안전한 연합학습 기술 연구 동향 및 분석

여웅기, 이준우*
중앙대학교, *중앙대학교

sam1705@cau.ac.kr, *jwlee2815@cau.ac.kr

A Survey on Trends and Analysis of Robust Federated Learning Techniques

Yeo Ung Gi, Lee Joon Woo*
Chung-Ang Univ., *Chung-Ang Univ.

요 약

연합학습은 다양한 분야에서 무궁무진한 활용 가능성을 가지고 있다. 특히, 연합학습은 다양한 분야에서 무궁무진한 활용 가능성을 가지고 있다. 특히, 여러 클라이언트나 기관에 분산된 데이터를 이용해 모델이 직접 학습을 수행하는 대신, 각 로컬 클라이언트가 학습을 진행한 후 그 파라미터를 글로벌 모델에 전달하여 최종적으로 글로벌 모델을 학습시키는 방식으로 로컬 데이터의 프라이버시를 보존하면서도 글로벌 모델이 로컬 데이터를 이용해 학습한 것과 유사한 성능을 얻을 수 있다. 그러나 여전히 프라이버시가 완전히 보존되지 않는 문제가 있으며, 다양한 공격이 등장하고 있다. 이를 방지하기 위해 안전하게 학습을 수행할 수 있는 연합학습 방법론들이 제안되었으며, 본 논문에서는 이러한 제안된 방법들과 그 연구 동향에 대해 조사하고 분석하였다.

Ⅰ. 서 론

머신러닝 모델이 학습에 사용하는 데이터에 따라 다양한 산업 분야에 적용될 수 있으며, 현재 큰 영향력을 미치고 있다. 그러나 기존 머신러닝은 여러 클라이언트에 분산된 데이터에 대해 학습을 진행하기 어렵다는 문제가 있다. 이는 각 데이터에 포함된 개인 정보가 학습 과정에서 노출되어 프라이버시를 보존하기 어렵기 때문이다. 이 문제를 해결하기 위해 여러 방법이 제안되었으며, 그중 연합학습은 가장 주목받는 방법이다. 연합학습은 여러 클라이언트나 기관에 분산되어 있는 데이터를 이용해 모델이 직접 학습을 수행하는 대신, 각 로컬 클라이언트가 학습을 진행한 후 그 파라미터를 글로벌 모델에 전달하여 최종적으로 글로벌 모델을 학습시키는 방식이다. 이를 통해 로컬 데이터의 프라이버시를 보존하면서도 글로벌 모델이 로컬 데이터를 이용해 학습한 것과 유사한 성능을 낼 수 있다. 그러나 연합학습조차도 여전히 프라이버시가 완전히 보존되는 것은 아닌데, 그 이유는 연합학습에 나쁜 영향을 줄 수 있는 다양한 공격이 등장했기 때문이다. 이를 방지하고자 안전하게 학습을 수행할 수 있는 연합학습 방법론들이 제안되었으며, 본 논문에서는 이러한 제안된 방법들과 그 연구 동향에 대해 자세히 설명할 것이다.

Ⅱ. 본론

안전한 연합학습을 수행하기 위해 안정성과 프라이버시 보존은 반드시 해결해야 할 문제이다. 이 문제를 해결하기 위해 데이터 프라이버시의 보호와 공격자의 공격 전략 식별에 대한 광범위한 연구가 진행되었다. 최근 연구들은 주로 모델 파라미터를 보호하여 공격자가 클라이언트의 데이터에 무단으로 접근하는 것을 방지하는 데 중점을 두고 있다.

FLAME 는 FLAME 는 셔플 모델에서 차등 프라이버시(Differential Privacy)보장하는 연합 학습을 위한 프레임워크로 [1], 제 3 자에 의존하지 않고 큐레이터 모델의 정확성과 로컬 모델의 강력한 프라이버시 보장을 결합한 것이다. FLAME 은 셔플 모델의 프라이버시 증폭 효과를 활용하여 큐레이터 모델의 정확성과 제 3 자가 없는 강력한 프라이버시를 모두 달성한다.

SplitFed 는 연합 학습과 분할 학습(Split Learning, SL)을 결합하여 두 접근 방식의 단점을 극복하는 새로운 분산 기계 학습 접근 방식이다 [2]. 연합 학습은 클라이언트에서 로컬 데이터를 사용해 전체 기계 학습 모델을 훈련한 후, 각 클라이언트에서 훈련된 모델을 서버로 전송하여 글로벌 모델을 형성한다. SplitFed 는 FL 과 SL 의 장점을 결합하여 병렬 처리를 통해 빠른 모델 훈련을 가능하게 하고, 네트워크를 분할하여 자원 제한적인 클라이언트에서도 효율적으로 작동한다. 또한, 차등 프라이버시와 PixelDP 를 통합하여 데이터 프라이버시와 모델의 안전성을 향상시킨다.

