THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube
  https://www.youtube.com/watch?v=Bty2G0HX7ck&t=43s
- Link slides:
  https://github.com/haindtdt/CS2205.RM/blob/main/CS2205.SEP2025.Nghien_Cuu_va_xay_dung_khung_hoc_sau_tuan_tu.pdf

- Họ và tên: Nguyễn Duy Hải
- MSSV: 250202007

- Lớp: 2205.RM
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 0
- Số câu hỏi QT của cả nhóm: 0
- Link Github:
  https://github.com/haindtdt/CS2205.RM

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

NGHIÊN CỨU VÀ XÂY DỰNG KHUNG HỌC SÂU TUẦN TỰ NÂNG CAO TÍNH BỀN VỮNG VÀ KIÊN CƯỜNG CHO HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG

TÊN ĐỀ TÀI TIẾNG ANH

A SEQUENTIAL DEEP LEARNING FRAMEWORK FOR ROBUST AND RESILIENT NETWORK INTRUSION DETECTION SYSTEM

TÓM TẮT

Trong bối cảnh an ninh mạng hiện đại, các hệ thống phát hiện xâm nhập (NIDS) truyền thống đang đối mặt với hai thách thức lớn: Thứ nhất là khả năng nắm bắt các mẫu tấn công phức tạp theo chuỗi thời gian của lưu lượng mạng, và thứ hai là sự mong manh trước các kỹ thuật tấn công đối kháng (Adversarial Attacks) nhắm vào chính mô hình AI. Đa số các mô hình hiện nay xử lý gói tin một cách rời rạc và dễ dàng bị đánh lừa bởi các nhiễu nhỏ trong dữ liệu đầu vào.

Đề tài này đề xuất nghiên cứu và xây dựng một khung học sâu tuần tự (Sequential Deep Learning Framework) sử dụng kiến trúc **Bi-Directional LSTM (Bi-LSTM)** kết hợp với cơ chế **Huấn luyện đối kháng (Adversarial Training)**. Mục tiêu cốt lõi là tạo ra một hệ thống NIDS không chỉ đạt độ chính xác cao trong việc phát hiện các tấn công tinh vi (tính Bền vững - Robustness) mà còn duy trì được hiệu năng ổn định khi chịu sự tác động nhiễu hoặc đánh lừa từ kẻ tấn công (tính Kiên cường - Resilience). Nghiên cứu sẽ được thực nghiệm trên tập dữ liệu chuẩn **CIC-IDS-2017** và kiểm chứng khả năng chống chịu thông qua các kịch bản tấn công giả lập sử dụng thuật toán FGSM. Kết quả nghiên cứu hứa hẹn mang lại giải pháp bảo mật tin cậy hơn cho các hạ tầng mạng trọng yếu.

GIỚI THIỆU

Sự bùng nổ của các cuộc tấn công mạng tiên tiến (Advanced Persistent Threats - APT) và mã độc đa hình đã làm lộ rõ điểm yếu của các hệ thống NIDS dựa trên chữ ký và các mô hình học máy truyền thống. Mặc dù Học sâu (Deep Learning) đã được ứng dụng rộng rãi để nâng cao khả năng phát hiện, đa số các nghiên cứu hiện tại vẫn xử lý lưu lượng mạng như các điểm dữ liệu tĩnh và rời rạc, bỏ qua tính chất liên kết chuỗi thời gian quan trọng của các luồng giao tiếp mạng (Network Flows). Điều này dẫn đến việc bỏ sót các hành vi tấn công diễn ra chậm hoặc kéo dài.

Hơn nữa, sự xuất hiện của lĩnh vực Học máy đối kháng (Adversarial Machine Learning) đặt ra một mối đe dọa nghiêm trọng mới. Kẻ tấn công có thể sử dụng các thuật toán như FGSM (Fast Gradient Sign Method) để tạo ra các mẫu dữ liệu nhiễu (perturbations) nhằm đánh lừa mô hình AI, khiến hệ thống phân loại sai lệch lưu lượng độc hại thành an toàn, trong khi con người vẫn nhìn thấy đó là dữ liệu bình thường.

