

MMDD-FLARE: 지역 및 환경 간 연합학습을 통한 메모리 기반 동적 악성코드 탐지

장현규¹, Segun Popoola², 이현우³

¹한국에너지공과대학교 에너지공학부 통합과정

²Anglia Ruskin University School of Computing and Information Science Senior Lecturer

³한국에너지공과대학교 에너지공학부 조교수

hkjang@kentech.ac.kr, segun.popoola@aru.ac.uk, hwlee@kentech.ac.kr

MMDD-FLARE: Memory-based Malware Dynamic Detection via Federated Learning Across Regions and Environments

Hyunkyu Jang¹, Segun Popoola², Hyunwoo Lee³

¹Dept. of Energy Engineering, Korea Institute of Energy Technology (KENTECH)

²School of Computing and Information Science, Anglia Ruskin University

³Dept. of Energy Engineering, Korea Institute of Energy Technology (KENTECH)

요 약

전 세계를 대상으로 하는 사이버 위협이 고도화되고 있으나, 국제적 개인정보 보호 규제로 인해 국가 간 보안 데이터 공유가 제한된다. 본 연구는 원본 데이터를 공유하지 않고 악성코드 탐지 모델을 학습할 수 있는 글로벌 연합 학습 프레임워크인 MMDD-FLARE를 제안한다. 난독화된 악성코드 탐지를 위해 메모리 기반 동적 분석과 경량화된 Dilated CNN 모델을 도입하고, 장거리 및 엄격한 방화벽 환경에서의 원활한 연결을 위해 WireGuard 기반 메시 오버레이 네트워크와 종단 간 암호화를 적용한 제로 트러스트 아키텍처를 구현하였다. 한국-영국 간 실제 테스트베드 실험 결과, Non-IID 환경에서의 성능 저하를 FedProx 알고리즘이 효과적으로 완화함을 확인하였으며, 대륙 간 네트워크 지연에도 불구하고 데이터 유출 없는 글로벌 협력 탐지 모델의 운용 가능성을 입증하였다.

1. 서론

사이버위협은 전 세계를 대상으로 빠르게 퍼지며, 특히 알려지지 않은 제로 데이 취약점 등은 그 파괴력이 커 공격을 탐지하고 차단하는 등 보안 대책의 중요성이 크다. 인공지능(AI)을 활용한 정보보안 시스템 또한 활발히 연구되고 있으며, 이들은 AI 학습 과정에서 많은 데이터를 필요로 한다.

그러나 한국의 개인정보보호법(PIPA)과 영국의 UK GDPR 등 엄격한 국제 규제로 인해 국가 간 보안 데이터 공유가 제한된다. 이에 연합 학습(Federated Learning, FL)이 원본 데이터 교환 없이 협력적 모델 학습을 가능하게 하는 핵심 기술로 주목받고 있다. 기존 연구들은 FL 을 침입 탐지에 적용하는 시도를 해왔으나[1][2][3], 몇 가지 한계가 존재한다. 첫째, 대부분의 연구가 IID(독립 동일 분포) 데이터를 가정하여 지역별 악성코드 분포 차이를 반영하지 못한다. 둘째, 트리 기반 모델(Random Forest, XGBoost)은

중앙집중식에서 우수하지만 단순 파라미터 평균화 메커니즘이 없어 FL 적용이 어렵다. 셋째, 기존 시뮬레이션은 국가 간 높은 네트워크 지연과 엄격한 방화벽(Hard NAT) 등 물리적 네트워크 제약을 무시한다.

본 논문은 이러한 한계를 해결하기 위해 한국-영국 간 실제 테스트베드를 통해 검증된 글로벌 연합 학습 기반 동적 악성코드 탐지 프레임워크를 제안한다. CIC-MalMem-2022 데이터셋을 활용하여 난독화된 악성코드를 탐지하며, 경량 신경망 분류기와 UDP 홀펀칭 및 종단 간 암호화를 활용한 보안 메시 오버레이 네트워크 아키텍처를 도입한다.

