘은 클라우드가 아니라 차량이나 장치 내에서 실행되며, 이러한 엣지(edge, 최종 단말기)에서는 대규모 연산 환경 제공이 어려운 한계가 있음
- 이런 경우 오작동 가능성이 높아 실시간으로 정확한 객체 판정을 할 수 있는 이미지 인식 알고리즘과 이를 지원할 고급 AI 전용 프로세서의 개발이 필요하다는 것이 구글의 제안임
- 일상생활 속에서 드론, 로봇, 자율운전자동차 등의 이용이 확산될 경우 AI 를 악용한 공격은 현실적 문제로 대두될 것이기 때문에 이를 방어하기 위한 대책 강구, 특히 이미지 인식 알고리즘의 정확도를 개선하는 것이 아주 중요한 과제가 된다는 것
- 이제는 보안업체뿐 아니라 해커들도 AI 를 이용하므로 이미지 알고리즘 정확도 개선 노력이 요구되며, 애드버세리얼 패치 기술도 계속 고도화될 것이기 때문에 향후 AI 를 이용한 공격과 방어 수단 개발의 치열한 전투가 본격적으로 시작될 전망

[참고문헌]

[1] Boing Boing, “Adversarial patches: colorful circles that convince machine-learning vision system to ignore everything else”, 2018. 1. 8.

[2] The Verge, “These stickers make computer vision software hallucinate things that aren’t there”, 2018. 1. 3.

[3] Google, “Adversarial Patch”, 2017. 12. 27.