수직 연합학습에서의 백도어 공격 연구

조윤기[1], 김현준[1], 한우림[1], 백윤흥[1]
[1]서울대학교 전기정보공학과, 반도체연구소
ygcho@sor.snu.ac.kr, hjkim@sor.snu.ac.kr, rimwoo98@snu.ac.kr, ypaek@sor.snu.ac.kr

A Study on Backdoor Attack against Vertical Federated Learning

Yun-gi Cho[1], Hyun-jun Kim[1], Woo-rim Han[1], Yun-heung Paek[1]
[1]Dept. of Electrical and Computer Engineering and Inter-university Semiconductor Research Center, Seoul National University

요 약

연합학습(Federated Learning)에서는 여러 참가자가 서로 간의 데이터를 공유하지 않고 협력하여 하나의 모델을 학습할 수 있다. 그 중 수직 연합학습(Vertical Federated Learning)은 참가자 간에 동일한 샘플에 대해 서로 다른 특성(Feature)를 가지고 학습한다. 또한 서로 다른 특성(Feature)에는 입력의 라벨(Label)도 포함하기 때문에 라벨을 소유한 참가자 외에는 라벨 정보 또한 접근할 수 없다. 이처럼 다양한 참가자가 학습에 참여하는 경우 악의적인 참가자에 의해 모델이 포이즈닝 될 여지가 존재함에도 불구하고 수직 연합학습에서는 관련 연구가 부족하다. 포이즈닝 공격 중 백도어 공격은 학습 과정에 관여하여 특정 입력 패턴에 대해서 모델이 공격자가 원하는 타겟 라벨로 예측하도록 오염시키는 공격이다. 수직 연합학습에서는 참가자가 학습과 추론 모든 과정에서 관여하기 때문에 백도어 공격에 취약할 수 있다. 본 논문에서는 수직 연합학습에서의 최신 백도어 공격과 한계점에 대해 분석한다.

1. 서론

AI는 4차 산업혁명 시대의 핵심 키워드로서 우리의 삶을 크게 변화시키고 있다. 이러한 눈부신 AI의 발전의 원동력에는 엄청난 양의 수집 데이터가 기반이 된다. 대표적인 예시로 최근 많은 주목을 받고 있는 대화형 인공지능 ChatGPT가 있다. 해당 서비스의 기반이 되는 모델인 GPT-3는 45TB의 거대한 텍스트 데이터를 통해 학습되었다. 이렇게 막대한 양의 데이터를 수집하여 학습하는 경우 민감한 개인 데이터가 활용될 가능성이 있고, 이러한 정보들이 모델에 내재되어 악의적인 공격자가 이를 침해할 여지가 존재한다. 이에 따라, AI 모델 개발 및 서비스 과정에서 사용된 수많은 개인정보들이 노출 및 악용되는 것을 방지하는 Privacy-Preserving AI 기술은 고도화된 AI 기술에 발맞춰 필수적이다. 또한, 이는 데이터에 기반한 AI의 잠재력을 극대화하며 AI가 미래 우리의 삶에 지속적으로 혜택을 누릴 수 있게 하는 Enabler라고 할 수 있다.

Privacy-Preserving 기술 중 하나인 연합학습(Federated Learning)은 개인의 학습 데이터를 직접적으로 공유하지 않으면서, 하나의 모델에 학습한 것처럼 성능 향상을 얻을 수 있다는 장점이 있다. 이를 위해 메인 서버가 존재하고 개인 사용자들은 학습 데이터가 아닌 관련 (gradients, weights, latent representation)를 서버와 교류한다. 연합학습은 크게 두 가지 종류가 존재하는데 하나는 수평 연합학습(Horizontal Federated Learning)이고 다른 하나는 수직 연합학습(Vertical Federated Learning)이다. 수평 연합학습에서는 참가자들이 각자 다른 샘플을 가지되 동일한 특성(Feature)과 데이터에 대한 라벨(Label)을 가진다. 이에 반해 수직 연합학습에서는 참가자들이 동일한 샘플에 대해서 다른 특성(Feature)을 가지게 되며 라벨 정보 또한 특정 참가자만 소유하고 있다.

이러한 방식은 개인의 데이터를 보호할 수 있지만, 다양한 참가자가 모델 학습에 영향을 끼치기 때문에 일부 악의적인 참가자로 인해 모델의 무결성을 침해하는 모델 포이즈닝 공격에 취약하다. 이러한 공격들은 수평 연합학습(Horizontal Federated Learning)에서 많이 연구되었지만, 수직 연합학습(Vertical Federated Learning)에서는 상대적으로 고려되지 않

고 있다.