본 프레임워크의 주요 기여는 다음과 같다. 첫째, 메모리 기반 동적 분석과 경량 Dilated CNN 을 결합하여 난독화된 악성코드를 효과적으로 탐지하는 FL 호환 분류기를 설계하였다. 둘째, WireGuard 기반 메시 오버레이 네트워크와 DERP 릴레이를 통해 방화벽 및 NAT 환경에서도 안전한 통신이 가능한 제로 트러스트 네트워크 아키텍처를 구현하였다. 셋째, Non-IID 데이터

환경에서 FedProx 알고리즘의 효과를 체계적으로 분석하고, 한국-영국 간 실제 대륙 간 테스트베드에서 실현 가능성을 검증하였다. 본 논문의 구성은 다음과 같다: 2 장에서 관련 연구를 검토하고, 3 장에서 제안 방법을 상세히 기술하며, 4 장에서 실험 결과를 분석하고, 5 장에서 결론 및 향후 연구를 제시한다.

2. 관련 연구

악성코드는 IDS 를 회피하기 위해 코드 난독화 기법이 고도화되고 있으며[4][5], 정적 분석만으로는 탐지가 어려워지고 있다. 동적 분석은 악성코드를 안전한 환경에서 실행하여 메모리 덤프, API 호출 등의 행위 정보를 수집하는 방법으로, CIC-MalMem-2022 데이터셋[6]은 메모리 기반 난독화 악성코드 탐지를 위한 대표적 벤치마크이다. 이 데이터셋은 58,596 개 샘플(Benign 29,298, Malware 29,298)로 구성되며, Volatility 프레임워크 기반 VolMemLyzer를 통해 추출된 55 개 특성을 포함한다. 특성은 프로세스, DLL, 핸들, API 통계 등 다양한 메모리 행위 지표를 포괄한다.

연합 학습은 원본 데이터 대신 로컬에서 학습된 모델 파라미터를 글로벌 서버에 전송하여 집계하는 방식으로[7][8], 데이터 주권을 보장한다. McMahan 등이 제안한 FedAvg[7]는 각 클라이언트의 로컬 모델 가중치를 데이터 크기에 비례하여 가중 평균하는 방식이지만, 클라이언트 간 데이터 분포가 상이한 Non-IID 환경에서는 수렴이 불안정해지는 문제가 있다[9]. 이를 해결하기 위해 Li 등이 제안한 FedProx[10]는 로컬 목적함수에 근접항($\mu/2 \cdot \|w - w_{\text{global}}\|^2$)을 추가하여 로컬 모델이 글로벌 모델로부터 과도하게 발산하는 것을 억제한다. 또한, 분산 환경에서의 다양한 하드웨어와 소프트웨어 호환성을 위해 gRPC 기반의 Flower AI 프레임워크[11]를 활용하였다.

메모리 기반 동적 분석은 실행 중인 프로세스의 메모리 스냅샷에서 악성 행위의 흔적을 추출하는 방식으로, 난독화된 악성코드라도 실행 시 메모리에 원본 코드가 복원되므로 탐지 우회가 어렵다. CIC-MalMem-2022 데이터셋의 55 개 특성은 프로세스 정보(pslist, psscan, dlllist), 네트워크 연결(netscan, connections), 레지스트리 작업(hivelist), 핸들 정보(handles), API 호출 빈도 등을 포함한다. 이러한 특성들은 악성코드의 런타임 행위를 다각도로 반영하여 정적 분석에서 탐지가 어려운 패킹, 다형성, 메타모픽 기법을 사용하는 악성코드도 효과적으로 식별할 수 있다. 기존 연구에서는 트리 기반 모델(Random Forest 등)이 이 데이터셋에서 99% 이상의 정확도를 달성하였으나[12][13], 이는 단일 머신 환경에서의 결과이며 FL 환경에서의 성능은 검증되지 않았다.

3. 제안 방법

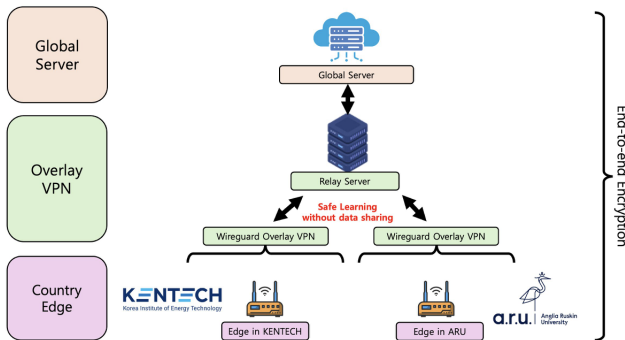
본 연구는 CIC-MalMem-2022 데이터셋을 기반으로 휘발성 메모리 덤프에서 추출된 55 개 특성을 사용하여 동적 탐지 시스템을 설계하였다. 기존 연구에서 트리 기반 모델이 우수한 성능을 보이지만[12][13], 이산적 구조로 인해 FL 의 파라미터 평균화에 적합하지 않다. 따라서 Mezina 등[14]이 제안한 Dilated CNN 분류기를 기본 모델로 채택하였다. 이 모델은 55 차원 특성 벡터를 입력으로 받아 Benign, Trojan, Spyware, Ransomware 의 4 클래스 다중 분류를 수행하며, FedAvg 스타일 집계와 자연스럽게 통합된다. 모델 구조는 복수의 1D 팽창 합성곱 층과 완전연결 층으로 구성되어, 넓은 수용 영역을 확보하면서도 경량화된 파라미터 수를 유지한다.

