연합 학습 기반 안드로이드 악성코드 탐지에서 label flipping attack 방어를 위한 FedLB 집계 기법

FedLB Aggregation for Defending Against Label Flipping Attacks in Federated Learning-based Android Malware Detection

장성준[1] · 오찬석[1] · 김주영[2] · 최석환[2]
Seong-Joon Chang, Chan-Seuk Oh, Ju-Young Kim and Seok-Hwan Choi

[1]연세대학교 전산학과
E-mail: {sj90724sj, chanseok_oh}@yonsei.ac.kr
[2]연세대학교 소프트웨어학부
E-mail: {juyoung0401, sh.choi}@yonsei.ac.kr

요 약

안드로이드 악성코드 데이터셋은 수집 시기와 조건에 따라 서로 다른 특성을 가지며, 이는 연합 학습 환경에서 데이터 이질성 문제로 이어질 수 있다. 특히 기존 FedAvg는 데이터 수를 기준으로 집계 가중치를 결정하므로, 악성 클라이언트의 조작된 업데이트에 취약할 수 있다. 이에 본 논문에서는 로컬 학습 loss를 기반으로 집계 가중치를 조정하는 FedLB를 제안하고, label flipping attack 환경에서의 방어 효과를 평가하였다. 실험 결과, FedLB는 모든 공격 조건에서 FedAvg보다 높은 Accuracy와 F1-score를 보이며 글로벌 모델의 성능 저하를 완화하였다.

키워드 : 안드로이드 악성코드 탐지, 연합 학습, 포이즈닝 공격, 손실 기반 집계

1. 서 론

안드로이드 애플리케이션 기반 악성코드 탐지는 지속적으로 변화하는 공격 양상에 대응해야 하며, 수집 시기와 조건이 다른 데이터가 함께 사용된다. 그러나 실제 환경에서는 기관 또는 시스템 간 데이터 공유에 제약이 있고, 데이터셋마다 수집 기간, 조건, 규모가 달라 큰 차이를 보인다[1][4]. 이러한 환경에서 연합 학습은 로컬 데이터를 직접 공유하지 않고도 글로벌 모델을 학습할 수 있어 안드로이드 보안 분야에 적합하다[2]. 반면, Non-IID 환경에서는 글로벌 모델의 일반화 성능이 저하되거나 특정 클라이언트의 업데이트가 전체 학습에 과도한 영향을 줄 수 있다. 또한 연합 학습은 클라이언트가 학습 결과만 서버에 전달하므로, 서버가 각 클라이언트의 데이터와 학습 과정의 신뢰성을 검증하기 어렵다. 이로 인해 악의적인 클라이언트의 조작된 업데이트가 글로벌 모델의 안정성을 저하시킬 수 있으며, 특히 label flipping attack은 라벨을 반전시켜 잘못된 업데이트를 유도하는 대표적인 공격 방식이다. 더욱이 데이터 규모와 분포가 서로 다른 환경에서는 단순 평균 기반 집계만으로 이러한 영향을 충분히 완화하기 어렵다.

따라서, 본 논문에서는 이러한 문제를 완화하기 위해, 연합 학습 기반 안드로이드 악성코드 탐지 환경에서 데이터 수만으로 집계 가중치를 결정하는 FedAvg 방식의 한계를 실험적으로 보인다. 또한, FedAvg의 한계를 보완하기 위해 로컬 학습 결과를 반영해 클라이언트 기여도를 조정하는 loss 기반 집계 방식인 FedLB를 제안하며, 공격 환경에서의 방어 효과와 글로벌 모델의 일반화 성능 변화를 실험적으로 분석한다.

감사의 글 : 본 연구는 20xx년 OO연구개발 지원사업에 의해 수행되었습니다. 이곳에는 감사의 글(사사표기)을 적는다. (감사의 글 스타일, 9pt)

2. 본 론

2.1 데이터셋

본 논문에서는 수집 시기 및 조건이 서로 다른 데이터셋을 지닌 현실적인 클라이언트 환경을 위하여 CIC-MalDroid 2020[2], KronoDroid[1], AndroZoo[3][4]를 로컬 데이터셋으로 사용하였다. 또한, 현실적인 환경과 유사한 평가를 위해 정상 애플리케이션과 악성 애플리케이션 비율이 4:1인 CIC-AndMal2017[5]을 테스트 데이터셋으로 사용하였다.

CIC-AndMal2017은 2015년부터 2017년, CIC MalDroid 2020은 2017년부터 2018년, KronoDroid는 2008년부터 2020년까지의 정상 애플리케이션과 2019년부터 2020년까지의 악성 애플리케이션, AndroZoo는 2020년 5월부터 2025년까지 수집한 애플리케이션을 포함한다. 본 논문에서는 로컬 데이터셋을 훈련 80%, 검증 10%, 테스트 10%로 분할하여 학습 및 검증에 사용하였다.

