연합학습에서의 보안 취약점 분석

최지헌, 유미리, 윤대건, 오상윤*
아주대학교
unidev@ajou.ac.kr, zztiok@ajou.ac.kr, kljp@ajou.ac.kr, *syoh@ajou.ac.kr

Analysis on Security Vulnerability in Federated Learning

Jiheon Choi, Miri Yu, Daegun Yoon, Sangyoon Oh*
Ajou Univ.

요 약

개인 데이터에 대한 프라이버시 침해 없이 분산 기계학습을 구현하기 위해 연합학습이 제안되었다. 기존 연합학습 기법의 개선을 통해 정확도 향상 및 수렴속도 향상을 목표로 하는 새로운 기법들이 등장하고 있어서, 이에 대한 보안 가이드라인이 필요한 상황이다. 본 논문에서는 연합학습 구조의 특징으로 나타나는 보안 취약점을 공격형태 별로 구분하고 이에 대한 대응방안을 고찰한다.

Ⅰ. 서 론

개인 데이터에 대한 프라이버시 침해 없이 기계학습을 구현하기 위해 연합학습[1]이 제안되었다. 연합학습은 데이터를 모으지 않는 기본 구조에서 데이터의 프라이버시가 보장되는 구조이나, 각 단계 별로 보안 취약점이 존재한다. 예를 들어, 클라이언트가 보내는 gradient 들의 재조합을 통해서 원본 데이터를 추측할 수 있는 점은 연합학습을 제안한 구글의 연구팀에서도 인지된 부분이며, 이를 위해 secure aggregation 및 differential privacy 방식을 제안하였다. 새로이 제안되는 연합학습 기법들은 최초 제안된 FedAvg[1] 기법을 수정하여 더 빠르게 수렴하거나 더 높은 정확도를 얻고자 하며, 기존 기법의 수정을 통해 이 목표를 달성하려 한다. 따라서, 이러한 새로운 기법들에는 새로운 보안 취약점들이 발생할 수 있어 이에 대한 가이드라인이 필요하다. 본 논문에서는 연합학습 기법과 기존 연구 분석을 기반으로 연합학습에서 발생 가능한 보안 취약점과 대응방안을 고찰한다.

Ⅱ. 본론

연합학습에서 클라이언트는 로컬에서 DNN (deep neural network) 학습을 수행하고 학습 결과로서의 모델 파라미터를 중앙 서버로 전송하며, 중앙 서버는 다수의 클라이언트로부터 전송받은 모델 파라미터를 집계하여 글로벌 모델을 생성한다. 이후 생성된 글로벌 모델은 클라이언트로 반환되어 로컬 모델을 갱신한다.

본 구조의 모든 단계에서 악성 행위자에 의해 예상치 못한 결과를 얻을 수 있으며, 연합학습 시스템을 대상으로 하는 공격은 다음 3가지로 구분할 수 있다. 첫째, 글로벌 모델을 생성하기 위해 집계 연산을 수행하는 서버에 의해 발생하는 경우이다. 특정 클라이언트로부터 수신한 gradient로부터 학습 데이터를 추출하거나 악성 서버가 실제 집계 연산을 수행하는 서버인 것으로 속일 수 있다. 특히, 서버는 학습을 위한 라운드에 참여하는 모든 클라이언트의 gradient를 확인할 수 있어 공격이 발생할 경우 파급력이 높다. 관련되어 gradient 값을 통해 학습 데이터를 유사한 형태로 생성하거나 복원이 가능한 것을 보인 연구[2]가 소개되어 있다.

두 번째로, 글로벌 모델 생성을 위한 학습에 참여하는 클라이언트에 의해 보안 취약점이 발생하는 경우이다. 이는 서버에게 정상적인 클라이언트인 것으로 속여 다양한 공격을 수행할 수 있어 해결하기 어려운 과제이다. 예를 들어, attacker 클라이언트는 공격이 수행되는 동안 victim 클라이언트가 글로벌 모델에만 접근이 가능하도록 하여 로컬 모델을 이용한 학습을 불가능하게 한다. 이처럼 클라이언트에 의해 발생하는 공격은 앞서 설명한 서버 측의 취약점과 다르게 글로벌 모델의 가용성과 학습 데이터의 기밀성을 침해하는 것을 목표로 한다. 예를 들어, 가용성을 침해하는 공격은 이후 설명할 inference 공격이, 그리고 기밀성을 침해하는 공격은 model inversion 공격이 대표적이다. 아래 표는 클라이언트로 인해 발생할 수 있는 위협을 정리하였다.

Threat	Affected Phase
Poisoning Attack [3, 4]	Training
Inference Attack [5, 6]	Training / Inference
Model Inversion [7]	Inference
Free-riding Attack [8]	Training

1) Poisoning Attack

단일 머신에서 수행하는 기계학습에서도 발생할 수 있는 공격이다. 연합학습에서는 단일 클라이언트의 로컬 모델 학습 과정에서 악성 데이터를 주입하여 학습 결과를 오염시킨다. 이로 인해 모델의 무결성을 보장할 수 없는 경우이다. 공격의 목표에 따라, 데이터 수집 과정에서 발생하는 data poisoning 방법과 모델 학습 과정에서 발생하는 model poisoning 방법으로 나눌 수 있다. 두 가지 방법 모두 글로벌 모델을 조작하여 정확도를 낮춘다.