제안 시스템은 국가 간 협력을 위해 글로벌 규모로 운영된다. 한국-영국 간 8,800km 이상의 거리로 인해 네트워크 지연(약 280ms)과 불안정한 연결이 발생하며, 지역별로 상이한 악성코드 분포가 Non-IID 환경을 형성한다. 이를 해결하기 위해 WireGuard 프로토콜 기반 메시 오버레이 네트워크를 구축하였다. 기존 Hub-and-Spoke VPN 과 달리 컨트롤 플레인과 데이터 플레인을 분리하여, STUN/ICE 기반 UDP 홀 펀칭[15]을 통해 가능한 경우 직접 P2P 연결을 수립한다. 다만 본 테스트베드에서는 KENTECH 와 ARU 양측 기관망의 Hard NAT 및 엄격한 방화벽 정책으로 직접 P2P 연결이 제한되어 DERP(Designated Encrypted Relay for Packets) 릴레이를 경유하는 방식으로 통신을 수행하였으며, 공용 DERP 의 장거리 대역폭 한계(1 Mbps 미만)를 보완하기 위해 한국 내 전용 릴레이 서버를 운영하였다. Non-IID 데이터 시뮬레이션을 위해 Dirichlet 분포($\alpha \in \{0.1, 0.3, 0.5\}$)를 적용하여 클라이언트 간 클래스 불균형을 체계적으로 제어하였다.

실험은 Flower 프레임워크를 활용하여 단일 장치 시뮬레이션과 글로벌 데모 두 가지로 진행하였다. 시뮬레이션은 NVIDIA DGX H100 에서 수행하였고, 글로벌 데모는 한국(KENTECH) 및 영국(ARU) 실제 장치를 사용하였다. 영국 클라이언트는 NVIDIA Jetson Orin Nano 엣지 디바이스로 구성하여 자원 제약 환경을 반영하였다. 주요 실험 변수는: 학습 라운드(50, 100, 500), 데이터 분포(IID, Dirichlet $\alpha \in \{0.1, 0.3, 0.5\}$), 연합 최적화 기법(FedAvg, FedProx $\mu \in \{0.01, 0.05, 0.1, 0.5\}$), 분류기 모델(Dilated CNN, FNN)이다.

Tailscale 서비스를 활용하여 방화벽 정책 변경 없이 안전한 통신을 보장한다. 각 노드는 Tailscale 컨트롤 플레인을 통해 인증 후 WireGuard 키 교환을 수행하며, 이후 데이터 플레인에서는 직접 P2P 통신 또는 DERP 릴레이를 통해 암호화된 패킷을 전송한다. 이

아키텍처는 기존 VPN 의 단일 장애점(Single Point of Failure) 문제를 해결하며, 노드 추가 시에도 기존 네트워크 설정을 변경할 필요가 없다. FL 학습 과정에서 클라이언트는 로컬 데이터로 모델을 학습한 후 모델 가중치만을 서버로 전송하므로, 원본 데이터는 각 클라이언트의 로컬 환경을 벗어나지 않는다. 서버는 수신된 가중치를 FedAvg 또는 FedProx 알고리즘으로 집계하여 글로벌 모델을 생성하고, 이를 다시 클라이언트에 배포하는 과정을 반복한다.



