한국컴퓨터정보학회 하계학술대회 논문집 제31권 제2호 (2023. 7)

AI 서비스 보안에 대한 자료 조사

김주원*, 박재경O
*한국폴리텍대학 사이버보안학과,
O한국폴리텍대학 사이버보안학과
e-mail: saykjw0707@gmail.com*, jakypark@gmail.comO

A Research the literature on AI service security

Juwon Kim*, Jaekyoung ParkO
*Dept. of Cybersecurity, Korea Polytechnic University,
ODept. of Cybersecurity, Korea Polytechnic University

● 요 약 ●

인공지능 (AI) 서비스는 현대 사회에서 중요한 역할을 맡고 있다. 그러나 이러한 서비스는 보안과 관련된 문제들을 가지고 있다. 본 논문은 AI 서비스의 보안과 관련된 문제와 해결책을 조사하고자 한다. AI 서비스의 개요와 대표적인 상용 서비스를 간략히 소개 후, AI 서비스에서 발생할 수 있는 보안상의 문제와 Chat GPT를 중심으로 한 보안 문제에 대해 다루고자 한다. 또한, 향후 AI보안 서비스 연구 분야와 적재적 기계학습 연구에 대한 전망을 살펴볼 예정이다. 이를 통해 안전하고 신뢰성 있는 AI 서비스를 제공하는데 기여하고자 한다.

키워드: 적대적 기계학습(Adversarial Machine Learning), 인공지능 기반 서비스 제공(AI 서비스)

I. Overview of AI services

인공지능(Artificial Intelligence) 서비스는 인공지능 기술을 활용하여 사용자에게 유용한 서비스를 제공하는 것이다. 인공지능 기술은 컴퓨터가 인간의 학습, 추론, 자연어 처리 등과 같은 지능적인 작업을 수행할 수 있도록 하는 기술이다.

1. AI 서비스 중 대표적인 상용서비스

i. Siri, Google Assistant, Alexa 등의 가상 비서 서비스: 음성 인식 기술을 활용하여 사용자의 명령을 인식하고 이에 대한 대화형태의 답변을 제공한다.

ii. Netfilex, Youtube, Amazon 등의 추천 알고리즘: 사용자의 이전 시청 기록, 검색 기록 등을 분석하여 맞춤형 추천을 제공한다.

iii. Facebook, Instagram, Twitter 등의 이미지 인식 기술: 사진이나 비디오의 이미지를 분석하여 자동으로 태그를 달거나, 검색 등의 서비스를 제공한다.

iv. Amazon Go, CJ ONE Smart Store 등의 자동 결제 및 인식 기술: 이미지 인식, 센서 기술 등을 활용하여 사용자가 제품을 선택하면 자동으로 결제가 이루어지는 서비스를 제공한다.

v. IBM Watson, Google Cloud AI 등의 인공지능 플랫폼: 기업이나 조직이 자신만의 AI 서비스를 개발할 수 있도록 도와주는 플랫폼 서비스를 제공한다.

vi. Chat GPT 등의 자연어 처리 기술을 기반으로 하는 챗봇 서비스: 대화형 인공지능 기술을 기반으로, 사용자의 질문에 대한 답변을 자연스러운 형태로 제공하며 사용자의 질문에 대한 자연어 처리를 수행하고 실시간으로 응답을 생성하여 다양한 정보를 제공한다.

II. AI 서비스의 보안상 문제점

AI 서비스는 개인정보를 다루는 등 민감한 정보를 처리하기 때문에 보안상의 문제가 발생할 수 있다. 이에 따라 AI 서비스를 이용하는 사용자와 서비스 제공자는 보안에 대한 책임을 지고 적절한 대응책을 마련해야 한다. 아래는 AI 서비스의 보안상 문제를 몇 가지 나열해 보았다.

i. 데이터 보호: AI 서비스는 많은 양의 데이터를 수집하고 처리한다. 이러한 데이터는 고객의 개인정보를 포함하고 있을 수 있기 때문에, 데이터의 안전한 보호가 매우 중요하다. 이러한 개인정보를 적절한 방법으로 암호화하고 보호해야 한다.

ii. 모델 보호: AI 모델에 대한 접근 권한이 없는 사람들이 모델에

대한 액세스를 획득하면, 모델을 악용하여 다양한 범죄 행위에 이용될 수 있다. 모델을 보호하기 위해 적절한 접근 제어와 모델 보호 메커니즘을 마련해야 한다.