2) Inference Attack

추론 공격은 공격자가 연합학습 프로세스 중에 gradient에서 학습 데이터를 추론한다. 클라이언트와 정직하지만 호기심 많은 (honest-but-curios)한 서버 모두 다른 클라이언트의 학습 데이터에 대한 공격을 수행할 수 있다. 공격자는 각 라운드에서 집계된 글로벌 모델의 파

라미터의 스냅샷을 저장하고, 연속 스냅샷 간의 차이를 이용하여 추론 공격을 수행할 수 있다. 공격의 종류에는 membership inference attack [5], property inference attack이 있다.

3) Model Inversion

공격자는 GAN (Generative Adversarial Network)을 활용하여 victim의 샘플을 재구성한다. [7]

4) Free-riding Attack

Free-rider[8]는 학습에 기여하지 않는 클라이언트를 말한다. 예를 들어, 충분하지 않은 연산 성능이나 충분한 데이터를 가지고 있지 않은 경우가 될 수 있다. 공격자는 stochastic 업데이트를 기반으로 평균이 0이고, 표준 편차가 하나인 가우시안(gaussian) 노이즈를 사용한다. 결과적으로 이를 통해 글로벌 모델의 무결성을 침해하는 것으로 볼 수 있다.

세 번째로, 서버와 클라이언트의 통신 구간에서의 문제이다. 처음 제안된 연합학습의 구조에서는 통신 구간이 암호화되지 않아 중간자 공격(man in the middle attack)이 발생하였다. 이를 개선하기 위한 연구로 동형 암호(homomorphic encryption)를 이용한 secure aggregation 방법[9]에 대한 연구를 구글 연구팀에서 제안하였다. 이 방법의 주요한 특징으로는 연산과정에서 암호문 해독을 위해서 연산을 수행하지 않기 때문에 효율적인 글로벌 모델의 집계가 가능하다.

지금까지 설명한 보안 취약점을 개선하기 위한 대응방안이 적용된 연구를 아래 표에 정리하였다. 기존의 연구에서는 주로 기밀성(confidentiality)을 지키기 위하여 암호기술과 차분프라이버시(differential privacy) 기술, 가용성과 무결성을 지키기 위하여 robust aggregation[13], adversarial training 등의 기술이 활용된다.

Research	Countermeasure / Mitigation
[10]	differential privacy, secure multiparty computation
[11]	trusted execution environment (IntelSGX)
[12]	homomorphic encryption

Ⅲ. 결론

본 논문에서는 연합학습 구조를 서버와 클라이언트, 통신 구간으로 나누어서 각각에서 나타나는 보안 취약점과 기존 연구에서 제안된 대응방안을 고찰하였다. 연합학습의 분산 환경으로 인해 기존의 기계학습과 비교해서 견고성(robustness)과 정보보호 모두에 약점을 가지고 있음을 보여준다. 실제 산업에 연합학습을 적용함에 있어서 보안을 고려한 새로운 구조의 제안을 기대한다.

ACKNOWLEDGMENT

본 논문은 2023년 SW중심대학사업(2022-0-01077) 및 2022년도 한국연구재단 기본연구사업(NRF-2021R1F1A1062779)의 연구비 지원으로 수행하였습니다.

참 고 문 헌

[1] McMahan, Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. "Communication-efficient learning of deep networks from decentralized data." In Artificial intelligence and statistics, pp. 1273-1282. PMLR, 2017.

[2] Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." Advances in neural information processing systems 32 (2019).

[3] Fang, Minghong, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. "Local model poisoning attacks to Byzantine-Robust federated learning." In 29th USENIX Security Symposium (USENIX Security 20), pp. 1605-1622. 2020.

[4] Tolpegin, Vale, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. "Data poisoning attacks against federated learning systems." In European Symposium on Research in Computer Security, pp. 480-501. Springer, Cham, 2020.

[5] Hu, Hongsheng, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. "Membership inference attacks on machine learning: A survey." ACM Computing Surveys (CSUR) 54, no. 11s (2022): 1-37.

[6] Usynin, Dmitrii, Alexander Ziller, Marcus Makowski, Rickmer Braren, Daniel Rueckert, Ben Glocker, Georgios Kaissis, and Jonathan Passerat-Palmbach. "Adversarial interference and its mitigations in privacy-preserving collaborative machine learning." Nature Machine Intelligence 3, no. 9 (2021): 749-758.

[7] Zhang, Yuheng, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. "The secret revealer: Generative model-inversion attacks against deep neural networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 253-261. 2020.

[8] Fraboni, Yann, Richard Vidal, and Marco Lorenzi. "Free-rider attacks on model aggregation in federated learning." In International Conference on Artificial Intelligence and Statistics, pp. 1846-1854. PMLR, 2021.

[9] Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. "Practical secure aggregation for privacy-preserving machine learning." In proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175-1191. 2017.

[10] Truex, Stacey, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. "A hybrid approach to privacy-preserving federated learning." In Proceedings of the 12th ACM workshop on artificial intelligence and security, pp. 1-11. 2019.

[11] Mo, Fan, Hamed Haddadi, Kleomenis Katevas, Eduard Marin, Diego Perino, and Nicolas Kourtellis. "PPFL: privacy-preserving federated learning with trusted execution environments." In Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, pp. 94-108. 2021.

[12] Zhang, Chengliang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. "BatchCrypt: Efficient homomorphic encryption for {Cross-Silo} federated learning." In 2020 USENIX annual technical conference (USENIX ATC 20), pp. 493-506. 2020.

[13] Pillutla, Krishna, Sham M. Kakade, and Zaid Harchaoui. "Robust aggregation for federated learning." arXiv preprint arXiv:1912.13445 (2019).