적대적 AI 공격 기법을 활용한 프라이버시 보호

이범기[1], 노현아[2], 최유빈[3], 이서영[4], 이규영[5]
[1] 전북대학교 컴퓨터공학과
[2]성신여자대학교 융합보안공학과
[3]이화여자대학교 휴먼기계바이오공학부
[4]동덕여자대학교 컴퓨터학과
[5]한국과학기술원 전산학부 정보보호대학원
jeongiun@naver.com, pineapple507@naver.com, bl00mfl0wer@naver.com,
zjarhk21@naver.com, leeahn1223@kaist.ac.kr

Privacy Protection using Adversarial AI Attack Techniques

Beom-Gi Lee[1], Hyun-A Noh[2], Yubin Choi[3], Seo-Young Lee[4], Gyuyoung Lee[5]
[1]Dept. of Computer Engineering, Chon-Buk National University
[2]Dept. of Convergence Security Engineering, Sungshin Women's University
[3]Dept. of Mechanical and Biomedical Engineering, Ewha Woman's University
[4]Dept. of Computer Science, Dongduk Women's University
[5]Graduate School of Information Security, KAIST

요 약

이미지 처리에 관한 인공지능 모델의 발전에 따라 개인정보 유출 문제가 가속화되고 있다. 인공지능은 다방면으로 삶에 편리함을 제공하지만, 딥러닝 기술은 적대적 예제에 취약성을 보이기 때문에, 개인은 보안에 취약한 대상이 된다. 본 연구는 ResNet18 신경망 모델에 얼굴이미지를 학습시킨 후, Shadow Attack을 사용하여 입력 이미지에 대한 AI 분류 정확도를 의도적으로 저하시켜, 허가받지 않은 이미지의 인식율을 낮출 수 있도록 구현하였으며 그 성능을 실험을 통해 입증하였다.

I. 서론

오늘날 강력한 얼굴인식 모델의 확산은 개인 정보 보호에 실질적인 위협을 가하고 있다. New York Times의 Kashmir Hill은 민간기업이 30억 개가 넘는 온라인 사진을 수집하고 사전 동의 없이 수백만 명의 시민을 인식할 수 있는 대규모 모델을 훈련한 Clearview.ai에 대해 보도했다.[1]

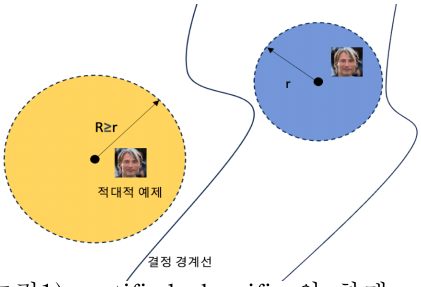
이에 우리는 민간인들이 승인되지 않은 얼굴인식 모델에 의해 식별되지 않도록 할 수 있는 도구가 필요하다고 생각하게 되었다. 본 논문에서는 이미지를 크게 왜곡시키거나 섭동의 시각적 특징이 강조되지 않으면서도, 승인되지 않은 얼굴인식 모델의 정확도를 떨어뜨리는 적대적 모델을 제안한다.

II. 제안모델

우리는 원본 이미지에 육안으로 구분이 불가한 미세 섭동(perturbation)을 의도적으로 추가하여, 딥러닝 모델이 입력 이미지에 대해 오분류를 일으키도록 하는 적대적 공격(Adversarial Attack) 기술을 사용하며, 이렇게 변조한 이미지를 적대적 예제(Adversarial example)라고 한다.

한편 다양한 적대적 공격방식이 제안되고 이러한 공격을 회피하기 위해 certified classifier가 제안되었다. 이는 이미지가 일정한 크기의 Lp-boundary 안에서 Adversarial example이 만들어질 수 없도록 수학적으로 보장하는 기법을 말한다. 이를 통해 특정 범위 안에서는 적대적 예제가 만들어지지 않음을 보장받게 된다. 하지만 이미지가 결정 경계(decision boundary)로 한참 멀리 떨어져 있다면 classifier는 노이즈를 섞은 이미지도 같은 레이블로 분류하게 된다. 게다가 높은 certification radius를 갖기 때문에, 적대적 예제임에도 certified 값이 커지게 된다.

그리하여 적대적 공격의 특성을 가지면서 certified defense 방식에서 높은 점수를 받을 수 있는 공격기법을 적용하게 되었다. 정상적인 이미지처럼 보이게 하는 특징(Imperceptibility), 타겟 클래스로 잘못 분류하도록 유도하는 특징(Misclassification), 높은 인증반경을 갖는 특징(Strongly certified)이 있는 shadow attack 기법을 차용하였다.



