AI 모델의 적대적 공격 대응 방안에 대한 연구

박재경[O], 장준서[*]
[O]한국폴리텍대학 사이버보안과,
[*]한국폴리텍대학 사이버보안과
e-mail: jakypark@kopo.ac.kr[O], junseo040108@gmail.com[*]

A Study on Countermeasures Against Adversarial Attacks on AI Models

Jae-Gyung Park[O], Jun-Seo Chang[*]
[O]Department of Cyber Security, Korea Polytechnics,
[*]Department of Cyber Security, Korea Polytechnics

● 요 약 ●

본 논문에서는 AI 모델이 노출될 수 있는 적대적 공격을 연구한 논문이다. AI 챗봇이 적대적 공격에 노출됨에 따라 최근 보안 침해 사례가 다수 발생하고 있다. 이에 대해 본 논문에서는 적대적 공격이 무엇인지 조사하고 적대적 공격에 대응하거나 사전에 방어하는 방안을 연구하고자 한다. 적대적 공격의 종류 4가지와 대응 방안을 조사하고, AI 모델의 보안 중요성을 강조하고 있다. 또한, 이런 적대적 공격을 방어할 수 있도록 대응 방안을 추가로 조사해야 한다고 결론을 내리고 있다.

키워드: 적대적 공격(Adversarial Attack), 보안(Security), 대응 방안(Countermeasure)

I. Introduction

ChatGPT나 이루다 등의 AI챗봇이 많아지면서 AI 모델에 대한 적대적 공격(Adversarial Attack) 사례 또한 많아지고 있다. 적대적 공격은 AI 모델을 속이거나 예측을 왜곡시켜 잘못된 결과가 나오도록 유도하는 해결되지 못한 보안 취약점 중 하나로, 이루다 1.0의 경우 이런 공격에 취약함을 볼 수 있었던 사례 중 하나이다. 이에 본 논문은 적대적 공격을 효과적으로 방어할 수 있는 방법을 연구해보고자 한다.

II. Preliminaries

1. Adversarial Attacks

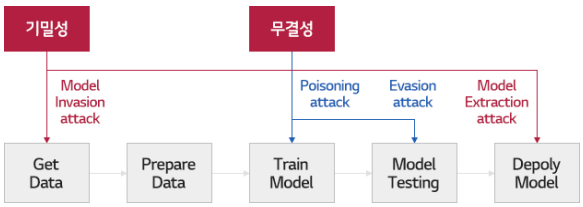

Fig. 1. Machine Learning and Adversarial Attacks (LG CNS Blog - AI Data)

그림 1과 같이 적대적 공격에는 4가지 종류가 있다. Poisoning Attack, Evasion Attack, Model Extraction Attack, Model Invasion Attack이 있다.

Table 1. Types of Adversarial Attacks

적대적 공격 종류	설명
Poisoning Attack	훈련 데이터에 악성 샘플을 삽입하여 모델의 학습 과정을 조작하는 공격
Evasion Attack	적대적 예제를 생성하여 모델의 검출 또는 분류를 회피하는 공격
Model Extraction Attack	AI 모델의 지식을 악의적인 공격자가 추출하는 공격
Model Invasion Attack	악의적인 AI 모델을 시스템에 침입시켜 보안을 우회하거나 악의적인 동작을 수행하는 공격

위와 같이 공격은 4가지 종류로 흔히 분류되며, 각각의 공격은 각자의 특이점을 가지고 있다. Poisoning Attack과 Model Invasion Attack은 모델 자체를 공격하며, Evasion Attack과 Poisoning Attack 둘 다 잘못된 모델을 AI에게 제공함으로써 공격하며, Model Extraction Attack과 Model Invasion Attack 둘 다 모델을 복제하기 위한 공격이다.

2. Adversarial Countermeasures

Adversarial Attack에 대응하는 방안 (Countermeasures)는 여러 가지가 있다. Adversarial Training은 적대적 공격으로부터 학습하게 하여 적대적 공격에 내성을 가지게 하는 대응 방법으로, 여러 AI가 이 기술을 사용하고 있다. 이를 통해 적대적인 환경에서도 AI는 안정적인 예측을 할 수 있게 된다. Adversarial Training 이외에도 AI 모델의 결과를 분석하거나 학습 값이 노출되지 않도록 하는 결괏값 차단과 적대적 공격(Adversarial Attack)을 판단하는 AI를 따로 두어 적대적 공격으로부터 AI 모델을 보호하는 방법도 있다. 또한, Chat AI는 금칙어 시스템과 비슷한 선제 보호 기법을 사용하여 먼저 공격을 감지하고 차단하는 방식 또한 사용하고 있다.

III. The Proposed Scheme

Chat AI를 보호하기 위해 적대적 공격 (Adversarial Attack)에 대한 방어책을 여러 가지 사용해야 한다고 판단한다. 특히 이미 공격에 취약한 이루다 1.0이나 Google의 인공지능 챗봇 Tay의 사례처럼 공격받았을 때 대응에 실패할 시 결과가 매우 위험하여서, 적대적 공격의 대응 방안은 필요한 상황이다. 추가로, AI의 불법 복제를 막기 위해 결괏값의 분석을 막고 AI 자체 데이터베이스에 대한 공격을 막을 수 있는 보안 조치를 해두어야 한다고 주장한다. 특히 최근 출시 되고 있는 Chat AI들의 보안이 특히나 중요하며, 악의적인 사용자가 데이터를 훼손시키거나 손상을 일으키는 상황이 발생하지 않도록 하여야한다.

IV. Conclusions

본 논문에서는 AI 모델에 대한 적대적 공격의 종류와 대응 방안에 대해서 다루었다. 적대적 공격은 AI 모델의 신뢰성과 무결성을 무너트릴 수 있는 위협이며, 이에 대응하는 방안을 연구하는 것이 필요하다.

AI 모델에 대한 적대적 공격 분야는 계속해서 연구되어야 하며, 견고한 방어 기법의 개발이 필요하다. AI의 보호는 데이터 유출을 방지하고 데이터의 무결성을 보장하며, 사용자의 악의적 행동으로부터 보호하는데 매우 중요하다. 효과적인 방어 전략의 구현을 통해 AI 모델의 적대적 공격에 대한 내성을 향상하게 시키고, AI 기술의 사용에 대한 신뢰를 갖출 수가 있다.

REFERENCES

[1] LG CNS Blog – AI/Data, https://www.lgcns.com/blog/cns-tech/ai-data/9616/

[2] Eykholt, Kevin, et al “Robust physical-world attacks in deep learning models”.arXiv preprint arXiv : 1707.08945, 2017

[3] Sandip Kundu, security and Privacy of Machine Learning Algorithms, ISQED 2019