AI 모델 탈취 공격 및 방어 기법들에 관한 연구

전소희*, 이영한*, 김현준*, 백윤흥*
*서울대학교 전기 · 정보공학부, 반도체공동연구소

shjun@sor.snu.ac.kr, yhlee@sor.snu.ac.kr, hjkim@sor.snu.ac.kr, ypaek@snu.ac.kr

A Study of AI model extraction attack and defense techniques

So-Hee Jun*, Young-Han Lee*, Hyun-Jun Kim*, and Yun-Heung Paek*
*Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center (ISRC),

Seoul National University

요 약

AI (Artificial Intelligence)기술이 상용화되면서 최근 기업들은 AI 모델의 기능을 서비스화하여 제공하고 있다. 하지만 최근 이러한 서비스를 이용하여 기업이 자본을 투자해 학습시킨 AI 모델을 탈취하는 공격이 등장하여 위협이 되고 있다. 본 논문은 최근 연구되고 있는 이러한 모델 탈취 공격들에 대해 공격자의 정보를 기준으로 분류하여 서술한다. 또한 본 논문에서는 모델 탈취 공격에 대응하기 위해 다양한 관점에서 시도되는 방어 기법들에 대해 서술한다.

1. 서론

최근 AI (Artificial Intelligence)에 대한 연구가 활발히 진행되면서, AI 모델을 통해 서비스를 제공하는 사업 또한 증가하는 추세이다. AI 모델은 데이터에 기반하여 데이터가 가지고 있는 특정 패턴에 대한 학습을 통해 정해진 작업을 수행한다. AI 모델의 학습이 데이터에 의존하기 때문에, 이러한 데이터의 양과 질이 AI 모델의 작업 성능에 큰 영향을 끼칠 수 있다. 하지만, 일반적인 상황에서 질 좋은 대량의 데이터를 수집하기 위해서는 많은 노력이 필요하다. 즉, 일반적인 상황에서 성능이 매우 높은 AI 모델을 학습하기 어렵다는 것을 의미한다. 그렇기 때문에 자본을 가진 여러 기업들이 자사가 수집한 대량의 데이터를 통해 AI 모델을 훈련시켜 고객에게 API 형태로 AI 모델 서비스를 제공하는 사업이 증가하고 있다. 이를 MLasS (Machine Learning as Service)라고 하며, Amazon, Microsoft 등과 같은 세계적인 대기업과 여러 중소기업에서 서비스를 진행하고 있다.

최근, 기업이 자원을 들여 학습시킨 AI 모델을 탈취하는 공격에 대한 연구가 활발히 진행되고 있다. 이는 공격자가 AI 모델을 탈취하는 것에 그치는게 아니라, 탈취한 AI 모델을 활용하여 추가적인 공격을 수행할 수 있기 때문에 큰 위협이 될 수 있다.

본 논문에서는 최근 AI 모델 탈취 공격 및 방어에 대한 동향을 서술한다.

2. AI 모델 탈취 공격 연구

AI 모델 탈취 공격은 AI 모델을 탈취하려는 공격자의 공격 대상이 되는 AI 모델에 대한 사전 지식 여부에 따라 분류할 수 있다. 여기서 사전 지식이란, 공격 대상이 되는 AI 모델의 구조, 학습 데이터 등이 될 수 있다. 본 논문에서는 공격에서 가정되는 공격자의 사전 지식의 양에 따라, 공격을 블랙박스 AI 모델 탈취 공격, 그레이박스 AI 모델 탈취 공격으로 분류한다.

2.1 그레이박스 AI 모델 탈취 공격

공격자가 공격 대상이 되는 AI 모델에 대하여 부분적인 사전 지식을 갖고 있을 때, 수행되는 공격이 그레이박스 AI 모델 탈취 공격에 속한다. 대표적인 공격으로는 [1]과 [2]가 있다.

