연합학습을 위한 클라이언트 데이터 보안 연구 동향 조사

손영진 [1], 박민정 [2], 채상미 [3*]
[1] 이화여자대학교 경영학부 박사과정
[2] 금오공과대학교 경영학과 교수
[3] 이화여자대학교 경영학부 교수

teumdal@ewhain.net, mjpark@kumoh.ac.kr, smchai@ewha.ac.kr

요 약

연합 학습(Federated Learning, FL)은 중앙 서버 없이 분산된 클라이언트들이 공동으로 모델을 훈련시키는 방식으로, 데이터를 로컬에서 학습시키기에 개인정보 보호의 이점을 제공한다. 그러나 연합 학습 환경에서도 여전히 데이터 보안을 위협하는 다양한 공격이 존재한다. 본 논문에서는 특히 개인 데이터 탈취와 관련된 개인 정보 보호, 보안을 주요 대상으로 공격기법과 대응 방안에 대한 연구를 소개하고 이를 통해 연합 학습에서 클라이언트 데이터 보호를 위한 지속적인 연구를 촉진하기 위한 기초를 제공한다.

1. 서론

연합된 클라이언트(엣지 노드)에서의 분산 기계 학습을 연합학습(federated learning)[1][2] 이라고 하며 연합 학습은 분산된 클라이언트들이 단일 글로벌 모델을 공동으로 훈련시킬 수 있는 새로운 패러다임을 제시한다. 연합학습에서 원시 데이터(raw data)는 여러 엣지 노드에서 수집되어 저장되며, 엣지 노드에서 중앙 저장소로 원시 데이터를 보내지 않고 분산된 엣지 노드 데이터로부터 머신러닝 모델이 학습된다. 엣지 장치의 컴퓨팅 리소스를 활용하여, 클라이언트 소유의 데이터에 직접 액세스하지 않고도 로컬 모델을 학습시키는 연합학습의 접근 방식은 데이터 보안의 관점에서 볼 때, 중요한 데이터 보호 메커니즘을 제공한다. 학습 시 클라이언트의 데이터는 로컬 환경에서 보호되며, 중앙 서버나 다른 클라이언트와 공유되지 않는다는 점에서 기본적으로 개인 데이터 보호의 이점을 가진다[3].

그러나 학습 과정의 반복, 모델 업데이트 교환 과정에서 악의적인 공격자에게 시스템이 노출될 수 있으며, 멤버십 추론 공격, GAN 기반 추론 공격과 같은 연합학습 환경에서 데이터 프라이버시 위협, 의도하지 않은 데이터 유출과 같은 보안 문제도 지속적으로 관심을 받고 있다. 연합학습의 환경도 클라우드 환경, IoT 기기 연합학습, 자율주행차 연합학습[4]과 같은 데이터 보안이 반드시 필요한 환경으로 점차 변화하고 있기에 데이터 보안을 중심으로, 현재의 연구 동향과 기술들을 살펴보고 연합학습의 보안 취약점 발굴 및 데이터 프라이버시 보장을 위한 연구를 지속하고자 한다.

2. 연합학습 환경에서의 클라이언트 데이터 보안 공격 기법

연합학습의 주요 문제로는 높은 통신 비용, 이기종 시스템, 이기종 데이터의 통계적 이질성과 데이터 보안이 존재한다. 본 논문에는 데이터 보안만을 살펴보았으며 특히 연합 학습에서의 개인 데이터 유노출과 관련된 개인 정보 보호, 보안을 주요 대상으로 한다.

2.1. 추론(Inference) 공격[5]

추론 공격은 훈련 데이터 세부 사항을 추론하는 공격으로 학습된 모델로부터 원본 훈련 데이터에 대한 정보를 추출하거나 근사화하는 방법을 사용한다.

2.1.1. 멤버십 추론(Membership inference) 공격[6]

기계 학습 모델이 특정 데이터 포인트를 학습에 사용했는지 여부를 추론하는 공격 유형으로 모델의 예측 또는 확률값을 분석하여 특정 입력 데이터가 학습 데이터셋의 일부였는지를 판단하여 특정 데이터 인스턴스가 훈련 데이터로 사용되었는지 여부를 판별하는 것을 목표로 한다. 멤버십 추론 공격의 성공과 연합학습 모델의 일반화 간극(훈련 정확도와 테스트 정확도 사이의 차이) 사이가 클수록 연합학습 모델은 멤

버십 추론 공격에 더 취약하다[7][8].