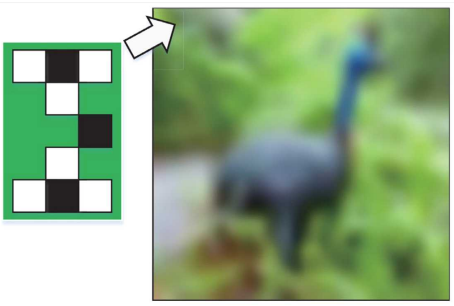
모델 포이즈닝 공격 중 백도어 공격은 특정 백도어 패턴을 모델이 입력 받으면 공격자가 의도하는 결과로 예측하도록 한다. 이를 위해 특정 백도어 패턴을 공격 타겟 라벨에 심는 과정이 필요하다. 수직 연합학습에서는 참가자가 라벨을 가지고 있는 경우에는 공격을 시도할 수 있지만, 더 어려운 가정인 라벨이 없는 참가자 경우를 고려한다. 또한 이와 같은 가정이 수직 연합학습의 목적 상 실제 환경과 훨씬 유사하다. 본 논문에서는 이와 같은 상황에서 최신 백도어 공격을 설명하고 한계점과 개선 방식들에 대해 분석한다.

2. 수직 연합학습 알고리즘

본 장에서는 수직 연합학습의 알고리즘 중 스플릿 러닝을 활용한 경우에 대해 서술한다. 학습은 다음과 같은 방식으로 진행된다[4]: 1)모델의 앞부분(입력단부터 중간단까지) 와 뒷부분(중간단부터 출력단까지) 나누어 모델의 앞부분은 각 참여자가 모델의 뒷부분은 서버가 가진다. 2) 각 참여자는 자신의 데이터를 모델의 앞부분을 통과시킨다. 그 결과, 참여자는 데이터에 대한 중간 표현(Latent Representation)을 가지게된다. 3) 생성된 중간 표현을 서버로 전송하고, 서버는 중간 표현을 이어(Concatenate) 뒷부분의 입력으로 사용한다. 4) 최종 출력과 서버의 라벨을 통해 그레디언트를 계산하여 업데이트하고 중간 표현에 대한 그레디언트를 해당하는 참여자에 돌려준다. 5) 각 참여자는 중간 표현 그레디언트를 이용하여 나머지 앞부분에 대한 업데이트를 마무리한다. 위 과정을 반복하여 모델을 학습할 수 있게된다.

3. 수직 연합학습에서의 백도어 공격

백도어 공격은 학습 데이터에 대상 라벨 데이터에 특정 입력 패턴을 섞고 추론 시 해당하는 입력 패턴에 대해서 해당 라벨로 추론하도록 한다. 이때 데이터의 본래 라벨과 관계없이 입력 패턴이 포함되어있을 경우 해당 라벨로 추론하도록 하는 것이 핵심이다. 기존 백도어 공격은 데이터셋의 특정 라벨의 데이터에 백도어 입력 패턴을 섞기 때문에 백도어 데이터가 모델 학습자의 데이터셋에 포함되도록 하는 가정이 필요하다. 그러나 수직 연합학습 같은 경우에는 데이터를 각 참여자가 가지고 있기 때문에 참여자가 의도적으로 모델에 백도어를 심을 수 있



(그림 1) 백도어 입력 패턴 (CIFAR10) [2]

다. 그러나 라벨을 알 수 없을 경우에는 백도어를 심을 데이터를 고를 수 없기 때문에 특정 수직 연합학습 상황에서는 어렵다.

이러한 상황을 해결하기 위해 LRBA[2]라고 불리우는 백도어 공격 방식이 제안되었다. 해당 공격 방식은 라벨 추론 공격[1]을 활용한다. LRBA는 수직 연합학습의 전체적인 학습이 끝난 뒤에 추가 학습을 통해 백도어를 심게 되는데, 자세한 공격 방식은 다음과 같다. 먼저 라벨 추론 공격을 통해 데이터셋의 라벨을 추론한다. 이 과정에서 자신이 소유하고 있는 모델의 앞부분 하단에 추론 모듈을 추가한다. 추론 모듈은 몇가지 레이어로 구성되며 데이터의 라벨을 추론하도록 한다. 이때 아주 최소의 데이터를 활용하여 추론 모듈을 준지도 학습방법으로 학습시킨다. 이후 학습된 추론 모듈을 통해 학습 데이터셋에 대해 라벨을 추론할수 있다. 2) 다음으로 백도어 공격으로 활용할 중간 표현을 생성한다. 이때 사용하는 수식은 다음과 같다.

$$H^* \leftarrow H^* - \eta \nabla_{H^*}(L(I(\theta_I, H^*), y^*)) \quad (1)$$

H^*는 백도어 중간표현을 의미하고, I와 θ_I는 추론 모듈과 파라미터를 의미한다. y^*는 백도어 타겟 라벨을 의미한다. 위의 수식을 통해 중간 표현을 업데이트하면 중간 표현이 타겟 라벨과 유사하도록 최적화된다. 이때 최적화 방식으로는 Stochastic Gradient Descent 가 사용된다. 3) 다음으로는 공격자가 자신의 모델 앞부분을 특정 입력 패턴에 대해 백도어 중간 표현을 출력하도록 재학습시킨다. 동시에 재학습 과정에서 백도어가 아닌 정상 입력에 대한 출력은 변화를 줄이려고 한다. 수식은 다음과 같다.

