연합학습환경에서KnowledgeDistillation기반사후방어기법의효과분석: FGSM및PGD공격시나리오를중심으로

황선준† , 전홍준*, 왕우석**, 이재훈***

연세대학교

† sunjun7559012@yonsei.ac.kr, *hongjun2295@yonsei.ac.kr,

**wangws1108@yonsei.ac.kr, jjahuni@yonsei.ac.kr

Post-hoc Defense with Knowledge Distillation in Federated Learning: An Empirical Study against FGSM and PGD Attacks

Sun Jun Hwang† , Hong Joon Jun*, Woo Seok Wang**, Jaehoon Lee***

Yonsei Univ.

요 약

연합학습(Federated Learning, FL)은 데이터 프라이버시를 보장하면서 분산 환경에서 모델을 학습할 수 있는 장점을 제공하지만, 적대적 예제(adversarial example)에 취약하다는 한계가 있다. 본 연구에서는 서버 측 Knowledge Distillation(KD)을 활용한 사후(post-hoc) 방어 전략을 제안하고, FGSM(Fast Gradient Sign Method)과 PGD(Projected Gradient Descent) 공격에 대한 효과를 실험적으로 분석하였다. MNIST, FashionMNIST, CIFAR-10, CIFAR-100, SVHN 다섯 가지 데이터셋을 대상으로 서버-사이드, 모든 클라이언트 공격(localALL), 단일 클라이언트 공격(localONE), 단일 클라이언트가 공격 데이터만 사용하는 경우(localONE_attackedOnly) 등 네 가지 시나리오에서 평가를 수행하였다. 실험 결과, PGD가 FGSM보다 훨씬 강력한 공격으로 나타났으며, 특히 서버-사이드와 localALL 시나리오에서 전역 모델의 성능 저하가 가장 크게 나타났다. 반대로 localONE 시나리오에서는 분산 구조의 특성상 공격 영향이 부분적으로 완화되었으나, localONE_attackedOnly에서는 공격 신호가 집중되어 성능이 크게 저하되었다. KD 기반 사후 업데이트는 clean 성능을 유지하면서 공격 완화 효과를 제공하였으며, 이는 FL 환경에서 경량적이고 범용적인 방어 전략으로 활용될 수 있음을 시사한다.

Ⅰ. 서 론

연합학습(Federated Learning, FL)은 클라이언트의 데이터를 중앙 서버로 수집하지 않고, 각 클라이언트가 로컬에서 모델을 학습한 뒤 그 결과를 서버에서 집계하는 방식으로 개인정보 보호와 분산 학습을 동시에 달성할 수 있는 새로운 패러다임이다 [1,2]. 데이터가 서버로 직접 전송되지 않기 때문에 개인정보 보호 측면에서 장점이 크며, 의료, 금융, 모바일 서비스, IoT 등 다양한 분야에서 빠르게 적용되고 있다. 그러나 FL은 구조적으로 다수의 참여자와 중앙 서버 간의 상호작용에 의존하기 때문에 다양한 보안 위협에 취약하다. 특히 적대적 예제(adversarial example)를 이용한 공격은 입력에 미세한 교란을 추가하여 모델이 잘못된 출력을 내도록 하는 기법으로, 딥러닝 모델의 신뢰성을 크게 위협한다 [3,4].

대표적인 적대적 공격 기법으로는 (1) FGSM(Fast Gradient Sign Method)과 (2) PGD(Projected Gradient Descent)가 있다. FGSM은 단일 스텝으로 손실 함수의 그래디언트 부호를 따라 입력을 교란하는 방식으로, 계산 효율성이 높지만 공격력이 제한적이다 [5]. 반면 PGD는 반복적으로 FGSM을 적용하면서 교란이 ϵ-범위를 벗어나지 않도록 투영(projection)하는 방식으로, FGSM보다 훨씬 강력한 공격으로 알려져 있다 [6].