(그림 1) 제안 시스템의 네트워크 구성도

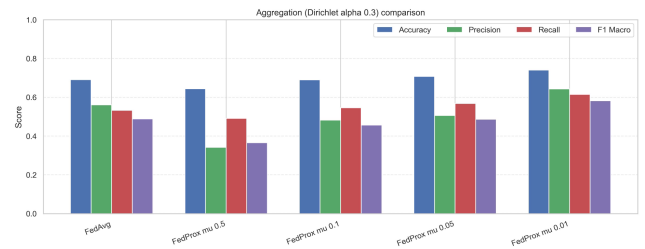
Dilated CNN 의 구조는 Mezina 등[14]의 구현을 따른다. 55 차원 특성 벡터를 입력으로 받아 복수의 1D 팽창 합성곱 층을 통해 넓은 수용 영역을 확보하면서도 경량화된 구조를 유지한다. 팽창 합성곱은 파라미터 수를 늘리지 않으면서 특성 벡터 내 장거리 의존 관계를 포착할 수 있다. 최종적으로 완전연결 층과 소프트맥스 출력 층이 4 클래스(Benign, Trojan, Spyware, Ransomware) 분류를 수행한다. FNN 대비 경량이면서도 비선형 결정 경계를 효과적으로 근사할 수 있어 연합 학습 환경에 적합하다.

4. 실험 결과

학습 라운드 실험에서는 50, 100, 500 라운드로 증가함에 따라 정확도와 Macro-F1 이 전반적으로 향상되는 경향을 보였다. 100 라운드는 글로벌 데모 비교 기준으로 사용되었으며, 본 연구의 기준(baseline)은 500 라운드 IID 환경에서 학습된 Dilated CNN 으로 설정하였다.

데이터 분포 실험에서 IID 대비 Dirichlet $\alpha=0.5$ 및 $\alpha=0.1$ 조건에서 성능 저하가 관찰되었으며, 특히 Trojan 과 Spyware 간 혼동이 증가하였다. 이는 두 클래스의 메모리 행위 패턴이 유사하여 클라이언트 간 클래스 불균형에 민감하기 때문으로 분석된다. 연합 최적화 기법 비교에서 FedProx 는 $\mu=0.01$ 일 때 Dirichlet $\alpha=0.3$ 환경에서 FedAvg 대비 가장 우수한 성능 개선을 보였다. 이는 근접항이 로컬 모델의 과도한 발산을

억제하여 글로벌 모델의 안정성을 향상시키는 것으로 분석된다. 반면 $\mu=0.05$, $\mu=0.1$ 과 같이 근접항 가중치가 커질수록 로컬 학습의 자유도가 제한되어 개선 폭이 감소하는 경향이 관찰되었다.



(그림 2) Dirichlet $\alpha=0.3$ 환경에서의 FedAvg 대비 FedProx 성능 비교

분류기 모델 비교에서 Dilated CNN 은 FNN(2 층, 4 층) 대비 전반적으로 우수한 성능을 보였다. FNN-4 층은 Dilated CNN 보다 낮은 성능을 보였으며, Dilated CNN 은 경량이면서도 팽창 합성곱을 통해 넓은 수용 영역을 확보하여 55 차원 메모리 특성 벡터 내 복잡한 관계를 효과적으로 학습하였다. 클래스별로는 Trojan 과 Spyware 가 상대적으로 낮은 Macro-F1 을 보여 향후 개선의 여지가 있다.

글로벌 테모에서 한국-영국 간 실제 인프라를 통해 연합 학습을 성공적으로 수행하였다. 동일 조건(100 라운드, Dilated CNN)에서 단일 사이트 환경은 4 시간 34 분 23 초가 소요된 반면, 대륙 간 글로벌 환경에서는 7 시간 09 분 49 초가 소요되어 약 +56.6%의 네트워크 오버헤드가 발생하였다. 지역별 DERP 릴레이 지연 측정 결과, 가장 가까운 아시아·유럽 DERP 까지의 왕복 지연이 글로벌 학습을 수행할 수 있는 실용적 수준임을 확인하였다. 모델 성능 자체는 시뮬레이션과 동일하게 유지되어, 네트워크 지연이 최종 모델 수렴에는 영향을 미치지 않음을 확인하였다.

Dirichlet $\alpha=0.3$ 조건의 aggregation 실험에서 FedProx $\mu=0.01$ 이 FedAvg 대비 가장 우수한 Macro-F1 을 보였으며, 특히 클라이언트 간 분포 편차에 민감한 소수 클래스(Trojan, Spyware)에서 개선 폭이 상대적으로 크게 나타났다. μ 값이 커질수록(0.05, 0.1) 근접항의 제약이 강해져 개선 폭이 감소하는 경향이 관찰되었다. 경량화된 모델 크기와 Flower 기반 gRPC 통신 덕분에 대륙 간 대역폭 제약 환경에서도 실용적인 시간 내에 AI 학습이 진행되었다.