2.1.2. GAN (Generative Adversarial Network) 기반 데이터 추론 공격[9]

GAN은 두개의 신경망, 생성자(generator)와 판별자(discriminator)로 구성되어 있다. 생성자는 실제 데이터와 유사한 데이터를 생성하려고 시도하고, 판별자는 입력 데이터가 실제인지 생성자가 만든 가짜인지를 판별하려고 시도하면서 서로 경쟁적으로 학습한다. 공격자가 GAN을 사용하여 학습된 모델로부터 원본 훈련 데이터를 복구하거나 근사화하여 민감한 정보를 추출하는 모델 역설계(Reverse Engineering) 공격, 원본 데이터셋의 통계적 특성을 학습하여 원본과 유사한 데이터를 생성하는 데이터 추론(Inference) 공격이 있다. GAN과 다중 작업 판별자를 결합한 mGAN-AI라는 프레임워크를 통해 클라이언트 신원을 판별하여 클라이언트 수준의 개인정보 복구를 가능하게 한 방법도 존재한다.

2.1.3. FL 학습 환경 구현시 의도하지 않은 데이터 유출[10]

[7]은 연합학습의 모델 가중치만을 고려하여 데이터를 추론할 수 있음을 보여주었으며, [11]에서는 클라이언트와 서버의 통신을 도청하여 클라이언트 모델을 재구성하는 방법으로 클라이언트가 서버로 전송한 모델 그라디언트를 반전시켜 로컬 데이터 수집을 할 수 있는 가능성을 보여주었다. 연합 학습은 원시 데이터 대신 기울기 정보, 계산된 최종 파라미터 값과 같은 모델 업데이트를 공유함으로써 각 장치에서 생성된 데이터를 보호하나 훈련 과정 전반에 걸쳐 모델 업데이트를 전달하면서 민감한 정보가 제3자나 중앙 서버에 공개될 수 있다[6].

중앙집중식 연합 학습 환경에서는 악의적이거나 승인되지 않은 사람의 중앙 서버 공격을 통해 클라이언트가 보낸 모든 로컬 모델을 가로채 클라이언트 원본 데이터의 일부를 재구성[10] 할 가능성도 존재한다.

3. 연합학습 환경에서의 클라이언트 데이터 보호 기법

3.1. Differential privacy (DP)[12]

DP(차등정보보호)는 데이터의 속성에 임의의 노이즈를 추가하여 각 사용자의 개인 정보를 보호한다[13]. 연합학습에서는 이에 더해 역 데이터 검색을 피하기 위해 참가자가 업로드한 매개변수에 노이즈를 추가하는 방식이 도입되었다.

3.2. SMC(Secure Multi-party Computation)[12]

다중 참가자가 모델이나 함수를 공동으로 계산하는 동안 입력을 보호하기 위해 도입되었으며 암호화 방식으로 통신이 보호된다. 연합학습에서는 입력 값인 클라이언트 업데이트를 보호하는 데 활용되고 있다.

3.3. Homomorphic Encryption (HE)

동형암호는 암호화 전의 연산과 암호화 상태의 연산 결과가 복호화시 같다는 동형성 특징을 지니고 있으며 사용자는 암호화한 데이터를 서버에 전송하고 그 연산 결과를 돌려받아 해독하여 값을 확인할 수 있다. 다른 암호들과 다르게 연산 과정 중 데이터가 외부에 노출되지 않아 개인 정보가 완전히 보호될 수 있다[14]. 다만 동형암호는 계산 시 높은 컴퓨팅 능력 및 통신 오버헤드를 요구하기 때문에 연합학습 기반 시스템에 적용하기 어려워 이러한 오버헤드를 줄이기 위한 연구[15]가 진행이 진행되었으며 추가적으로 보안성을 높이기 위해 [16]는 연합학습의 각 클라이언트가 서로 다른 동형암호 개인 키를 사용하는 분산 암호화 시스템을 제안하였다.

3.4. Trusted execution environment (TEE)

신뢰실행환경은 다른 프로세스나 내부 운영체제의 접근을 차단하여 기기 내부에 로드되는 사용자의 코드와 데이터를 보호하는 메인 프로세서의 보안 영역이다[17]. 연합학습에서 TEE는 연합학습에 사용되는 기기를 보호하여 신뢰를 구축하며, 이는 공격자가 개인 정보에 접근하는 것을 효과적으로 방지한다.

3.5. Adversarial training[18]

적대적 학습 공격은 기계 학습 모델에 적대적인 샘플을 주입하여 모델을 속이는 것을 목표로 한다. 공격자는 원래의 입력 데이터에 작은 변화를 가하여 모델을 오분류하게 만드는 데이터를 삽입하여 연합학습 모델의 견고성에 영향을 미치도록 한다. 이러한 공격을 막기 위해 연합학습 모델 훈련 단계 시작부터 공격의 모든 순열을 미리 시도하여 알려진 적대적 공격에 대해 대비한다.

3.6. Blockchain[19]

블록체인은 위변조 방지, 공동 유지보수, 추적성 등의 기능을 통해 중앙 서버를 대체하여 완전한 탈중앙

화 연합학습을 구현하고 협력 환경에서 데이터 세트 기여에 대한 인센티브를 제공할 수 있다. 암호화 기술을 결합하여 검증 가능한 계산의 실현 및 기존 연합학습에서는 해결하기 힘든 보안 및 개인 정보 보호 위협에 대응이 가능하다. 이러한 장점으로 최근 연합학습에서 블록체인 도입의 연구가 증가하고 있다.

