프라이버시 보호를 위한 적대적 AI 공격기법의 연구

조혜원[1], 한지은[2], 김민솔[3], 박주현[4], 홍예령[5], 이규영[6]
[1]이화여자대학교 전자전기공학과 학부생
[2]강남대학교 데이터사이언스학과 학부생
[3]홍익대학교 컴퓨터공학과 학부생
[4]동덕여자대학교 데이터사이언스학과 학부생
[5]성신여자대학교 AI융합학부 학부생
[6]한국과학기술원 정보보호대학원 박사수료
jhw5974@gmail.com, gks3177@naver.com, minsolkim0920@gmail.com,
hhyyn00@gmail.com, iamyerung@gmail.com, leeahn1223@kaist.ac.kr

A Study on Adversarial AI Attack Techniques for Privacy Protection

Hye-Won Jo[1], Ji-Eun Han[2], Min-Sol Kim[3],
Ju-Hyeon Park[4], Ye-Ryeong Hong[5], Gyu-young Lee[6]
[1]Dept. of Electronic & Electrical Engineering, Ewha Woman's University
[2]Dept. of Data Science, Kangnam University
[3]Dept. of Computer Engineering, Hongik University
[4]Dept. of Data Science, Dongduk Women's University
[5]Dept. of AI Convergence, Sungshin Women's University
[6]Graduate School of Information Security, KAIST

요 약

AI 기술이 발전함에 따라 개인정보 유출 및 딥페이크 등의 프라이버시 침해가 심각한 사회적 문제로 대두되고 있다. 본 연구는 프라이버시 보호를 위해 FGSM, One-Pixel, Deepfool, JSMA 등 적대적 공격기법을 소프트웨어로 구현하고, 이를 활용하여 얼굴인식 AI 시스템을 공격하는 실험을 수행하였으며, 그 결과 정보 보호를 위한 최적의 적대적 공격 방안을 도출하였다.

I. 서론

최근 AI 얼굴인식 기술의 빠른 발전으로 딥페이크와 같은 악용 사례가 급증하면서, 프라이버시 침해가 심각한 사회문제로 떠오르고 있다.

본 논문에서는 원본 이미지에 아주 미세한 섭동을 가해 AI 시스템의 이미지 인식능력을 교란하는 적대적 공격기술을 프라이버시 보호에 활용한다.[1]

이를 위해 적대적 공격기법을 구현하고, 관련 실험에 적용하여 가장 효과적인 공격법을 선정하였다.

II. 관련 연구

적대적 예제는 원본 이미지에 미세한 노이즈인 섭동을 의도적으로 추가하여 변조한 이미지로서, 인간의 시각으로는 원본과 차이를 거의 느끼지 못하지만, 딥러닝 모델에는 심각한 오분류를 일으킨다. 적대적 예제를 생성하는 주요 기법들은 아래와 같다.

(1) FGSM (Fast Gradient Sign Method)

$$adv_x = x + \varepsilon * sign(\nabla_x J(\theta, x, y)) \qquad (1)$$

식(1)과 같이, 손실 함수(J)에 대해 입력이미지(x)의 그라디언트를 계산한 후, 그 부호 방향으로 ϵ만큼의 일정한 잡음을 적용한다.

(2) One_Pixel

$$X_{Adv}[i,j] = X[i,j] + \Delta p \qquad (2)$$

식(2)에서 (i,j)는 선택된 픽셀의 좌표, Δp는 해당 픽셀의 변형된 값을 나타낸다. 단 한 개의 픽셀 변경만으로 적대적 이미지를 생성한다는 획기적인 발상에서 출발한 기법이다.

(3) Deepfool

$$f_i(x) + \nabla_x f_i(x) \cdot r = 0 \qquad (3)$$

입력 데이터가 모델의 결정 경계를 넘어가도록 하는 방식이며, 각 클래스 i에 대한 결정 경계를 상기 수식으로 근사하고, 이를 통해 경계를 넘기 위한 최소 변형인 r을 계산한다.