5. 결론

본 연구는 국가 간 데이터 규제를 준수하면서 악성코드 탐지 모델을 협력적으로 학습할 수 있는 글로벌 연합 학습 프레임워크 MMDD-FLARE 를 제안하였다. 메모리 기반 동적 분석과 Dilated CNN

분류기를 통해 난독화된 악성코드를 탐지하고, WireGuard 기반 네트워크 아키텍처로 제로 트러스트 통신을 보장하였다. 실험 결과, Non-IID 환경에서 FedProx가 성능 저하를 효과적으로 완화하였으며, 한국-영국 간 실제 테스트베드에서 데이터 유출 없는 글로벌 협력 탐지 모델의 실현 가능성을 입증하였다. 향후 연구에서는 다양한 악성코드 유형과 추가적인 국가 노드로의 확장과 함께, 데이터 이질성으로 인한 성능 저하를 완화하고 실제 글로벌 환경에서도 높은 탐지 정확도를 유지할 수 있는 보다 견고한 학습·집계 방법을 개발할 필요가 있다.

이 연구는 2025 년도 산업통상자원부 및 한국산업기술기획평가원(KEIT) 연구비 지원에 의한 연구임(과제번호 RS-2025-02653102)

참고문헌

- [1] V. Mothukuri et al., "Federated-learning-based anomaly detection for IoT security attacks," *IEEE Internet of Things J.*, vol. 9, no. 4, pp. 2545–2554, 2021.
- [2] S. A. Rahman et al., "Internet of things intrusion detection: Centralized, on-device, or federated learning?" *IEEE Network*, vol. 34, no. 6, pp. 310–317, 2020.
- [3] H. Liu et al., "Blockchain and federated learning for collaborative intrusion detection in vehicular edge computing," *IEEE Trans. Vehicular Tech.*, vol. 70, no. 6, pp. 6073–6084, 2021.
- [4] M. Botacin and H. Gomes, "Cross-regional malware detection via model distilling and federated learning," in *Proc. RAID*, pp. 97–113, 2024.
- [5] M. Botacin et al., "One size does not fit all: A longitudinal analysis of Brazilian financial malware," *ACM Trans. Privacy and Security*, vol. 24, no. 2, pp. 11:1–11:31, 2021.
- [6] Canadian Institute for Cybersecurity (UNB), "CIC-MalMem-2022: Malware Memory Analysis Dataset," 2022. [Online]. Available: <https://www.unb.ca/cic/datasets/malmem-2022.html>
- [7] H. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," *arXiv:1602.05629*, 2016.
- [8] X. Pei et al., "A knowledge transfer-based semi-supervised federated learning for IoT malware detection," *IEEE Trans. Dependable and Secure Computing*, vol. 20, no. 3, pp. 2127–2143, 2022.
- [9] Z. Lu et al., "Federated learning with non-iid data: A survey," *IEEE Internet of Things J.*, vol. 11, no. 11, pp. 19188–19209, 2024.
- [10] T. Li et al., "Federated optimization in heterogeneous networks," *arXiv:1812.06127*, 2020.
- [11] D. J. Beutel et al., "Flower: A friendly federated learning research framework," *arXiv:2007.14390*, 2022.
- [12] T. Carrier, "Detecting obfuscated malware using memory feature engineering," *Univ. of New Brunswick*, 2021.
- [13] M. Dener et al., "Malware detection using memory analysis data in big data environment," *Applied Sciences*, vol. 12, no. 17, p. 8604, 2022.
- [14] A. Mezina and R. Burget, "Obfuscated malware detection using dilated convolutional network," in *Proc. ICUMT*, pp. 110–115, 2022.
- [15] B. Ford et al., "Peer-to-peer communication across network address translators," in *Proc. USENIX Annual Tech. Conf.*, pp. 179–192, 2005.
- [16] Siemens AG and Ponemon Institute, "Caught in the crosshairs: Are utilities keeping up with the industrial cyber threat?" White paper, October 2019.
- [17] S. Gibbs, "Triton: hackers take out safety systems in 'watershed' attack on energy plant," *The Guardian*, December 2017.