[1]은 공격 대상이 되는 AI 모델에게 데이터를 미세하게 변화시키면서 반복적으로 데이터를 보내 돌아오는 결과값이 변화하는 양상을 통해 AI 모델 탈취 공격을 수행하는 방법을 제안하였다. 이 방법을 통해 공격 대상이 되는 AI 모델의 가중치 값을 정확하게 알아낼 수 있다는 것을 보여주었다. 하지만, 제안된 방법을 수행하기 위해서는 공격 대상이 되는 AI 모델 구조에 대한 사전 지식이 요구되며 공격 대상이 되는

AI 모델에게 많은 데이터를 보내야 한다.

[2]는 공격 대상이 되는 AI 모델에게 인위적으로 생성된 데이터를 보내 돌아오는 결과값을 토대로 데이터셋을 구성하여 공격자의 AI 모델을 학습시켜 AI 모델 탈취 공격을 수행하는 방법을 제안하였다. 제안된 공격 방법은 AI 모델의 결정 경계선 (Decision boundary) 근처에 있는 데이터가 AI 모델의 결정 경계선에 많은 정보를 내포하고 있다는 점을 활용하였다. 제안된 방법은 공격자가 공격 대상이 되는 AI 모델이 학습한 데이터의 도메인에 대한 사전 지식을 갖고, 이와 같은 도메인에 속하는 적은 양의 데이터를 기반으로 인위적인 데이터를 생성하여 [1]보다 비교적 적은 데이터를 보내어 효율적으로 AI 모델 탈취 공격을 수행할 수 있음을 보여주었다.

2.2 블랙박스 AI 모델 탈취 공격

공격자가 공격 대상이 되는 AI 모델에 대하여 어떠한 사전 지식도 갖지 않을 때, 수행되는 공격이 블랙박스 AI 모델 탈취 공격에 속한다. 대표적인 공격으로는 [3]과 [4]가 있다.

[3]은 공격자가 공격 대상이 되는 AI 모델에 대한 아무런 사전 지식도 갖지 않기 때문에 일반적으로 쉽게 구할 수 있는 공공의 데이터를 사용하여 AI 모델 탈취 공격을 수행하였다. 제안된 공격 방법은 공공의 데이터를 부분 집합으로 나누어 이 부분 집합을 반복적으로 공격 대상이 되는 AI 모델에 보내어 돌아온 결과를 토대로 데이터 셋을 구성하여 공격자의 AI 모델을 훈련시킨다. 이 때, 효율적인 AI 모델 탈취 공격을 위하여 공격자의 AI 모델의 결과를 바탕으로 공공 데이터의 부분집합 중에서 공격 대상이 되는 AI 모델에 보낼 데이터를 선택한다. 제안된 방법은 공격 대상이 되는 AI 모델에 대한 사전 지식 없이도 AI 모델 탈취가 가능하다는 것을 보여주었다. 또한, 공격자가 탈취한 공격에 대하여 Adversarial example 을 생성하고 이를 탈취 공격 대상이었던 AI 모델에 보내어 AI 모델 탈취 공격뿐만 아니라 우회 공격 (Evasion attack) 또한 수행 가능함을 보여주었다.

[3]이 공공 데이터를 사용하여 블랙박스 AI 모델 탈취 공격에 성공하였다면, [4]는 공격 대상이 되는 AI 모델이 학습한 데이터 도메인에 속하는 데이터를 생성해 블랙박스 AI 모델 탈취 공격을 수행하였다. 제안된 공격 방법은 AI 모델은 학습 데이터와 같은 도메인에 속하는 데이터에 대해서 높은 Confidence 를 갖는 결과값을 갖는다는 점을 활용하였다. 생성 모델 (Generative model)을 사용하여 공격 대상이 되는 AI 모델이 높은 Confidence 를 갖는 데이터를 생성하였고 생성된 데이터를 바탕으로 공격자의 AI 모델을 학습 시켜 높은 성능을 보여주었다.

