인공지능 기반 얼굴 인식 시스템 사이버 공격 구현 및 연구

김동연, 손인수*

동국대학교

2019111885@dongguk.edu isohn@dongguk.edu*

The Implementation and Study of Cyber Attacks on AI-Based Facial Recognition Systems

Dongyeon Kim, Insoo Sohn*

Division of Electronics and Electrical Engineering
Dongguk University

요 약

인공지능 모델의 취약점을 악용한 사이버 공격이 증가하면서, 얼굴 인식 시스템의 안전성과 신뢰성에 대한 우려가 커지고 있다. 특히, 적대적 공격(adversarial attack)은 인공지능 모델의 판단을 왜곡하여 잘못된 인식을 유도할 수 있다. 본 연구에서는 FGSM(Fast Gradient Sign Method) 공격을 Rasberry Pi 4 와 카메라를 이용하여 YOLOv5 모델에 적용하고, 그 결과를 소개한다.

I. 서론

인공지능 기술의 발전은 다양한 분야에 큰 영향을 미치고 있으며, 그 중 얼굴 인식 시스템은 주목받는 기술 중 하나이다. 얼굴 인식은 여러 목적으로 사용되며, 이로 인해 보안 위협에 대한 안정성이 필수적이다. 그러나 얼굴 인식 시스템은 적대적 공격에 취약하며, 이는 AI 시스템의 보안과 신뢰성에 중요한 문제로 대두되고 있다. 본 연구에서는 이와 같은 얼굴 인식 시스템에 대한 적대적 공격 중 하나인 FGSM(Fast Gradient Sign Method)[1] 공격을 실험적으로 구현하고, 그 영향성을 평가하고자 합니다. FGSM 공격을 수행하기 위해 Raspberry Pi 와 카메라를 활용하여 실시간 객체 탐지 모델인 YOLOv5 [2]를 대상으로 실험을 진행하였으, ε 값의 변화를 통해 공격 강도를 조절하고 그 결과를 분석하였다. 이 연구는 FGSM 공격이 얼굴 인식 시스템의 성능에 미치는 영향을 실험적으로 분석함으로써, AI 시스템의 보안 취약성을 확인하고 향후 대응 방안을 모색하는 데 목적이 있다.

그림 1. 실험에 사용된 하드웨어(라즈베리파이 4B, 로지텍 웹캠 C922)

II. 본론

FGSM 은 신경망의 Gradient 를 활용하여 입력 이미지에 미세한 노이즈를 추가함으로써, AI 모델의 잘못된 예측을 유도하는 대표적인 적대적 공격 방법이다. 이 공격은 모델의 손실 함수 $J(\theta,X,y)$의 Gradient 를 기반으로 생성된 노이즈를 입력 데이터 X 에 더하는 방식으로 수행되며, 이는 다음과 같이 정의된다.

$$X'=X+\epsilon\cdot sign(\nabla\ XJ(\theta,X,y))$$

여기서 ϵ 값은 입력 이미지에 추가되는 노이즈의 크기를 결정하며, 값이 커질수록 공격 강도가 증가하게 된다. FGSM 공격의 주요 목표는 입력 이미지가 모델에 의해 잘못 분류되도록 하는 것이며, 본 연구에서는 이 공격 방법을 활용하여 YOLOv5 모델에 대한 적대적 예제를 생성하고 그 영향을 분석하였다.

III. 실험 결과 및 분석

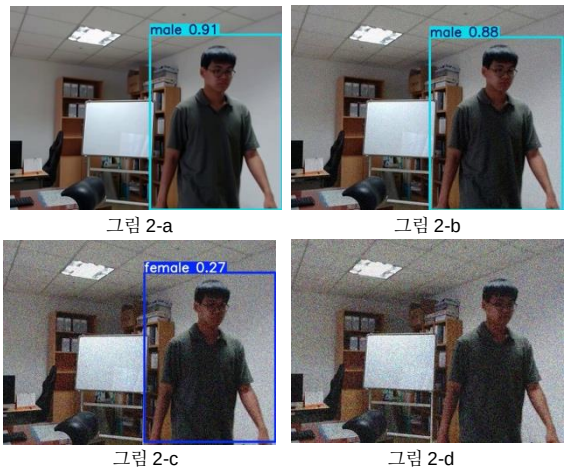

그림 2-a 그림 2-b

그림 2-c 그림 2-d

그림 2. FGSM 의 공격 강도에 따른 모델의 결과

FGSM 공격을 통해 각기 다른 ε 값(0.00, 0.05, 0.125, 0.30)을 적용한 이미지를 생성하였다. 그림 2-a 부터 그림 2-d 까지는 각각 해당 ε 값에 대해 변형된 이미지를 YOLOv5 에서 탐지한 결과를 나타낸다. 이를 통하여 ε 값이 증가할수록 YOLOv5 모델의 객체 탐지 성능이 점차 저하됨을 확인할 수 있다.

그림 2-b ε = 0.05 의 경우 모델이 대부분의 객체를 정확하게 탐지하였으나, 그림 2-c ε = 0.125 에서는 오분류가 발생하기 시작하였다. 그림 2-d ε = 0.30 에서는 다수의 객체를 잘못 탐지하거나 탐지를 실패하는 사례를 관찰할 수 있다. 이러한 결과는 ε 값이 모델의 예측에 미치는 영향을 명확히 확인 할 수 있다.