$$x^{\text{adv}} = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \quad \cdots \quad (1)$$

$$x^{(t+1)} = \Pi_{B_\infty(x,\epsilon)}(x^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^{(t)}, y))) \quad \cdots \quad (2)$$

이러한 공격이 FL 환경에서 발생할 경우, 서버 또는 클라이언트 위치에 따라 전역 모델에 미치는 영향이 달라진다. 따라서 공격자의 위치와 전략에 따른 위협 수준을 정량적으로 평가하는 것은 매우 중요하다. 한편, 기존 방어 기법인 차등 개인정보보호, 암호화 기반 보안, 적대적 학습은 성능 저하·계산 비용·특정 공격에 대한 과적합 등 한계가 있다 [7,8]. Knowledge Distillation(KD)은 교사 모델의 soft label을 학생 모델이 학습하게 하는 방식으로 [9], 원래는 모델 경량화 목적이지만 최근 연구에서는 적대적 공격에 대한 완화 효과도 보고되고 있다 [10,11]. 본 연구에서는 서버 측에서 KD를 사후(post-hoc) 적용하여, 공격 이후에도 전역 모델의 성능을 보존할 수 있는지를 검증하였다.

Ⅱ. 본론

본 연구에서는 다섯 가지 벤치마크 데이터셋(MNIST, FashionMNIST, CIFAR-10, CIFAR-100, SVHN)을 대상으로 FGSM과 PGD 공격을 수행하였다.

실험은 네 가지 시나리오에서 이루어졌다. 첫째, 서버-사이드(Server-side) 공격은 중앙 서버가 전역 모델에 대해 직접 적대적 예제를 생성하는 방식으로, 모든 클라이언트에 영향을 미쳐 가장 치명적인 결과를 초래한다. 둘째, localALL 공격은 모든 클라이언트가 적대적 예제를 사용해 학습을 수행하는 경우로, 전체 네트워크가 동시에 오염되는 상황을 반영한다. 셋째, localONE 공격은 단일 클라이언트만이 적대적 예제를 사

Dataset	시나리오	EPS	Baseline	FGSM	PGD
MNIST	Server-side	0.1	0.9949	0.7230	0.0012
	Local All	0.1	0.9949	0.7892	0.1712
	Local One	0.1	0.9949	0.9511	0.8210
	Local One Only	0.1	0.9949	0.7735	0.1230
Fashion MNIST	Server-side	0.06	0.9304	0.0922	0.0000
CIFAR-10	Server-side	0.0157	0.9072	0.1222	0.0000
CIFAR-100	Server-side	0.0157	0.6500	0.0805	0.0018
SVHN	Server-side	0.0157	0.9526	0.5012	0.499

표 1. 대표 epsilon에서의 시나리오별 성능 요약

용하여 업데이트를 제출하는 경우로, 분산 평균화 덕분에 공격 영향이 희석될 수 있다. 마지막으로 localONE_attackedOnly 시나리오는 단일 클라이언트가 오직 공격 데이터만 사용하여 학습하는 경우로, 공격 신호가 집중적으로 반영되어 localONE보다 더 큰 성능 저하를 유발할 수 있다.

각 데이터셋은 FedAvg [1,2] 알고리즘을 기반으로 학습되었으며, epsilon 값은 데이터셋 특성에 맞추어 설정하였다. MNIST의 경우 0.1, FashionMNIST는 0.06, CIFAR-10과 CIFAR-100은 0.0157, SVHN은 0.0157을 대표값으로 선택하였다. 본 연구의 핵심은 이 네 가지 시나리오에서 FGSM과 PGD의 상대적 공격력을 비교하고, 서버 측 KD 기반 사후 업데이트가 공격 완화에 기여하는지를 확인하는 것이다.

실험 결과는 표 1에 요약되어 있다. 각 데이터셋과 시나리오에서 baseline 성능과 공격 이후 성능을 비교하면, PGD가 FGSM보다 훨씬 더 치명적인 성능 저하를 유발함을 명확히 확인할 수 있다.

