연합학습 모델에 대한 공격 및 방어 기술 동향에 관한 연구

임채문, 박수현*, 김중헌
고려대학교, *숙명여자대학교

{anscodla0314, joongheon}@korea.ac.kr, *soohyun.park@sookmyung.ac.kr

A Study on the Attack and Defense Mechanisms Against Federated Learning Model

Chaemoon Im, Soohyun Park*, Joongheon Kim
Korea Univ., *Sookmyung Women's Univ.

요 약

연합 학습은 개별 사용자의 데이터를 직접 드러내지 않고서도 효율적인 학습을 가능하게 한다. 그러나 연합학습 알고리즘의 구조적 특징을 악용하여 연합학습의 보안성을 위협하는 다양한 공격 기법이 존재한다. 본 논문은 연합학습 모델에 대한 공격 기법 및 방어 기법에 대해 논하고, 연합학습 모델의 방어 기법에 대한 향후 연구 방향성을 제시한다.

Ⅰ. 서 론

연합 학습(Federated Learning)은 개별 모델의 데이터를 통해 중앙 모델을 학습하는 기계학습의 기법이다 [1]. 연합 학습은 그 구조상 개별 모델의 학습에 사용된 데이터가 외부로 유출되지 않으므로 데이터의 유출에 따른 보안 문제에서 비교적 자유롭다고 알려져 있다. 특히 보안이 더더욱 중요한 의료 데이터 등의 경우에 있어, 연합 학습은 개별 데이터의 유출을 최소화하면서도 다수의 로컬 모델 간 일반화 가능한 모델 학습을 가능하게 한다.

그러나 연합학습 역시 종래의 인공신경망 기반 인공지능 기술과 마찬가지로 외부의 공격에서 자유롭지 않다 [2]. 다양한 공격 기법들은 학습된 모델의 오작동을 유도하거나, 학습에 사용된 개별 사용자의 데이터를 역으로 유추하는 데까지 다양한 방식으로 연합학습의 구조상 보안성을 위협한다. 이러한 공격은 향후 연합 학습을 통한 서비스 제공 및 상용화에 있어 큰 장애물로 작용할 수 있다. 이러한 공격으로부터 연합학습 모델 및 학습 데이터의 보안을 유지하기 위해, 다양한 방어 기법이 고안된 바 있다.

본 논문에서는 연합학습 모델의 간략한 특성과, 해당 특성으로 인해 발생하는 다양한 공격 기법 및 그를 방지하기 위한 방어 기법에 대해서 논한다. 아울러 이러한 공격 및 방어 기법에 대한 고찰을 토대로 연합학습의 보안성 강화를 위한 향후 연구 방향성을 제시함으로써 논문을 맺는다.

Ⅱ. 본론

i) 연합학습 모델의 특징 및 취약점

연합 학습은 세 단계로 이루어진다. 첫 번째로 로컬 모델은 개별 학습 데이터를 사용해 학습을 진행한다. 두 번째로 각 로컬 모델은 학습을 통해 얻은 정보를 중앙 모델에 업로드한다. 마지막으로 중앙 모델은 개별 모델의 학습 정보를 통해 학습을 진행하고, 개별 모델은 중앙 모델의 파라미터를 다시 전송받아 첫 번째 단계를 반복한다. 공격자들은 연합학습 과정에서 학습 데이터를 통해 생성된 정보가 업데이트된다는 점을 이용하여, 해당 정보를 조작하거나 탈취함으로써 연합학습 모델에 공격을 가한다.

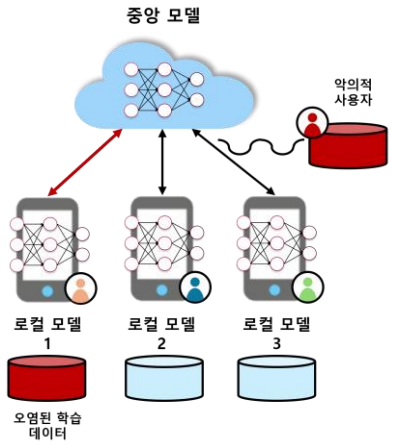

그림 1. 연합학습 공격 기법 모식도

ii) 연합학습 모델에 대한 공격 기법

인공 신경망 기반 인공지능은 학습을 위해 다수의 데이터를 필요로 하고, 이 때 데이터의 편향성은 곧 모델의 편향성을 좌우한다. 이러한 인공 신경망의 특징에 착안하여 등장한 Poisoning Attack 은 학습 데이터를 오염시켜 모델의 성능을 감소시키는 공격 기법이다 [3]. 악의적인 사용자는 연합학습에 사용되는 개별 데이터를 오염시키거나, 로컬 모델의 그래디언트에 노이즈를 추가하는 방식으로 중앙 모델의 학습을 방해하고 모델의 오류 강건성을 약화시킨다.

그래디언트 및 학습 정보를 공유한다는 연합학습의 구조적 특징으로부터 이점을 얻는 또다른 공격 기법으로는 Inference Attack 이 있다. 악의적인 사용자는 각 로컬 모델이 중앙 모델로 업로드하는 그래디언트를 탈취하여, 해당 그래디언트를 통해 개별 사용자의 사적 데이터를 추론할 수 있다 [4]. 해당 공격 기법은 개별 사용자의 데이터가 노출되지 않는다는 연합학습의 장점을 무력화시킨다.

