XAI 를 활용한 적대적 공격 탐지 연구 동향 분석

전아영 [1], 이연지 [2], 이일구 [3]
[1] 성신여자대학교 융합보안공학과 학부생
[2] 성신여자대학교 융합보안공학과 박사과정
[3] 성신여자대학교 융합보안공학과 교수

20221127@sungshin.ac.kr, cselab.lyj@gmail.com, iglee@sungshin.ac.kr

Research trend analysis on adversarial attack detection utilizing XAI

A-Young Jeon[1], Yeon-Ji Lee[1], Il-Gu Lee[1]
[1]Department of Convergence Security Engineering, Sungshin Women's University

요 약

인공지능 기술은 사회 전반에 걸쳐 다양한 분야에서 활용되고 있다. 그러나 인공지능 기술의 발전과 함께 인공지능 기술을 악용한 적대적 공격의 위험성도 높아지고 있다. 적대적 공격은 작은 왜곡으로도 의료, 교통, 커넥티드카 등 인간의 생명과 안전에 직결되는 인공지능 학습 모델의 성능에 악영향을 미치기 때문에 효과적인 탐지 기술이 요구되고 있다. 본 논문에서는 설명 가능한 AI 를 활용한 적대적 공격을 탐지하는 최신 연구 동향을 분석한다.

1. 서론 및 배경

인공지능(Artificial Intelligence, AI) 기술은 사회 전반에 걸쳐 다양한 분야에서 발전하고 있다. 특히 이미지 인식, 자연어 처리, 예측 모델링 등의 영역에서 인공지능은 효율성과 정확성을 크게 향상시켰다. 그러나 인공지능을 접목한 시스템의 발전과 함께 인공지능 기술을 악용한 적대적 공격의 위험성도 높아지고 있다[1].

인공지능은 다양한 유형의 적대적 공격(adversarial attack)에 취약하며, 이러한 공격은 사회적으로 큰 위협을 초래한다[2]. 예를 들어, 인공지능 기반 이미지 인식 시스템에 대한 적대적 공격은 얼굴 인식 기술을 활용한 보안 시스템이나 교통 감시 시스템의 정확성을 떨어뜨려 사회적 혼란을 야기할 수 있다. 또한, 인공지능 기반의 예측 모델링 시스템에 대한 적대적 공격은 금융, 의료, 제조 등 여러 산업 분야에서 부정확한 의사결정을 유발할 수 있다. 적은 왜곡으로도 인공지능의 판단에 큰 영향을 미칠 수 있다는 점에서 적대적 공격은 인공지능 기술의 신뢰성과 안전성에 대한 심각한 위협이 되고 있다.

설명 가능한 AI(eXplainable AI, XAI)는 인공지능 모델이 생성한 결과를 사람이 이해하고 신뢰할 수 있게 돕는 인공지능 모델이다. 학습과정에서 데이터로부터 다양한 패턴을 학습하고, 분석하여 사람이 이해하기 어려운 기계학습의 처리 과정과 규칙을 시각적으로 이해할 수 있는 그래프나 차트의 형태로 도출한다. 적대적 공격은 모델의 신뢰성을 낮추는 것을 목적으로 하는 공격이므로, XAI 를 활용하면 모델의 신뢰성을 개선하고, 나아가 기존의 탐지 모델이 파악하지 못한 잠재적인 취약점 탐지에 도움을 줄 수 있다[3].

본 연구에서는 XAI 기술을 활용하여 적대적 공격을 탐지하는 종래의 연구들을 분석하고, 종래 기술의 한계점과 시사점에 대해 정리하였다.

2. XAI 를 활용한 적대적 공격 탐지 기술

Ben Pinhasov et al[3]의 연구에서는 딥페이크(Deepfake)를 탐지하기 위한 탐지 모델에 적대적 공격이 진행되는 것을 방어하기 위해 XAI 를 도입하였다. 해당 연구는 XAI 모델을 통해 미리 정상 데이터와 비정상 데이터의 분석 결과를 전이 학습 모델인 resnet 을 통해 학습하였다. 이후 학습한 resnet 과 일반적인 딥페이크 탐지기를 결합하여 분류를 진행한다. 입력된 이미지의 딥페이크 여부와 함께 적대적 공격 여부를 결과값으로 확인할 수 있으며, XAI 기반의 적대적 공격 탐지 모델은 약 83.45%의 정확도를 보였다. 하지만 제안하는 방법의 경우 딥페이크 탐지 모델과 XAI 기반의 적대적 공격 탐지 모델을 모두 활용하므로 시스템이 복잡하다.