백도어 공격(Backdoor Attack)은 모델이 특정 패턴이 포함된 입력에 대해 의도적으로 잘못된 출력을 내도록 하는 공격이다. 이는 모델의 전반적인 성능에 영향을 주지 않으면서 특정 입력에 대해 원하는 출력이 나오게 한다. 이 논문에서는 데이터 이질성이 백도어 공격의 성공에 중요한 요소임을 발견하고, 이를 방어하기 위한 몇 가지 전략을 제안한다 [3]. 첫째, 중앙 서버가

소규모의 글로벌 IID(Intrinsically Identical Distribution) 데이터셋을 유지하여 모든 참여 클라이언트의 업데이트된 가중치를 통합하기 전에 훈련함으로써 악성 클라이언트의 마지막 배치에서 발생하는 과적합을 방지하는 방법을 제안한다. 둘째, 클라이언트 선택을 다양화하는 메커니즘을 구현하여 각 라운드마다 동일한 클라이언트를 선택하지 않도록 스케줄링 정책을 적용함으로써 악성 클라이언트가 연속적으로 선택되는 것을 방지한다. 셋째, 공격자가 글로벌 데이터 분포를 추정하지 못하도록 글로벌 데이터 분포를 숨기거나 왜곡하여 공격자가 잘못된 분포를 기반으로 공격을 설계하게 만드는 전략을 제안한다. 이러한 방어 전략들은 데이터 이질성이 높은 환경에서도 연합 학습 시스템의 견고성을 강화시킨다.

앙상블 연합 학습(Ensemble Federated Learning, EFL)은 악성 클라이언트로부터 연합 학습 모델을 보호하기 위해 제안된 새로운 접근 방식이다 [4]. 기존의 연합 학습은 단일 글로벌 모델을 학습하는데, 이는 악성 클라이언트에 의해 쉽게 손상될 수 있다. EFL 은 이러한 문제를 해결하기 위해 여러 글로벌 모델을 학습하고 그 결과를 결합하는 방식을 사용한다. 구체적으로, 각 글로벌 모델은 클라이언트의 무작위 서브샘플을 사용하여 학습되며, 각 테스트 예제에 대해 여러 글로벌 모델의 예측 레이블 중 다수결로 최종 예측 레이블을 결정한다. 이러한 접근 방식은 다수의 글로벌 모델이 정상 클라이언트를 기반으로 학습되므로, 악성 클라이언트의 영향을 받지 않도록 보장할 수 있다. EFL 은 기본 연합 학습 알고리즘과 결합하여 사용되며, 이를 통해 연합 학습의 보안을 강화하고, 악성 클라이언트의 공격으로부터 모델의 예측 정확도를 보호할 수 있다.

FedFed 는 데이터 이질성을 해결하기 위해 제안된 연합 학습 프레임워크로 [5], 정보 병목 (Information Bottleneck, IB) 방법에서 영감을 받아 성능 민감 피처와 성능 강건 피처로 데이터를 분리하여 처리한다. 각 클라이언트는 로컬 데이터를 성능 민감 피처와 성능 강건 피처로 나눈 후, 성능 민감 피처는 랜덤 노이즈를 추가하여 보호한 후 전역적으로 공유하고, 성능 강건 피처는 로컬에 유지한다. 이를 통해 클라이언트는 로컬 데이터와 공유된 데이터를 사용하여 모델을 훈련하여 데이터 이질성을 완화할 수 있다.

Ⅲ. 결론

공격 전략이 다양해지고 발전함에 따라, 공격자로부터 클라이언트의 프라이버시를 보존하는 것이 중요해지고 있다. 이에 따라 안전한 연합학습을 수행하는 기술이 여러 분야에서 활발히 활용될 것으로 전망된다. 연합학습에 관한 많은 선행 연구가 진행되었지만, 안전성 관점에서 프라이버시를 어떻게 보존할 것인지에 관한 연구는 아직 타 분야에 비해 많이 부족한 실정으로, 더 많은 연구가 필요한 분야이다. 본 논문에서는 안전한 연합학습을 위한 기술 연구 동향에 대해 조사하고 분석하였다.

참 고 문 헌

[1] Liu, Ruixuan, et al. "Flame: Differentially private federated learning in the shuffle model." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 10. 2021.

[2] Thapa, Chandra, et al. "Splitfed: When federated learning meets split learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 8. 2022.

[3] Zawad, Syed, et al. "Curse or redemption? how data heterogeneity affects the robustness of federated learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 12. 2021.

[4] Cao, Xiaoyu, Jinyuan Jia, and Neil Zhenqiang Gong. "Provably secure federated learning against malicious clients." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 8. 2021.

[5] Yang, Zhiqin, et al. "FedFed: Feature distillation against data heterogeneity in federated learning." Advances in Neural Information Processing Systems 36 (2024).