PHƯƠNG PHÁP NGĂN CHẶN TẤN CÔNG ĐẦU ĐỘC CHO BỘ KHUNG LIÊN KẾT PHÁT HIỆN TÁC NHÂN ĐE DOẠ SỬ DỤNG CHIẾN LƯỢC KHÔNG GIAN TIỀM ẨN

Trần Đức Lương - 230202028

Tóm tắt

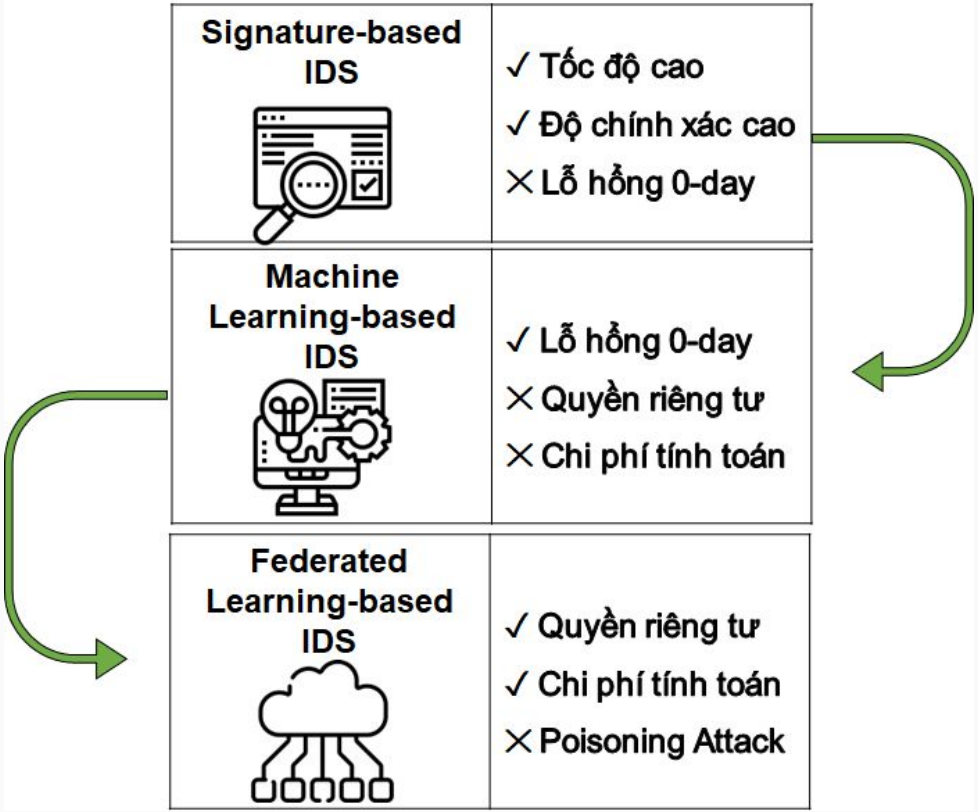
- Lớp: CS2205.CH181
- Link Github: https://github.com/ducluongtran9121/CS2205.CH181
- Link YouTube video: …
- Thành viên:

Trần Đức Lương

MSHV: 230202028

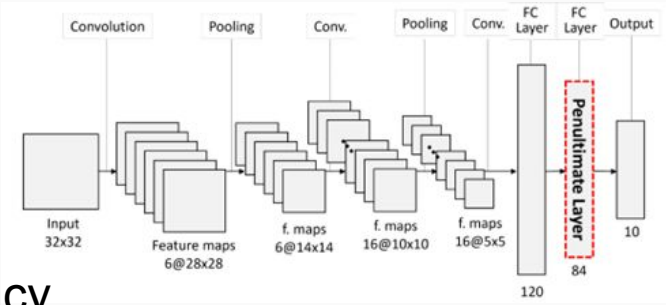
Email: luongtd.18@grad.uit.edu.vn

Giới thiệu

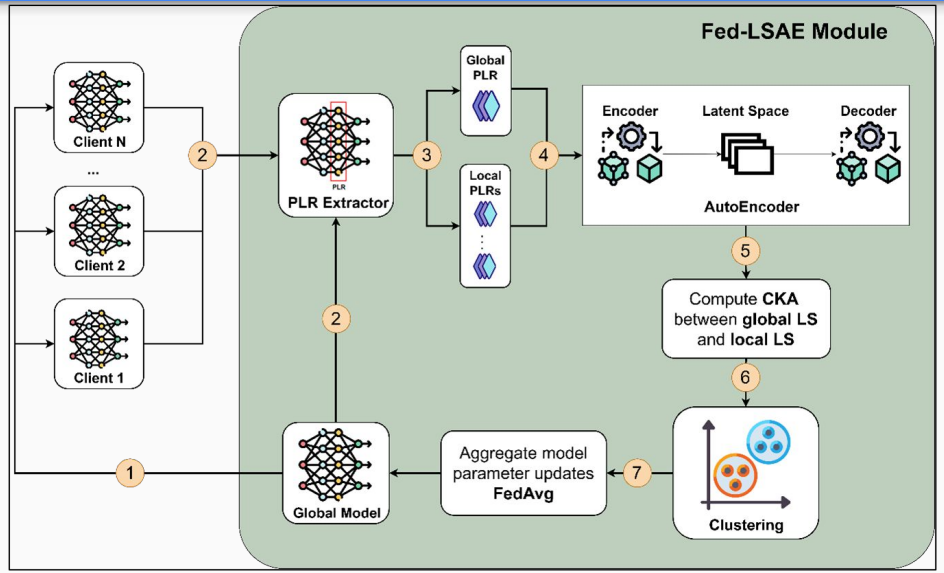

Signature-based IDS
✓ Tốc độ cao
✓ Độ chính xác cao
✕ Lỗ hổng 0-day
Machine Learning-based IDS
✓ Lỗ hổng 0-day
✕ Quyền riêng tư
✕ Chi phí tính toán
Federated Learning-based IDS
✓ Quyền riêng tư
✓ Chi phí tính toán
✕ Poisoning Attack

Giới thiệu

- Nhiều công trình [1,2,3] chống tấn công đầu độc trong FL được công bố.
- Gần đây, có phương pháp phân tích **không gian tiềm ẩn (Latent space)** để ngăn chặn **poisoning attack** trong FL.
- **Penultimate Layer Representation (PLR):**
  - FLARE [4]: Cần bộ dữ liệu bổ trợ trên máy chủ
  - FedCC [5]: Không toàn diện khi chỉ tính Accuracy
  - Chỉ đánh giá trên model poisoning



Giới thiệu



- **Đầu vào (Input):** N tổ chức huấn luyện cộng tác mô hình FL-based IDS trong môi trường tích hợp với mô-đun Fed-LSAE để chống tấn công đầu độc.
- **Đầu ra (Output):** Mô hình FL-based IDS đã được huấn luyện an toàn và có hiệu suất cao sau khi loại bỏ các thực thể tấn công đầu độc.

Mục tiêu

- Nghiên cứu sâu về 3 loại tấn công đầu độc điển hình, bao gồm cả đầu độc dữ liệu và mô hình, nhằm phá hoại ML-based IDS dựa trên FL trong bối cảnh IoT.
- Thiết kế một mô-đun phát hiện đầu độc mới có tên là Fed-LSAE bằng cách phát hiện các tham số mô hình độc hại thông qua Latent Space dựa trên AutoEncoder và PLR.
- Tiến hành một số kịch bản thử nghiệm để cho thấy hiệu quả của việc phòng thủ chống lại các cuộc tấn công đầu độc thông qua phân tích chuyên sâu về hai bộ dữ liệu về các cuộc tấn công mạng IoT với các mô hình ML khác nhau. Từ đó so sánh với các giải pháp trước đó như FLARE [4], FedCC [5], ...