iii. 새로운 취약점 대응: AI 서비스는 계속해서 새로운 취약점과 보안 위협에 노출될 수 있다. 따라서, 보안 이슈를 모니터링하고 신속하게 대응해야 한다. 이를 위해 보안 업데이트와 패치를 시스템에 적용하는 것이 중요하다.

iv. 인공지능 기술의 부작용: AI 서비스는 학습 데이터나 학습 모델의 품질에 따라 예측력이 크게 영향을 받을 수 있다. 만약 학습 데이터나 학습 모델의 품질이 떨어진다면, 잘못된 예측을 할 수 있다. 이러한 부작용을 방지하기 위해 적극적인 모니터링과 문제 해결 방안을 마련해야 한다.

v. 사람의 개입 부재: AI 서비스는 사람의 개입 없이 자동화된 프로세스를 수행한다. 따라서, 인공지능이 예측한 결과가 잘못되었을 경우, 심각한 결과를 가져올 수 있다. 이를 방지하기 위해 AI 서비스가 수행하는 작업을 감시하고, 문제가 발생할 경우 적극적인 개입을 할 수 있는 구조를 구축해야 한다.

1. AI 서비스의 '인공지능 기술의 부작용'

AI 서비스에서 인공지능 기술의 부작용은 예측이나 결정을 하는 과정에서 예기치 않은 결과를 가져올 수 있는 위험성을 말한다. 이러한 부작용은 아래와 같은 형태로 나타날 수 있다.

i. 데이터 바이어스(Data Bias): AI 모델이 학습하는 데이터에 바이어스가 존재할 경우, AI 서비스가 정확한 예측을 하지 못할 수 있다.

ii. 과적합(OverFitting): AI 모델이 학습 데이터에 과하게 적합되어, 새로운 데이터에 대한 일반화 성능이 떨어지는 현상을 말한다. 이러한 경우 AI 서비스가 정확한 예측을 하지 못할 수 있다.

iii. 블랙박스(Black Box): AI 모델이 내부적으로 어떻게 작동하는지 이해하기 어려운 경우를 말한다. 이 경우 AI 서비스가 어떻게 예측을 내리는지 설명할 수 없기 때문에, 예측 결과에 대한 신뢰도가 떨어질 수 있다.

iv. 민감도(Sensitivity): AI 모델이 학습 데이터에서 민감하게 반응하는 패턴을 찾아내는 경우를 말한다. 예를 들어, 인종이나 성별들과 같은 민감한 특성을 고려하지 않고 AI 모델을 학습시키면, 이러한 특성이 결과에 영향을 미치게 된다.

2. Chat GPT에서 보안상 문제점

Chat GPT는 인공지능 언어모델로써, 다양한 보안상 문제점이 존재할 수 있다. 아래에 Chat GPT의 대표적인 보안상 문제를 나열해 보았다.

i. 개인정보 보호: Chat GPT를 이용한 대화에서 사용자의 개인정보(이름, 주소, 전화번호 등)가 노출될 수 있다. 이러한 경우 개인정보 보호를 위한 적절한 보호 조치가 필요하다.

ii. 악용 가능성: Chat GPT는 일반적인 대화를 모사하기 때문에, 악용되어 부적절한 대화가 이루어질 수 있다. 예를 들어, 사기나 성희롱 등에 이용될 수 있다.

iii. 허위 정보 생성 가능성: Chat GPT는 대화의 흐름에 맞춰 자연스러운 문장을 생성하기 때문에, 허위 정보가 생성될 가능성이 있다. 이는 정보 신뢰성과 문제로 작용할 수 있다.

iv. 공격 가능성: Chat GPT를 이용한 대화에서 악의적인 사용자가 입력한 입력값이나 명령어를 통해 서버에 해킹이 가능한 취약점이 존재할 수 있다.