4. 결론 및 향후 연구방향

GDPR과 같은 데이터 보호법의 도입은 연합학습에서 클라이언트 데이터 보호의 중요성을 강조하고 있다. 연합학습 방법론이 데이터의 프라이버시를 보호하면서 AI 알고리즘을 위한 학습을 가능하게 하는 방법인 만큼 최근의 연구 동향은 학습능력 강화 및 성능 강화를 위한 연구가 주를 이루고 있다. 그러나 최근 기기의 이질성, 데이터의 이질성 및 클라우드 환경으로 발전하는 IT 산업의 방향을 비추어 볼 때, 이러한 다양한 환경에서 발생할 수 있는 보안 취약점 및 데이터 프라이버시 보장을 위한 연구가 필요하다. 이는 연합학습 방법론이 주목받고 있는 고유의 목적을 보존하기 위한 필수적인 연구이며 향후 연합학습의 실용성을 확장하는데도 필요한 연구이다. 본 연구는 연학학습 환경에서 나타날 수 있는 보안취약점에 대한 연구 동향을 제시함으로써 이러한 연구 필요성을 환기하고 향후 추가적인 연구 방향을 제시하는데 목적이 있다. 향후 연합학습에 대한 데이터 보안성 강화를 위해서는 클라이언트 기기의 이질성, 이기종 데이터 학습 문제, 그리고 클라우드와 같은 다양한 학습 환경에서 보안 취약점을 분석하고 데이터 프라이 버시 강화와 학습의 효율성을 제고라는 두가지 목적을 모두 충족시킬 수 있는 다양한 연구가 필요하다.

참고문헌

[1] S. Wang et al., "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1205-1221, June 2019, doi: 10.1109/JSAC.2019.2904348.

[2] McMahan, B., & Ramage, D. (2017). Federated learning: Collaborative machine learning without centralized training data. Google Research Blog, 3.

[3] Ghosh, A., Chung, J., Yin, D., & Ramchandran, K. “An efficient framework for clustered federated learning.” Advances in Neural Information Processing Systems, 33, 2020, 19586-19597.

[4] Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., & Jararweh, Y. (2022). Federated learning review: Fundamentals, enabling technologies, and future applications. Information processing & management, 59(6), 103061.

[5] 오석환, 정송헌, & 김경백. (2023). 데이터 중독 공격 방어를 위한 신뢰도 점수 기반 연합학습. 디지털콘텐츠학회논문지, 24(6), 1317-1326.

[6] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated learning: Challenges methods and future directions", IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, 2020.

[7] Nasr, M., Shokri, R., & Houmansadr, A. (2019, May). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE symposium on security and privacy (SP) (pp. 739-753). IEEE.

[8] Truex, S., Liu, L., Gursoy, M. E., Yu, L., & Wei, W. (2019). Demystifying membership inference attacks in machine learning as a service. IEEE Transactions on Services Computing, 14(6), 2073-2089.

[9] Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., & Qi, H. (2019, April). Beyond inferring class representatives: User-level privacy leakage from federated learning. In IEEE INFOCOM 2019-IEEE conference on computer communications (pp. 2512-2520). IEEE.

[10] Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019, May). Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE symposium on security and privacy (SP) (pp. 691-706). IEEE.

[11] Song, M., Wang, Z., Zhang, Z., Song, Y., Wang, Q., Ren, J., & Qi, H. (2020). Analyzing user-level privacy attack against federated learning. IEEE Journal on Selected Areas in Communications, 38(10), 2430-2444.

[12] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619-640.

[13] Dwork, C. (2006, July). Differential privacy. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

[14] 남기빈, 조명현, 김현준, 백윤흥. (2021). 동형암호를 활용한 딥러닝 모델 학습에 대한 연구. 한국정보처리학회 학술대회논문집, 28(1), 113-116.

[15] Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., & Liu, Y. (2020). {BatchCrypt}: Efficient homomorphic encryption for {Cross-Silo} federated learning. In 2020 USENIX annual technical conference (USENIX ATC 20) (pp. 493-506).

[16] J. Park, N. Y. Yu and H. Lim, "Privacy-Preserving Federated Learning Using Homomorphic Encryption With Different Encryption Keys," 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea,

Republic of, 2022, pp. 1869-1871, doi: 10.1109/ICTC55196.2022.9952531.

[17] 주유연, 백윤흥. (2022). 신뢰실행환경을 활용한 딥러닝 추론에 관한 연구. 한국정보처리학회 학술대회논문집, 29(2), 234-236

[18] Buckman, J., Roy, A., Raffel, C., & Goodfellow, I. (2018, February). Thermometer encoding: One hot way to resist adversarial examples. In International conference on learning representations.

[19] Y. Chen, Y. Gui, H. Lin, W. Gan and Y. Wu, "Federated Learning Attacks and Defenses: A Survey," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 4256-4265, doi: 10.1109/BigData55660.2022.10020431.