2.2 집계 기법

연합 학습에서 글로벌 모델 집계 방식은 각 클라이언트의 업데이트 반영 비율을 결정하는 핵심 요소이다. 기존 FedAvg는 데이터 수를 기준으로 집계 가중치를 결정하며, 글로벌 모델은 식 (1)과 같이 계산된다.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} w_{t+1}^{k} \quad (1)$$

여기서 K는 전체 클라이언트 수, n_k는 k번째 클라이언트의 데이터 수, w_{t+1}^k는 로컬 학습 후 모델 파라미터를 의미한다. 그러나 이러한 방식은 데이터 수가 큰 클라이언트의 업데이트를 더 크게 반영하므로, 악성 또는 편향된 클라이언트의 영향이 과도하게 반영될 수 있다. 이를 완화하기 위해 본 논문에서는 로컬 학습 loss를 기반으로 가중치를 조절하는 FedLB를 제안한다. FedLB는 각 클라이언트의 로컬 학습 결과를 반영하여 글로벌 모델을 식 (2)와 같이 집계한다.

$$w_{t+1} = \sum_{k=1}^{K} \left(\frac{(\frac{1}{L_k + \epsilon})^{\gamma}}{\sum_{i=1}^{K} (\frac{1}{L_i} + \epsilon)^{\gamma}} \right) w_{t+1}^k \quad (2)$$

여기서 L_k는 로컬 loss, ϵ은 수치적 안정성을 위한 상수, γ는 가중치 조절 민감도를 나타낸다. 이와 같이 FedLB는 데이터 수 대신 로컬 학습 loss를 반영하여 클라이언트 기여도를 조정함으로써, 공격 환경에서 FedAvg보다 성능 저하를 완화할 수 있다.

2.3 실험 구성

본 논문에서는 데이터 이질성이 존재하는 연합 학습 기반 안드로이드 악성코드 탐지 환경에서 FedAvg와 FedLB의 성능을 비교하였다. 각 클라이언트는 서로 다른 데이터셋을 로컬 데이터셋로 사용하여 데이터 이질성이 존재하는 Non-IID 환경을 구성하였다. 실험은 기준 학습 단계, 공격 적용 단계, 방어 적용 단계의 세 단계로 구성하였다. 기준 학습 단계에서는 정상 클라이언트만을 사용한 FedAvg 학습을 수행하였다. 공격 적용 단계에서는 각 클라이언트를 개별적으로 악성 클라이언트로 설정하고 FedAvg 환경에서 label flipping attack을 적용하였다. 방어 적용 단계에서는 동일한 공격 조건에서 FedLB의 방어 성능을 평가하였다. 모든 실험은 동일한 모델 및 학습 조건에서 집계 방식만 다르게 적용하였으며, Accuracy와 F1-score를 기준으로 성능을 비교하였다.

2.4 실험 및 성능 비교

표 1은 악성 클라이언트별 FedAvg와 FedLB의 성능 비교를 나타낸 것이다. 표 1을 보면, 기존 FedAvg 기법을 활용한 정상 FL 환경에서는 Accuracy 70.61%, F1-score 84.97%로 나타났다. 그러나 동일한 환경에서 label flipping attack이 적용되면 성능 저하가 뚜렷하게 나타났다. 표 1의 결과에 따르면, FedAvg 기법이 적용된 공격 환경에서 글로벌 모델의 성능은 Accuracy 평균 27.47%, F1-score 평균 46.71%로 나타났다. 반면, 동일한 공격 조건에서 FedLB를 적용한 경우 Accuracy는 평균 69.32%, F1-score는 평균 83.81%로 모든 공격 조건에서 FedAvg 대비 높은 성능을 보였다. 이는 loss 기반으로 집계 가중치를 조정하는 것이 악성 클라이언트의 영향 완화에 효과적일 수 있음을 보여준다.

3. 결 론

본 논문에서는 데이터 이질성이 존재하는 연합 학습 기반 안드로이드 악성코드 환경에서 label flipping attack에 대한 방어 기법으로 FedLB를 제안하였다. 기존 FedAvg는 데이터 수를 기준으로 집계 가중치를 결정하므로, 악성 클라이언트의 조작된 업데이트가 글로벌 모델에 크게 반영될 수 있지만, FedLB는 각 클라이언트의 로컬 학습 loss를 기반으로 집계 가중치를 조정하여 이러한 영향을 완화하고자 하였다. 실험 결과, FedLB는 FedAvg 대비 높은 회복량을 보였으며, 이를 통해 FedLB는 데이터 이질성이 존재하는 연합 학습 기반 안드로이드 악성코드 탐지 환경 내 공격에서 보다 강건한 집계 방식으로 활용될 수 있음을 확인하였다.

표 1. 악성 클라이언트별 FedAvg와 FedLB 성능 비교

Malicious client	Scenario	Accuracy	F1-score
None	Normal FL	0.7061	0.8497
CIC MalDroid 2020	FedAvg under attack	0.0579	0.5645
	Proposed	0.4991	0.7083
KronoDroid	FedAvg under attack	0.3991	0.4578
	Proposed	0.7777	0.8934
AndroZoo	FedAvg under attack	0.3670	0.3790
	Proposed	0.8027	0.9127

참 고 문 헌

[1] A. Guerra-Manzanares, H. Bahsi, and S. Nõmm, “Kronodroid: Time-based Hybrid-Featured Dataset for Effective Android Malware Detection and Characterization,” Computers & Security, vol. 110, pp. 102399, 2021.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273-1282, 2017.

[3] S. Mahdavifar, A. F. A. Kadir, R. Fatemi, D. Alhadidi, and A. A. Ghorbani, “Dynamic Android Malware Category Classification using Semi-Supervised Deep Learning,” in the Proceedings of the 18th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC), Aug. 17 - 24, 2020.

[4] M. Alecci, P. J. Ruiz Jiménez, K. Allix, T. F. Bissyandé, and J. Klein, “AndroZoo: A Retrospective with a Glimpse into the Future,” in the Proceedings of the 2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR), 2024.

[5] F. Mercaldo and A. Santone, “Deep Learning for Image-Based Mobile Malware Detection,” Journal of Computer Virology and Hacking Techniques, 2020.

[6] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, “Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification,” in the Proceedings of the 52nd IEEE International Carnahan Conference on Security Technology (ICCST), Montreal, Quebec, Canada, 2018.