3. Chat GPT의 보안상 문제점 해결 방안

i. 개인정보 보호를 위한 암호화 기술 도입: Chat GPT를 이용한 대화에서 사용자의 개인정보가 노출되는 것을 막기 위해, 암호화 기술을 도입하여 보안성을 강화할 수 있다.

ii. 악용 가능성 방지를 위한 사용자 인증 시스템 구현: Chat GPT를 이용한 대화에서 악의적인 사용자로부터 서비스를 보호하기 위해, 사용자 인증 시스템을 구현하여 인가된 사용자만이 대화를 이용할 수 있도록 할 수 있다.

iii. 허위 정보 생성 가능성 방지를 위한 데이터 검증 및 필터링: Chat GPT를 이용하여 생성된 대화에서 허위 정보가 생성될 가능성을 최소화하기 위해, 데이터 검증 및 필터링 시스템을 도입하여 정확한 정보만을 대화에 이용할 수 있도록 할 수 있다.

iv. 공격 가능성 방지를 위한 보안 업데이트 및 감사: Chat GPT를 이용한 대화에서 서버에 해킹 등의 공격이 발생하지 않도록, 보안 업데이트 및 감사를 적극적으로 시행하여 취약점을 최소화할 수 있다.

v. 모델 자체의 보안성 강화: Chat GPT 모델 자체의 보안성을 강화하는 방안으로, 모델의 학습 데이터를 보호하고, 모델 자체를 암호화하거나, 접근 권한을 제한하는 등의 방식을 적용할 수 있다.

4. AI 서비스 중 Chat GPT에 보안 대책 관련 논문 및 연구

IBM의 연구자들이 작성한 논문은 Chat GPT와 같은 자연어 처리 모델을 안전하게 학습시키기 위한 방법으로, 페더레이션 러닝과 차등 프라이버시를 결합한 방법을 제안한다. 페더레이션 러닝은 분산된

데이터에서 중앙 집중적으로 모델을 학습시키는 것이 아니라, 분산된 데이터를 가지고 로컬 모델을 학습시키고 중앙 서버에 보고하는 방식을 사용한다. 이를 통해 데이터의 보안과 프라이버시를 보호할 수 있다. 하지만 로컬 모델에서 보고하는 정보에는 여전히 개인정보가 포함될 수 있다. 이를 해결하기 위해 차등 프라이버시를 사용한다. 차등 프라이버시는 개인정보를 포함한 데이터를 다른 데이터와 섞는 것으로, 개인정보를 보호하면서도 모델의 학습 성능을 유지할 수 있다. 이 논문에서는 이러한 방법들을 결합하여, 자연어 처리 모델을 안전하게 학습시키는 방법을 제안한다. Chat GPT와 같은 모델에서도 이러한 방법을 적용함으로써 보안과 프라이버시를 보호할 수 있다.

III. 향후 AI 보안 서비스 연구 분야

i. Adversarial Machine Learning: 악의적인 공격자들이 AI 모델을 해킹하거나 조작하는 것을 막는 기술이다. 이를 위해서는 새로운 안전성 검사 방법론과 방어전략이 필요하다.

ii. 프라이버시 보호: AI 모델이 개인정보를 처리할 때는 프라이버시 보호가 매우 중요하다. 데이터에 포함된 개인정보를 보호하면서도 AI 모델이 효과적으로 작동할 수 있는 방법을 연구하는 것이 필요하다.

iii. 페더레이션 러닝: 페더레이션 러닝은 분산된 데이터에서 중앙 집중적으로 모델을 학습시키는 것이 아니라, 로컬 모델을 학습시키고 중앙 서버에 보고하는 방식을 사용한다. 이를 통해 데이터의 보안과 프라이버시를 보호할 수 있다.

iv. AI 모델 해킹 방지: AI 모델은 보안성이 좋다고는 할 수 없다. 따라서 AI 모델 해킹 방지 기술을 연구하여 보안성을 강화할 필요가 있다.

IV. Adversarial Machine Learning에 관한 연구

Adversarial Machine Learning은 AI 모델을 공격하는 악의적인 공격자를 막는 기술이다. 이 분야에서의 연구는 크게 두 가지로 나뉜다.

i. 새로운 안전성 검사 방법론과 방어 전략을 개발: 악의적인 공격자들은 다양한 방법을 사용하여 AI 모델을 조작하거나 해킹한다. 이를 방지하기 위해서는 새로운 안전성 검사 방법론과 방어 전략이 필요하다. 예를 들어, 생성적 적대 신경망(GAN)을 이용하여 이미지나 텍스트를 조작하는 것을 막기 위해, 안전성 검사 방법론으로는 적대적 샘플링 방법을, 방어 전략으로는 적대적 예방 학습 방법을 사용할 수 있다.

ii. 기존의 AI 모델에 대한 안전성을 평가하고 강화: 이를 위해 기존의 AI 모델에 대한 취약성을 분석하고, 그에 따른 새로운 방어전략을 개발하는 것이 중요하다. 예를 들어, 이미지 분류 모델의 취약성을 분석하여 그에 대한 방어 전략을 개발하는 것이 있다.