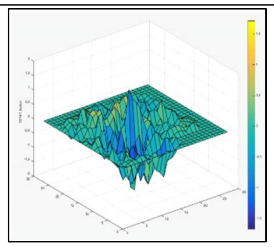
(4) JSMA (Jacobian-based Saliency Map Attack)

식(4)와 같이, 자코비안 행렬은 신경망의 출력에 대한 입력의 편미분들로 이루어진 행렬이며 입력 이

미지의 각 픽셀이 모델의 출력에 미치는 영향을 계산한다.

$$\nabla F(X) = \begin{bmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_1}{\partial X_2} & \cdots & \frac{\partial Y_1}{\partial X_n} \\ \frac{\partial Y_2}{\partial X_1} & \frac{\partial Y_2}{\partial X_2} & \cdots & \frac{\partial Y_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_m}{\partial X_1} & \frac{\partial Y_m}{\partial X_2} & \cdots & \frac{\partial Y_m}{\partial X_n} \end{bmatrix}$$

식(4) 자코비안 행렬



(그림1) 살리언시맵

그림(1)과 같이, 살리언시맵은 자코비안 행렬로 계산된 각 픽셀의 영향을 돌출 형태로 정량적으로 나타내는 시각적 지도이며 최종적으로 영향이 큰 픽셀 2개를 선택하여 적대적 이미지를 생성한다.

III. 제안 이론

적대적 공격은 딥러닝모델을 교란하고 그 분류성능을 저하시키기 위해 연구되어 왔지만, 이를 프라이버시 보호와 같은 긍정적인 방향으로 이용할 수도 있다. 즉, 원본이 아닌 적대적 예제로 생성한 이미지를 사용함으로써, 인터넷 상에서 활동하는 얼굴인식 시스템에 사용자의 이미지가 수집/식별/추적되지 않도록 만들 수 있다. 이는 얼굴인식기술을 광범위하게 사용하는 현대 사회에서 프라이버시 보호를 위한 강력한 도구가 될 수 있다.[1]

IV. 실험

(1) 실험 환경

CelebA-HQ 및 CIFAR-10 데이터셋을 사용하였다. 그림의 결과에서 좌측은 원본이미지, 우측은 생성한 적대적 이미지를 뜻하고, 이미지의 좌상단 숫자는 모델이 예측한 분류클래스의 번호이다.

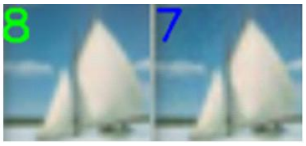
(2) 실험 결과

그림(2)~(5) 모두 원본이미지와 적대적 이미지의 클래스 번호가 다르며, 이는 AI 분류모델이 적대적 이미지를 원본과 다른 클래스로 오분류했음을 의미한다. 즉 적대적 공격이 성공한 것이다.

(그림2) FGSM 결과

(그림3) One Pixel 결과

(그림4) Deepfool 결과



(그림5) JSMA 결과

표(1)과 같이 공격정확도는 FGSM과 JSMA가 매우 높았다. 그러나 원본과 적대적 이미지 간의 외형적 차이를 육안으로 점검한 결과, FGSM 기법으로 생성한 적대적 이미지는 원본의 외형을 훼손한 정도가 매우 큰 반면, One Pixel과 JSMA는 외형적 차이를 거의 발견하기 어려웠다.

<표 1> 적대적 공격 유형 별 성능 측정 결과

No	Method	Attack Accuracy (%)	Visual Difference
1	FGSM	94.9029	매우 높음
2	One Pixel	50.255	거의 없음
3	Deepfool	67.91	높음
4	JSMA	98.11	거의 없음

V. 결론

본 연구를 통하여 섭동량을 최소화하는 L2 Norm 계열의 FGSM 기법보다, 섭동을 가하는 대상 픽셀수를 최소화하려는 L0 Norm 계열의 JSMA 공격기법이 원본 이미지의 왜곡을 최소화하면서도 인식모델의 분류정확도를 극도로 낮추는 데 매우 효과적인 적대적 공격기술인 것을 실험을 통해 확인하였다.

이로써 JSMA 적대적 공격 모델이 외부 공격자의 이미지 변조를 막는 동시에 사용자의 익명성을 보장하는 프라이버시 보호에 가장 적합한 것으로 판단하며, 향후에는 Visual Difference 성능을 정량적으로 평가하기 위하여 투입한 섭동량과 변경한 픽셀수를 공격기법 별로 측정하는 실험을 추가할 예정이다.

ACKNOWLEDGEMENT

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1] 이범기 “적대적 AI 공격 기법을 활용한 프라이버시 보호”, 정보처리학회 ACK 2023 학술대회.
[2] Alex Vakanski, Special Topics: Adversarial Machine Learning, University of Idaho, 2021.
[3] paperswithcode.com/dataset/celeba-hq
[4] https://paperswithcode.com/dataset/cifar-10