이를 구체적으로 평가하기 위해, FGSM 공격에 따른 YOLOv5 모델의 성능 변화를 정확도를 통해 분석하였다. 실험 결과, ε 값이 증가함에 따라 모델의 성능이 점차 저하되는 경향을 확인할 수 있다. 그림 3-a 과 그림 3-b 은 각각 ε = 0.00, ε = 0.125 일때의 탐지에 대한 Confusion Matrix 이다. ε = 0.125 의 경우, "female" 클래스에서 89%의 정확도를 보였으나 "background" 클래스와의 혼동이 두드러졌으며, "male" 클래스는 64%의 정확도가 나타났고, 나머지는 주로 "background"로 잘못 예측되었음을 확인할 수 있다. 반면, ε = 0.0 인 원본 이미지에 대한 모델의 성능은 매우 우수하였으며, 모든 클래스에서 높은 정확도를 기록하였다. 이러한 결과는 ε 값이 증가할수록 탐지 실패율이 증가하고, 모델 성능이 공격으로 인해 감소했음을 의미한다.

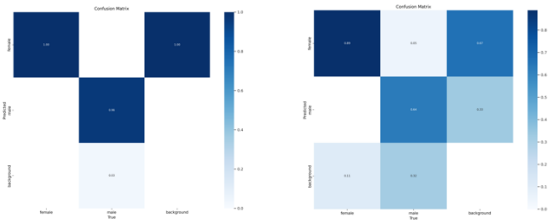

그림 3-a 그림 3-b

그림 3. Confusion Matrix 를 통한 FGSM 공격 모델 간 비교

각 ε 값에 따른 이미지의 주파수 성분 변화를 분석하기 위해 Fast Fourier Transform(FFT)을 사용하였다. FFT 는 시간 도메인에서 주파수 도메인으로 변환하여 이미지의 주파수 특성을 분석할 수 있는 방법이다. 그림 4-a 과 그림 4-b 는 각각 그림 2-a 와 그림 2-d 의 FFT 변환 이미지를 나타낸다.

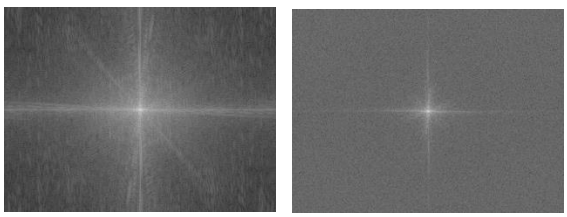

그림 4-a 그림 4-b

그림 4. FFT 를 통한 FGSM 정상 데이터 와 공격 데이터 비교

그림 4-a 과 그림 4-b 는 각각 그림 2-a 와 그림 2-d 의 FFT 변환 한 이미지로, ε 값이 증가함에 따라 그림 4-a 와는 달리 그림 4-b 에서 사진의 주파수가 전체적으로 감소하고 주변부에 많은 노이즈가 발생했음을 확인할 수 있다. 이는 ε 값이 클수록 노이즈에 의해 주파수 성분이 크게 변화함을 시각적으로 확인할 수 있다.

IV. 결론

본 연구에서는 FGSM 공격을 통해 얼굴 인식 시스템의 보안 취약성을 실험적으로 분석하였다. Raspberry Pi 와 카메라를 활용한 실시간 객체 탐지 모델인 YOLOv5 를 대상으로, ε 값을 조절하여 FGSM 공격의 강도를 변화시켰고, 이에 따른 모델의 성능 변화를 평가하였다.

Confusion Matrix 를 통한 YOLOv5 모델 분석과 FFT 를 통한 이미지의 주파수 성분 분석을 통해 ε 값이 증가함에 따라 이미지의 주파수 성분이 크게 변화하고, 이에 따라 노이즈가 증가함을 확인할 수 있다. 실험 결과, ε 값이 증가할수록 모델의 탐지 성능이 점차 저하되는 것을 확인 할 수 있었다.

이러한 FGSM 공격에 대응하기 위해, 학습 이미지에 노이즈를 첨가한 예제를 포함하여 학습하는 방법이 제시되었다. Madry et al. [3] 은 "adversarial training" 기법을 통해 FGSM 공격에 견고한 모델을 개발하였고, Goodfellow et al. [4]은 손실 함수에 적대적 예제를 활용하여 모델의 견고성을 높이는 방법을 제안했다. 또한, "Trans-IFFT-FGSM: A novel fast gradient sign method for adversarial attacks" [5] 논문에서는 단순한 FGSM 의 변형을 넘어 주파수 도메인에서의 조작을 통해 모델에 더 큰 혼란을 줄 수 있다는 점을 강조하고 있다. 본 연구는 FGSM 공격이 AI 기반 시스템의 보안 취약성에 미치는 영향을 FFT 를 통해 수학적으로 분석함으로써, 적대적 공격에 대응하기 위한 방어 메커니즘 개발의 필요성을 강조하며, 향후 연구에서 다양한 공격 기법에 대한 방어 전략을 모색하는 것이 중요함을 시사한다. 특히, 실시간 시스템의 보안성을 강화하는 방안이 향후 연구의 중요한 과제가 될 것이라 생각된다.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252328).

참 고 문 헌

[1] https://tutorials.pytorch.kr/beginner/fgsm_tutorial.html
[2] https://github.com/ultralytics/yolov5
[3]Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks." arXiv preprint arXiv:1706.06083.
[4]Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). "Explaining and Harnessing Adversarial Examples." arXiv preprint arXiv:1412.6572.
[5] Naseem, M.L. Trans-IFFT-FGSM: a novel fast gradient sign method for adversarial attacks. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-18475-7