Chase Walker et al.[4]은 XAI 가 정상 이미지와 적대적 샘플을 분석했을 때, 서로 다른 특성을 갖는 것에 초점을 맞추고 XAI 기반의 적대적 샘플 탐지 연구를 수행하였다. 분석 결과에 따르면 정상 데이터는 주요 대상에 초점을 맞추어서 핵심 피처가 조밀하게 모여 있는 반면에 적대적 샘플은 광범위한 섭동 노이즈로 인해 주요 속성이 희박하고, 분산된 특성을 보였다. XAI 기반의 적대적 샘플 탐지 모델을 생성하기 위해 XAI 의 분석 결과를 토대로 이진분류 알고리즘을 제안하였다. 결과적으로 제안하는 방법은 탐지 정확도를 최대 99%까지 개선했다. 하지만 이 방법은 학습 모델이 매우 복잡하다.

<표 1> 선행 문헌의 기여점 및 한계점

Title	기여점	한계점
XAI-Based Detection of Adversarial Attacks on Deepfake Detectors[3]	입력된 이미지의 딥페이크 여부와 함께 적대적 공격 여부를 결과값으로 확인할 수 있으며, XAI 기반의 적대적 공격 탐지 모델은 약 83.45%의 정확도를 보임.	딥페이크 탐지 모델과 XAI 기반의 적대적 공격 탐지 모델이 각각 존재한다는 점에서 높은 복잡성을 갖으며, 데이터셋에 대한 자세한 정보가 없고, 유독 낮은 성능을 보이는 모델에 대한 부가적인 설명이 부족
Adversarial Pixel and Patch Detection Using Attribution Analysis[4]	XAI 기반의 적대적 샘플 탐지 모델 생성을 위해 XAI 의 분석 결과를 토대로 이진분류 알고리즘을 제안하였고, 결과적으로 제안하는 방법은 최대 99%의 탐지 정확도를 보임.	분류기 생성과 XAI 분석, 그리고 분석 결과를 토대로 재학습을 시키는 과정까지 포함하므로 학습 모델의 복잡성이 높음.

표 1 은 분석한 선향 문헌의 기여점과 한계점을 비교 분석한 표이다. 종래 연구를 분석해본 결과 두가지 연구에는 공통된 한계점이 있다. 첫째, 모델의 복잡성 문제이다. 적대적 공격에 가장 취약한 드론, 센서, 자율주행 등의 IoT 장치는 컴퓨팅 자원에 제약이 있기 때문에 제시하는 것처럼 복잡한 모델은 적합하지 않다. 둘째, 데이터셋의 신뢰성 문제이다. 충분한 양의 데이터로 학습하고 평가하였는지 알 수 있는 지표가 없고, 전문적인 데이터셋인지 알기 어렵다. 셋째, 데이터셋 라벨링 대한 문제가 있다. 데이터셋의 라벨 생성을 위해 공격을 탐지하고 분류하기 위해 전문 인력이 필수적으로 요구되어 비효율적이다. 따라서 비지도 학습 기반의 데이터 라벨링 등 자동화 기술의 도입이 필요하다.

3. 결론 및 향후 연구

본 연구에서는 적대적 공격을 탐지하기 위해 XAI 를 활용한 종래의 연구를 분석하고 한계점과 연구 방향을 도출하였다. 선행 연구들은 상당히 복잡한 모델을 제시하고 있으나, 자율 주행과 같은 IoT 분야에는 복잡한 모델이 적합하지 않고, 한정된 데이터셋과 라벨링 측면에서 한계가 있음을 확인하였다.

이에 따라 향후에는 IoT 에서 활용할 수 있는 적대적 공격 경량 탐지 기법에 대한 연구가 필요하며, 충분한 데이터셋 선정과 자동화된 데이터 라벨링 기술을 활용한 데이터셋 구축 연구가 진행되어야 한다

Acknowledgements

본 논문은 2024 년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Qiu, Shilin, et al. "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, volume. 9, no. 5, 2019.

[2] Chakraborty, Anirban, et al. "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, volume. 6, no. 1, pp. 25-45, 2021.

[3] Ben Pinhasov, Raz Lapid, Rony Ohayon, Moshe Sipper and Yehudit Aperstein, “XAI-Based Detection of Adversarial Attacks on Deepfake Detectors,” Cryptography and Security, Mar, 2024..

[4] Chase Walker, Dominic Simon, Sumit Kumar Jha and Rickard Ewetz, “Adversarial Pixel and Patch Detection Using Attribution Analysis,” MILCOM 2023 - 2023 IEEE Military Communications Conference (MILCOM), usa, Oct, 2023.