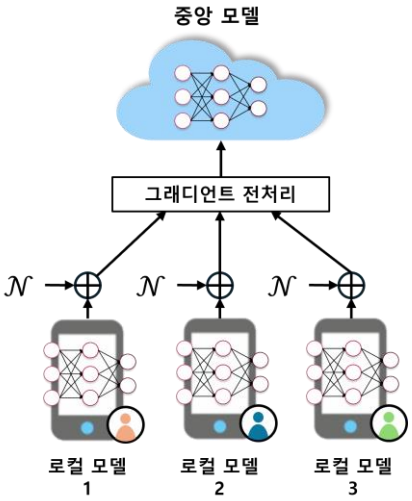

그림 2. 연합학습 방어 기법 모식도

iii) 연합학습 모델에 대한 방어 기법

연합 학습 모델을 대상으로 한 공격 기법에 대항하여, 다양한 방어 기법 또한 연구되고 있다.

모델의 견고성을 약화시키는 Poisoning Attack 에 대처하기 위해, 개별 사용자가 전송하는 그래디언트를 전처리하여 공격을 방어하는 알고리즘들이 연구되었다 [5], [6]. 해당 알고리즘들은 사용자로부터 수집된 그래디언트를 분석하여, 비정상적인 그래디언트를 걸러내어 모델의 학습 시 비잔틴 강건성(Byzantine Tolerance)을 높인다. 해당 알고리즘을 적용한 연합학습 모델은 일부 로컬 모델에서 발생하는 장애나 공격이 학습에 미치는 영향을 최소화하여 외부의 공격에도 학습이 원활하게 진행될 수 있도록 한다.

개별 학습자의 데이터 보안성을 지키기 위한 동형 암호(Homomorphic Encryption)나 SMC (Secure Multiparty Computation) 등 기존의 전통적인 암호화 기법들은 연합학습에 있어 그리 성공적이지 못했는데, 이는 연합 학습을 수행하는 사용자의 수가 증가함에 따라 계산의 복잡도가 상승하기 때문이다 [7]. 해당 문제를 극복하기 위한 보안 기법으로 Differential Privacy (DP)를 들 수 있다 [8]. DP 는 개별 사용자의 그래디언트를 의도적으로 변조함으로써 해당 사용자의 개인 정보를 식별할 수 있는 가능성을 낮추는 기법으로, 연합학습에서는 개별 로컬 모델의 그래디언트에 노이즈를 삽입하는 것으로 구현된다 [8]. DP 는 악의적인 사용자 뿐만 아니라 악의적인 중앙 모델이 있는 상황에서도 효율적으로 대응할 수 있다는 장점이 있다. 중앙 모델의 관리자가 악의적인 의도로 개별 사용자의 그래디언트를 활용해 Inference Attack 을 수행하려고 하더라도, 그래디언트에 삽입된 노이즈 때문에 개별 사용자의 원본 그래디언트를 식별할 확률이 감소하기 때문이다.

Ⅲ. 결론

본 논문에서는 연합학습에 있어 발생할 수 있는 공격 가능성 및 이에 따른 방어 기법에 대해 논의하였다. 인공신경망 기반 인공지능 기술이 사회 전면에서 빠르게 도입되는 현 상황에서, 인공지능의 보안을 유지하기 위한 기술 역시 심각하게 논의되어야 하는 상황이다. 그러나 다양한 보안 기법을 적용함에 있어, 사용자의 원본 학습 정보를 왜곡하면 왜곡할수록 모델의 보안성은 증대되지만 모델의 정확도 및 성능은 감소하게 된다 [9]. 이러한 보안성과 정확도 간 Trade-Off 역시 보다 안전한 연합학습을 달성하기 위해 극복해야 할 문제로 남아 있다.

ACKNOWLEDGMENT

본 연구는 2022 년 한국연구재단의 지원을 받아 수행됨 (NRF 2022R1A2C2004869). 본 논문의 교신저자는 김중헌임.

참 고 문 헌

[1] D. Kwon, J. Jeon, S. Park, J. Kim and S. Cho, "Multiagent DDPG-Based Deep Learning for Smart Ocean Federated Learning IOT Networks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9895-9903, October 2020.

[2] P. Kairouz, et al. "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*. vol. 14. no. 1-2. pp. 1-210, November 2021.

[3] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru and B. Li, "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," *In Proc. IEEE Symposium on Security and Privacy (SP)*, San Francisco, USA, May 2018, pp. 19-35.

[4] L, Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," *In Proc. Advances in Neural Information Processing Systems (NIS),* Vancouver, Canada, December 2019, pp. 14747-14756.

[5] S. Shen, S. Tople, and P. Saxena, "AUROR: Defending Against Poisoning Attacks in Collaborative Deep Learning Systems," *In Proc. Annual Conference on Computer Security Applications (ACSAC)*. Los Angeles, USA. December 2016, pp. 508-519.

[6] P. Blanchard, E.M.E. Mhamdi, R. Guerraoui and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," *In Proc. Annual Conference on Computer Security Applications (ACSAC)*, Long Beach, USA. December 2017, pp. 119-129.

[7] L. Lyu et al., "Privacy and Robustness in Federated Learning: Attacks and Defenses," *IEEE Transactions on Neural Networks and Learning Systems,* Early Access.

[8] A El. Ouadrhiri and A. Abdelhadi. "Differential Privacy for Deep and Federated Learning: A Survey," *" IEEE Access,* vol. 10, pp. 22359-22380, April 2020.

[9] M. Kim, O. Günlü and R. F. Schaefer, "Federated Learning with Local Differential Privacy: Trade-Offs Between Privacy, Utility, and Communication," *In Proc. IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, June 2021, pp. 2650-2654.