그림 1과 같이 머신러닝에서 적대적인 공격이 무엇이며 이를 어떻게 방어할 수 있는지에 대한 내용을 도식화하고 있다.

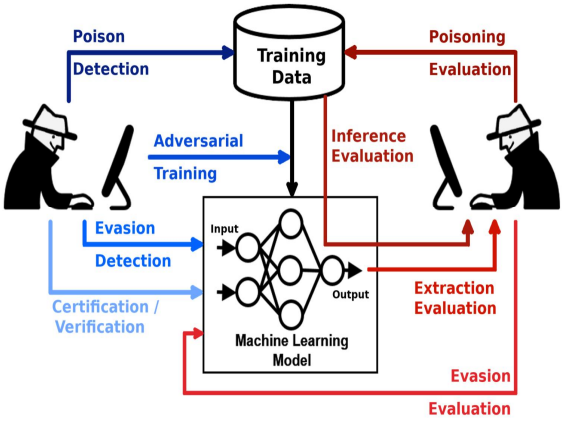

Fig. 1. Ritika. What are adversarial attacks in machine learning and how to prevent them?

V. Conclusions

AI 보안은 AI모델을 보다 안전하게 사용하고 보호하기 위한 기술 분야이다. 이 분야에서 다양한 연구들이 이루어지고 있으며, 이를 통해 AI 모델을 안전하게 사용할 수 있는 기술들이 개발되고 있다. Adversarial Machine Learning 분야에서는 적대적 공격에 강건한 AI 모델을 개발하기 위한 다양한 방법들이 제시되고 있다. 이를 통해 적대적 공격에 대한 대처 기술이 발전하고 있으며, 안전하고 신뢰성 높은 AI 시스템 개발에 큰 기여를 할 것으로 예상된다. AI 모델 해킹과 보안 문제에 대한 대응책으로는 모델 검증 및 모델 해킹에 대한 대응 기술이 연구되고 있다. 이를 통해 AI 모델의 안전성과 신뢰성을 검증하고, 해킹에 대한 대응 기술을 개발하여 보다 안전하게 사용할 수 있는 AI 모델을 만들기 위한 노력이 이루어지고 있다. AI 서비스에서의 보안과 프라이버시 문제는 매우 중요한 문제다. 이에 대한 대응책으로는 데이터 프라이버시 보호와 모델 보안 기술이 개발되고 있다. 또한, 페더레이션 학습과 같은 새로운 학습 방법이 제시되어 데이터 보안과 프라이버시를 보호하는 기술적인 문제들에 대한 대처가 이루어지고 있다.

마지막으로, AI 보안 분야에서는 향후 연구 방향으로 AI 모델의 안전성과 신뢰성을 보장하기 위한 보안 기술 및 대응책에 대한 연구가 계속 이루어져야 한다. 또한, 보안과 프라이버시를 보호하면서도 AI 모델의 성능을 개선할 수 있는 방법들이 개발되어야 한다. 종합적으로, AI 보안 분야에서는 다양한 문제들에 대한 대응책과 기술적인 해결책이 제시되고 있으며, 이를 통해 AI 모델을 보다 안전하게 사용할 수 있도록 하는 노력이 계속해서 이루어져야 할 것으로 보인다.

REFERENCES

[1] “Adversarial Machine Learning: A Survey” by Bo Li, Yevgeniy Vorobeychik, and Xiangyang Li (2018)
[2] “Deep Learning with Differential Privacy” by Martin

Abady, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang (2016)

[3] “Membership Interface Attacks Against Machine Learning Models” by Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami (2017)

[4] “Privacy-Preserving Deep Learning” by Yang Zhang, Shuang Wang, and Yongdong Wu (2018)

[5] “A Survey of Deep Learning for Scientific Discovery” by Maithra Raghu, Eric Schmidy, and Behnam Neyshabur (2019)

[6] “Federated Learning: Strategies for Improving Communication Efficiency” by Jakub Konečnỳ, Brendan Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon (2016)

[7] "Differentially Private Learning with Shadow Data" by Kamalika Chaudhuri and Anand D. Sarwate (2014)

[8] "Towards Federated Learning at Scale: System Design" by H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas (2019)

[9] "A Survey on Deep Learning for Named Entity Recognition" by Chenliang Li, Zhenghao Liu, and Wenjie Li (2020)