(그림1) certified classifier의 한계

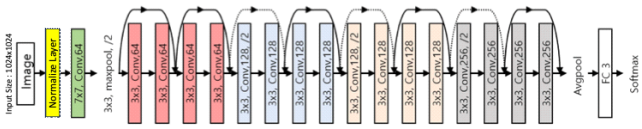
$$\max_{\delta} L(\theta, x+\delta) - \lambda_c C(\delta) - \lambda_{tv} TV(\delta) - \lambda_s Dissim(\delta)$$ [3]

이를 구현하기 위해서 기존의 적대적 공격방식인 Loss 값을 최대로 하는 식은 유지한 채 그림자와 같이 gray scale로 보이게 하는 similar perturbation, 인접한 픽셀 간의 차이를 작게 하도록 Total Variation을 활용한 smoothing, 입력되는 perturbation의 절대값이 커지지 않게 하기 위해 각 색상 채널별로 평균값이 작아지게 하는 color regularizer를 추가로 넣어 진행하였다.

III. 실험

3.1 데이터셋 및 모델 구축

본 연구는 높은 분류성능을 얻기 위해 18개 레이어로 경량화하고 ImageNet 데이터를 사전 학습한 ResNet18 모델을 사용하였다. 여기에 Workstation의 T4 GPU를 사용하여 CelebA-HQ 이미지 2만 개에 대한 전이학습을 실시하였다. 학습 후 1만 개의 테스트 데이터로 성능을 측정한 결과, 89.7119%의 분류 정확도를 보였다.



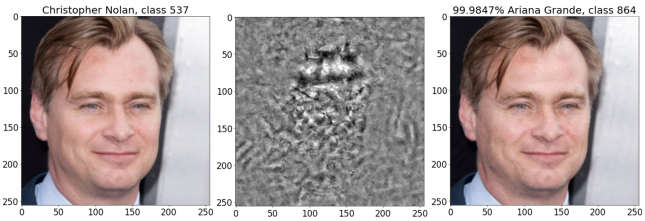
(그림2) Normalize layer를 추가한 ResNet18

본 연구의 핵심인 Shadow attack 기술을 이용한 Adversarial example 생성 작업에는 신경망 학습과정이 없으며, 학습을 완료한 신경망을 기준으로 분류예측값을 변화시키는 과정을 반복적으로 실행하기에, ResNet과 같은 집적화된 모델을 사용하더라도 수행속도 및 실행효율 관련한 영향을 받지 않는다.

3.2 실험 결과

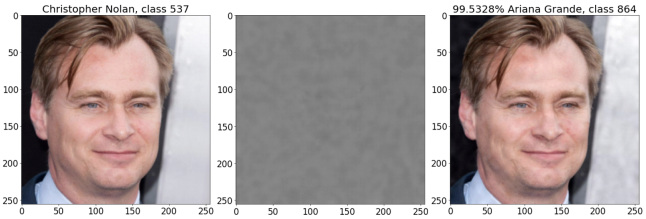
Shadow Attack 적대적 공격기법을 얼굴판별 모델인 ResNet18에 적용해 보았다. 입력 이미지를 지정한 클래스로 인식하게 하는 Targeted attack 방식으로 크리스토퍼 놀란을 아리아나 그란데로 인식하도록 타겟 클래스를 지정하였다.

그림3과 같이 섭동의 각 픽셀 범위를 epsilon으로 제약 조건을 설정하면 섭동이 불규칙적으로 생성되며, 공격 이미지의 흰색 배경에서 섭동 형태의 줄무늬가 적용된 것을 확인할 수 있다. 하지만 여전히 사람이 사진 간의 변화를 감지하기 어려운 상태다. 그 결과 99.9847% 확률로 타켓 클래스로 인식했다.



(그림3) 제약 조건이 적용된 Shadow Attack

제약을 제거하면 그림4와 같이 자연스러운 회색조 섭동이 생성되어 인간이 사진 간의 차이를 더 인지하기 힘들게 된다. 그 결과 99.5328%의 조금 더 낮은 확률로 아리아나 그란데라고 인식했다.



(그림4) 제약 없이 적용된 Shadow Attack

IV. 결 론

본 논문에서는 개인정보보호 및 이미지처리 분야에 적용할 수 있는 이미지 기반의 심층신경망에 대한 적대적 샘플 공격과 CelebA-HQ 데이터셋에 대한 얼굴 분류 모델을 구현하였다.

이로써 입력 이미지에 대한 안정성과 신뢰성이 한층 강화된 것을 확인할 수 있었다. 즉, 외부 공격자의 이미지 변조를 어렵게 만들어 무결성을 보장하면서, 사용자 식별을 어렵게 하여 익명성을 일부 보장한 것이다. 이러한 환경에서 사용자들은 안전하게 이미지를 공유하고 활용할 수 있게 될 것이다.

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

참고문헌

[1] HILL, K. The secretive company that might end privacy as we know it. The New York Times (January 18 2020)

[2] paperswithcode.com/dataset/celeba-hq

[3] Semantic adversarial examples with spoofed robustness certificates, Amin Ghiasi, ICLR 2020