|  |  |  |
| --- | --- | --- |
| **클라우드 기반 머신러닝 서비스 보안 프레임워크** | | |
|  | | |
| **영문제목** | | |
|  | | |
| **요 약**  AI 기술이 적용된 서비스 제공에 강제되는 높은 메모리 사용량을 해결하기 위해 일반적으로 클라우드 컴퓨팅 기술을 이용한다. 클라우드 기반 서비스는 애플리케이션 개발자로 하여금 메모리 사용량에 대한 걱정을 덜어주어 성능적인 부분을 좀 더 신경 쓸 수 있게 하며 이 용자는 편리하게 양질의 서비스를 제공받을 수 있게 한다. 하지만 보안 대책이 미흡한 클라우드 서비스는 서비스를 제공받아 얻는 이익만을 생각하기에는 보안사고로 인한 피해가 막대할 수 있다. AI 기술이 인간의 삶에 깊이 파고든 현 상황에서 우리가 사용하는 AI 기술이 적용된 애플리케이션 그 중에서 많은 부분을 차지하고 있는 클라우드 기반 애플리케이션의 보안은 그 중요도가 높다고 할 수 있다. 이를 위해 본 논문에서는 클라우드 기반 AI 서비스를 분석하여 어떤 공격이 이루어질 수 있는지 분석하고 그에 대한 연구된 방어법들의 효과를 확인하여 효과적인 것들을 선별하고 접목시키는 시도를 한다. | | |

**1. 서 론**

인공지능(AI)에 대한 활발한 연구를 통해 비약적인 기술 발전이 이루어졌다. 여러 기업들은 모델을 스스로 훈련(train)시킬 능력이 부족한 사용자에게 기업의 모델을 서비스로 제공하는데, 이 때 해당 모델의 파라미터(parameter)나 훈련 당시 사용한 데이터(data) 등 모델에 대한 정보는 사용자로부터 공개되지 않지만, 사용자는 모델을 이용하여 입력(input) 값에 대한 출력(output) 값을 얻을 수 있고, 이를 이용해 모델의 기밀성을 저하시킬 수 있다. 모델의 기밀성이 낮아지면 모델을 복사(copy)하여 비슷한 성능의 모델을 만들어 기업의 경제 활동에 피해를 입힐 수 있고, 모델의 파라미터(parameter)를 손상시키거나 원래의 모델에 악의적인 입력 값을 넣음으로써 모델의 성능을 떨어뜨릴 수 있다. 또한, 모델을 서비스하는 과정에서 많은 메모리 사용은 불가피하기 때문에 기업에서는 클라우드 서비스를 많이 이용한다. 하지만 클라우드 서비스를 이용함으로써 여러 보안 문제가 발생할 수 있다.

본 논문에서는 먼저 AI 모델이 갖는 여러 취약점들을 다양한 논문들로부터 찾아 분석하고, 그 취약점들을 이용하여 실제로 공격이 가능한지 실험한다. 실험의 결과 실제로 공격이 제대로 이루어진다면 해당 취약점을 방어할 수 있는 방어법을 다른 논문으로부터 발표되거나 해당 취약점의 알고리즘을 분석하여 알아내고 그 방어법을 직접 적용해 봄으로써 그 효과를 확인한다. 마지막으로, 앞선 취약점들을 방어할 수 있는 방어법을 적용한 하나의 프레임워크를 제시한다.

**2. AI 모델 취약점**

본 논문에서 다룰 취약점은 2가지가 있으며, 그 종류는 유사 데이터(surrogate data)를 이용한 모델 하이퍼 파라미터 탈취 공격(model hyperparameter stealing attack) [1], 클라우드 서버 관리자에 의한 개인 데이터(private data) 탈취 공격[2]이 있다.

**2.1 모델 하이퍼파라미터 탈취 공격(model hyperparameter stealing attack)**

모델 하이퍼파라미터 탈취 공격 [1]을 이용해 모델의 정보를 도용하게 되면 학습자의 지적 재산권과 알고리즘의 기밀성이 저하되고, 이후 공격자가 회피 공격(evasion attack), 반전 공격(inversion attack) 등의 공격을 수행할 수 있다. 이 때, 유사 데이터(surrogate data)란 예를 들어 강아지의 종을 분류하는 모델이 있을 때, 공격을 위해 강아지가 아닌 고양이의 사진을 입력으로 넣는데, 이 때 고양이의 사진이 유사 데이터이다. 고양이의 사진을 입력 값으로 넣어 나온 출력 값으로 새로운 모델을 만들 수 있고, 이 모델과 원래의 모델인 강아지의 품종을 분류하는 모델이 비슷한 성능을 낼 수 있다는 것이 알려져 있다 [1].

3. 각 취약점에 대한 대응 방안

본 장에서는 2. AI 모델 취약점에서 소개한 공격기법들에 대해 본 논문에서 채택한 대응방안에 대해 소개한다.

3.1 신뢰 실행 환경[Trusted Execution Environment] 클라우드 서버 관리자에 의한 데이터 탈취 공격에 대한 대응법으로 신뢰 실행 환경을 제시할 수 있다.

**[학술대회 논문작성시 유의사항]**

1. 논문 페이지 수

- 참고문헌, 부록 포함하여 2~3쪽

2. 용지 및 여백처리

- 용지: A4 세로방향

- 여백: 위 쪽 30mm, 아래 쪽 20mm,

왼 쪽 10mm, 오른 쪽 10mm

3. 논문구성

- ①∼③항목은 1 단(column)으로 구성

1. 국문제목
2. 영문제목

③ 요약

- ④∼⑦항목은 2 단으로 구성

④ 본문

* 장 및 절에 해당되는 번호는 아라비아 숫자로

각각 1., 1.1 등과 같이 표기

* 그림의 명칭은 하단에, 표는 상단에 각각

그림 1 및 표 1로 표기

⑤ 참고문헌

* 본문중에 참고문헌 번호를 쓰고, 그 문헌을

참고문헌란에 인용한 순서대로 기술.

* 기술 순서는 저자, 제목, 학술지명, 권, 호,

쪽수, 발행년도 순으로 작성.

⑥ 부록(해당사항이 있는 경우만 작성)

⑦ 학부생/주니어 논문은 필요 시 시연동영상 URL

기재 가능.

4. 기타

- 논문작성폰트 임의사용 가능, 글자크기는 9pt이상

- 파일은 PDF로 제출 권장합니다.

- 사사문구는 추후 출판용 제출 시 기재합니다.

- 샘플논문(KCC2012발표):

<http://www.kiise.or.kr/conference02/data/sample1.pdf>

- 논문심사는 저자와 심사위원 상호 비공개로 진행됩니다. 따라서, 심사용(저자정보 삭제)과 출판용(저자정보 포함)으로 나눠 제출합니다. 최초 투고 시에는 심사용을, 최종본 제출 기간에는 출판용을 업로드 하시면 됩니다.