분석 결과, MNIST와 FashionMNIST에서는 작은 epsilon 값에서도 서버-사이드 PGD 공격이 전역 모델을 사실상 무력화하였다. CIFAR-10과 CIFAR-100 역시 baseline 대비 큰 폭의 성능 저하를 보였으며, SVHN에서도 동일한 경향이 확인되었다. 시나리오를 비교하면 서버-사이드와 localALL 공격이 가장 치명적이었으며, localONE에서는 분산 구조의 평균화 효과로 인해 전역 모델의 성능이 상당 부분 유지되었다. 그러나 localONE_attackedOnly의 경우 공격 신호가 집중되어 localONE보다 더 큰 성능 저하가 관찰되었다. 이는 공격자의 전략에 따라 단일 클라이언트 공격도 전역 모델에 심각한 영향을 줄 수 있음을 시사한다.

KD 기반 사후 업데이트를 적용한 경우, clean 성능이 크게 훼손되지 않으면서 공격 영향이 완화되는 경향이 확인되었다. 특히 localONE과 같은 시나리오에서는 KD의 방어 효과가 분명하게 나타났다. 그러나 서버-사이드 PGD처럼 매우 강력한 공격에 대해서는 KD만으로 충분하지 않았으며, 추가적인 방어 기법과의 결합이 필요함을 보여주었다.

Ⅲ. 결론

본 연구에서는 FL 환경에서 FGSM과 PGD 공격을 네 가지 시나리오(서버-사이드, localALL, localONE, localONE_attackedOnly)로 나누어 실험하고, 서버 측 KD 기반 사후 방어 전략의 효과를 분석하였다. 실험 결과, PGD가 FGSM보다 훨씬 강력한 공격임이 재확인되었으며, 서버-사이드와 localALL 시나리오에서 전역 모델이 가장 심각한 성능 저하를 보였다. 반대로 localONE 시나리오에서는 분산 구조의 특성상 공격 영향이 희석되어 성능이 비교적 안정적으로 유지되었으나, localONE_attackedOnly는 공격 신호가 집중되어 더 큰 피해를 주었다. KD 기반 사후 업데이트는 clean 성능을 유지하면서 공격 완화 효과를 제공하였으며, FL 환경에서 경량적이고 범용적인 방어 전략으로 활용될 수 있는 가능성을 보여주었다. 향후 연구에서는 KD를 다른 방어 기법과 결합한 하이브리드 전략을 탐색하고, 비동기적·비IID 데이터 분포를 반영한 현실적 환경에서의 유효성을 검증하는 것이 필요하다.

참 고 문 헌

[1] Kairouz, P., et al., Federated Learning: Strategies, Applications, and Future Directions, Proceedings of the IEEE, 2021.

[2] Li, T., et al., Federated Learning in Practice: Reflections and Projections, arXiv preprint arXiv:2012.02783, 2020.

[3] Szegedy, C., et al., Intriguing properties of neural networks, ICLR, 2014.

[4] Yuan, X., et al., Adversarial Examples: Attacks and Defenses for Deep Learning, IEEE TNNLS, 2019.

[5] Goodfellow, I. J., Shlens, J., & Szegedy, C., Explaining and Harnessing Adversarial Examples, ICLR, 2015.

[6] Madry, A., et al., Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR, 2018.

[7] Sun, J., et al., Federated Learning Vulnerabilities, Threats, and Defenses: A Survey, ACM Computing Surveys, 2023.

[8] Zhang, C., et al., A Survey on the Security of Federated Learning, IEEE TPDS, 2022.

[9] Hinton, G., et al., Distilling the Knowledge in a Neural Network, NeurIPS Deep Learning Workshop, 2015.

[10] Gou, J., et al., A Survey on Knowledge Distillation: Fundamentals, Applications and Future Directions, ACM Computing Surveys, 2021.

[11] Shen, Y., et al., Knowledge Distillation Meets Adversarial Robustness: A Comprehensive Survey, IEEE Access, 2023.