$$\min_{\theta^*} L(\theta^*;D,D^*) = \frac{1}{N}\sum_{x \in D} \| B(\theta^*,x) - B(\theta,x) \|_2 + \frac{1}{N}\sum_{x \in D^*} \| B(\theta^*,x) - H^* \|_2 \quad (2)$$

이때 B, D, D^*는 각각 모델 앞부분과 정상 입력, 백도어 입력을 의미한다. 이 과정을 통해 공격자의 모델 앞부분은 백도어 입력에 대해서는 타겟 라벨과 유사한 중간 표현을 생성하게 된다.

4. 한계점과 개선 방안

수직 연합학습에서 백도어 공격은 현재까지 여러 개선사항이 존재한다. 본 장에서는 개선사항들을 소개하고 이에 대한 연구방향성을 제시한다. 먼저 라벨이 없는 참가자의 경우 백도어를 삽입해야할 데이터를 특정할 수 없다. 이런 문제를 해결하기 위해 라벨 추론 공격을 도입했을 경우, 라벨이 없다는 상황을 라벨이 있지만 일정 노이즈가 섞여있는 상황으로 바꾸어줄 수 있지만, 노이즈가 백도어 공격에 어느정도 영향을 줄 수 있는 지에 대한 고려가 필요하다. 또한 현재 LRBA[2]는 모델의 뒷부분이 아닌 앞부분에만 백도어를 심는다. 이에 따라서 학습의 참여자가 늘어나 공격자의 영향력이 줄어들 경우 공격 성공률이 현저하게 줄어들 것이다.

라벨 노이즈 라벨 노이즈가 섞여있는 데이터셋에 대한 학습은 정통적인 딥러닝 분야의 챌린지 중 하나이다. 실제로 데이터셋에 약간의 노이즈 라벨의 존재가 모델 성능의 큰 하락을 발생시킨다는 것이 입증되었다. 그러나 현재 수직 연합학습의 백도어 공격에서는 이러한 노이즈 라벨이 존재함에도 불구하고 고려가 되지 않고 있다. 실제로 LRBA[2]에서는 추론 모델을 직접 이용할 분 추론 모델에서 생기는 노이즈에 대해서는 따로 처리하지 않는다. 이런 부분을 개선한다면 백도어 공격 성능이 더욱 강해질 것이다.

전체 모델에 대한 공격 현재 수직 연합학습의 백도어 공격에서는 모델 전체가 아닌 앞부분에 대한 제한적인 공격을 진행한다. 물론 앞부분의 변화만으로 훌륭한 공격 성능을 보여줬지만, 실제적으로는 뒷부분이 변화하지 않는다. 이와 달리 기존 백도어 공격 시 패턴이 모델 전체에 심어진다[3]. 따라서 학습 참여자가 늘어날 경우 공격 성능이 급감하는 이유는 일부 모델에만 백도어가 심어지기 때문일 수 있다. 이를 해결하기 위해 모델 전체에 백도어 패턴을 심을 수 있는 방안이 필요하다.

5. 결론

현재 수직 연합학습에서는 수평 연합학습과 달리 보안 관점에서의 연구가 많이 진행되지 않았다. 본 논문에서는 수직 연합학습에서의 보안 이슈 중 백도어 공격에 대해 다루었다. 이 중 현재 수직 연합학습에서의 백도어 공격은 여러 문제점을 가지는데, 각 문제점들에 대해 분석하고 이에 대한 해결 방안을 제시하였다. 이러한 연구들을 통해 수직 연합학습에서의 백도어 공격 취약점을 분석하게되고, 결과적으로 여러 위협들에 대한 예비책을 수립하는데 도움이 될 것 이다.

5. ACKNOWLEDGEMENT

이 논문은 2023년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음" (IITP-2023-2020-0-01602). 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2023-0-00516, 국가통계데이터에 적용 가능한 차등 정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결)

참고문헌

[1] Fu, Chong, et al. "Label inference attacks against vertical federated learning." 31st USENIX Security Symposium (USENIX Security 22). 2022.

[2] Gu, Yuhao, and Yuebin Bai. "LR-BA: Backdoor attack against vertical federated learning using local latent representations." Computers & Security (2023): 103193.

[3] Hong, Sanghyun, Nicholas Carlini, and Alexey Kurakin. "Handcrafted backdoors in deep neural networks." Advances in Neural Information Processing Systems 35 (2022): 8068-8080.

[4] Wei, Kang, et al. "Vertical federated learning: Challenges, methodologies and experiments." arXiv preprint arXiv:2202.04309 (2022).