Nội dung và Phương pháp

Nội dung 1: Tìm hiểu về phương pháp học cộng tác Federated Learning. Ứng dụng với Machine Learning-based IDS.

→ *Phương pháp:* Tìm hiểu cách cài đặt Federated Learning và ứng dụng vào Machine Learning-based IDS.

Nội dung 2: Tìm hiểu về tấn công đầu độc trong mô hình học cộng tác.

→ *Phương pháp:*

- Tìm hiểu về các phương pháp để tấn công đầu độc bao gồm tấn công đầu độc trên dữ liệu (data poisoning) và tấn công trên mô hình (model poisoning).
- Triển khai các cuộc tấn công đầu độc trên mô hình Federated Learning cơ bản.

Nội dung 3: Tìm hiểu về Latent space thông qua AutoEncoder và Penultimate Layer Representation (PLR).

→ *Phương pháp:*

- Tìm hiểu cơ sở lý thuyết, nghiên cứu và các ứng dụng liên quan đến PLR.
- Tìm hiểu cơ sở lý thuyết, nghiên cứu, thực nghiệm và các ứng dụng liên quan đến AutoEncoder. Sử dụng AutoEncoder để trích xuất latent space của PLR.

Nội dung và Phương pháp

Nội dung 4: Thực nghiệm và đánh giá kết quả.

→ *Phương pháp:*

- Triển khai 3 cuộc tấn công: Label-Flipping (lật nhãn), tấn công bằng mẫu sinh đối kháng GAN và model poisoning.
- Xây dựng phương pháp phát hiện đầu độc Fed-LSAE dựa trên Latent Space.
- Thực nghiệm các kịch bản tấn công khác nhau để đánh giá về hiệu năng, độ chính xác và khả năng phát triển của hệ thống Fed-LSAE so với các công trình khác như FLARE [4], FedCC [5].

Kết quả dự kiến

- Xây dựng các kịch bản tấn công đầu độc, khiến hiệu suất mô hình FL-based IDS giảm đáng kể.
- Tích hợp Fed-LSAE vào hệ thống để chống tấn công đầu độc, đưa mô hình đạt hiệu năng cao 90% trở lên.
- Fed-LSAE có khả năng phát hiện các loại tấn công đầu độc khác nhau, có hiệu suất cao hơn các giải pháp trước đó như FedCC, FLARE trong ngữ cảnh an toàn thông tin.

Tài liệu tham khảo

- [1] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, “Poisoning Attack in Federated Learning using Generative Adversarial Nets,” in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, Aug. 2019.
- [2] N. C. Vy, N. H. Quyen, P. T. Duy, and V.-H. Pham, “Federated Learning-Based Intrusion Detection in the Context of IIoT Networks: Poisoning Attack and Defense,” in *Network and System Security*, Cham, 2021, pp. 131–147. doi: 10.1007/978-3-030-92708-0_8.
- [3] Z. Zhang, Y. Zhang, D. Guo, L. Yao, and Z. Li, “SecFedNIDS: Robust defense for poisoning attack against federated learning-based network intrusion detection system,” *Future Gener. Comput. Syst.*, vol. 134, pp. 154–169, Sep. 2022.
- [4] N. Wang, Y. Xiao, Y. Chen, Y. Hu, W. Lou, and Y. T. Hou, “FLARE: Defending Federated Learning against Model Poisoning Attacks via Latent Space Representations,” in *ASIA CCS ’22: ACM Asia Conference on Computer and Communications Security, Nagasaki, Japan, 30 May 2022 - 3 June 2022*, 2022.
- [5] H. Jeong, H. Son, S. Lee, J. Hyun, and T.-M. Chung, “FedCC: Robust Federated Learning against Model Poisoning Attacks.” arXiv, Dec. 04, 2022. Accessed: Mar. 13, 2023. [Online]. Available: http://arxiv.org/abs/2212.01976.