3. AI 모델 탈취 방어 연구

AI 모델 탈취 공격이 활발하게 연구되면서 그에 따른, AI 모델 탈취 방어 연구 또한 활발하게 진행되고 있다.

[5]는 AI 모델 탈취 공격자가 데이터를 미세하게 변경해가며 데이터를 보낸다는 점을 활용하여 보호 대상이 되는 AI 모델에게 미세한 차이를 보이는 데이터가 지속적으로 들어오면 AI 모델 탈취 공격으로 탐지하는 방어 방법을 제안하였다. 제안된 방법은 방어 대상이 되는 AI 모델에게 입력되는 데이터 간의 거리를 계산하고 최소 거리들의 분포가 정규 분포로부터 벗어난 정도에 기반하여 공격을 탐지한다. 제안된 방어 방법을 통해 [1], [2]을 방어할 수 있음을 보여주었다.

[6]은 AI 모델 탈취 공격이 공격 대상 AI 모델이 학습하지 않은 분포에 속하는 데이터를 사용한다는 점을 활용하여 보호 대상이 되는 AI 모델에 이상 분포에 속하는 데이터가 들어오면 공격자의 데이터라 판단하고 모델의 결과값에 변형을 주어 결과를 보낸다. 결과값을 변형시킬 때, 정상적으로 학습된 AI 모델의 결과값과 일부러 잘못된 방향으로 학습시킨 AI 모델의 결과값을 활용하기 때문에 공격자가 정상적으로 학습된 AI 모델의 실제 결과값을 유추하기 어렵게 만들어 방어하였다.

4. 결론

본 논문에서는 이미 학습되어 상용화되고 있는 AI 모델을 악의적인 목적으로 탈취하는 AI 탈취 공격에 대한 공격 및 방어 기법에 대해 서술하였다. AI 모델이 상업적으로 사용되면서 그에 대한 위협도 증가하고 있다. AI 모델을 서비스화하는 사업은 앞으로도 크게 증가할 것이며 안전한 서비스 제공을 위해서는 AI 모델 탈취 공격과 같은 여러 위협에 대한 연구 및 방어에 대한 연구가 지속적으로 진행되어야 한다.

5. ACKNOWLEDGEMENT

이 논문은 2021 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2020R1A2B5B03095204)을 받았으며, 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2018-0-00230, (IoT 총괄/1 세부) IoT 디바이스 자율 신뢰보장 기술 및 글로벌 표준기반 IoT 통합보안 오픈 플랫폼 기술개발 [TrusThingz 프로젝트]) 지원을 받았고, 2021 년도 BK21 FOUR 정보기술 미래인재 교육연구단 지원을 받았으며, 2021 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-01840, 스마트폰의 내부데이터 접근 및 보호 기술 분석)

참고문헌

[1] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. “Stealing machine learning models via prediction APIs.” In Proceedings of the 25th USENIX Conference on Security Symposium (SEC'16). USENIX Association, USA, 2016. 601–618.

[2] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. “Practical Black-Box Attacks against Machine Learning.” In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17). Association for Computing Machinery, New York, NY, USA, 2017, 506–519.

[3] Pal, S., Gupta, Y., Shukla, A., Kanade, A., Shevade, S., & Ganapathy, V. “ActiveThief: Model Extraction Using Active Learning and Unannotated Public Data.” Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 2020, 865-872.

[4] Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, Marius Popescu, “Black-Box Ripper: Copying black-box models using generative evolutionary algorithms”, Advances in Neural Information Processing Systems, 2020, 20120-20129.

[5] Juuti, S. Szyller, S. Marchal and N. Asokan, "PRADA: Protecting Against DNN Model Stealing Attacks," 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, 2019, pp. 512-527.

[6] S. Kariyappa and M. K. Qureshi, "Defending Against Model Stealing Attacks With Adaptive Misinformation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 767-775.