Do đó, tính cấp thiết của đề tài nằm ở việc giải quyết đồng thời hai vấn đề: (1) Khai thác thông tin ngữ cảnh theo thời gian của lưu lượng mạng để phát hiện chính xác hơn thông qua mô hình tuần tự, và (2) Trang bị "kháng thể" cho mô hình AI để chống lại các cuộc tấn công lừa đảo. Đề tài tập trung nghiên cứu kiến trúc mạng nơ-ron tuần tự kết hợp với kỹ thuật phòng thủ chủ động để xây dựng một hệ thống NIDS thế hệ mới.

MỤC TIÊU

1. **Nghiên cứu kiến trúc mô hình:** Xây dựng mô hình học sâu tuần tự sử dụng mạng nơ-ron hồi quy hai chiều (Bi-LSTM) để trích xuất đặc trưng ngữ cảnh và phân loại lưu lượng mạng theo chuỗi thời gian.
2. **Phát triển cơ chế phòng thủ:** Tích hợp kỹ thuật Huấn luyện đối kháng (Adversarial Training) để nâng cao tính kiên cường (Resilience), giúp hệ thống nhận diện chính xác ngay cả khi dữ liệu đầu vào bị kẻ tấn công làm nhiễu.
3. **Đánh giá thực nghiệm:** Triển khai hệ thống, đánh giá hiệu năng trên tập dữ liệu chuẩn CIC-IDS-2017 và chứng minh khả năng chống chịu vượt trội trước các thuật toán tấn công

đối kháng so với các mô hình truyền thống.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung nghiên cứu:

- Tìm hiểu đặc trưng chuỗi thời gian của dữ liệu mạng và kiến trúc Bi-LSTM.
- Phân tích cơ chế tấn công đối kháng (Adversarial Attack) và phòng thủ.
- Xây dựng quy trình xử lý dữ liệu và huấn luyện mô hình.

Phương pháp nghiên cứu:

- **Phương pháp thu thập và tiền xử lý dữ liệu:**
  - ✓ Sử dụng bộ dữ liệu **CIC-IDS-2017**.
  - ✓ Chuẩn hóa dữ liệu (Min-Max Scaling).
  - ✓ Tạo chuỗi dữ liệu (Sequence Generation) sử dụng kỹ thuật **Cửa sổ trượt (Sliding Window)** để gom nhóm các gói tin theo dòng thời gian.
- **Phương pháp mô hình hóa:**
  - ✓ Sử dụng kiến trúc **Bi-LSTM** (Bi-Directional Long Short-Term Memory) để học sự phụ thuộc dữ liệu từ cả hai hướng quá khứ và tương lai.
  - ✓ Lớp phân loại (Dense Layer + Softmax) để đưa ra quyết định cuối cùng.
- **Phương pháp tăng cường tính Kiên cường:**
  - ✓ Sử dụng thư viện **IBM Adversarial Robustness Toolbox (ART)** để sinh các mẫu tấn công giả lập (FGSM).
  - ✓ Thực hiện **Adversarial Training**: Trộn lẫn dữ liệu sạch và dữ liệu đối kháng để huấn luyện lại mô hình.
- **Công cụ triển khai:** Python, TensorFlow/Keras, Google Colab Pro.

KẾT QUẢ MONG ĐỢI

1. Mã nguồn hệ thống (Source Code): Một chương trình hoàn chỉnh thực hiện quy trình từ tiền xử lý, huấn luyện mô hình Bi-LSTM đến phát hiện xâm nhập.

2. Hiệu năng mô hình (Robustness): Đạt độ chính xác (Accuracy) và F1-Score **> 95%** trên tập dữ liệu gốc, vượt trội hơn các mô hình máy học truyền thống.

3. Khả năng chống chịu (Resilience): Duy trì độ chính xác **> 85%** khi đối mặt với các kịch bản tấn công đối kháng giả lập (trong khi các mô hình không được bảo vệ thường giảm xuống dưới 50%).

TÀI LIỆU THAM KHẢO

[1]. Soumyadeep Hore, Jalal Ghadermazi, Ankit Shah, Nathaniel D. Bastian: A sequential deep learning framework for a robust and resilient network intrusion detection system. Comput. Secur. 144: 103928 (2024) *(Giải thích: Bài báo gốc của đề tài - Bắt buộc)*

[2]. Zongwen Yu, Ping Yi, Jiacheng Ye, Wei Wang, Yuming Qu: NIDS-Vis: Improving the generalized adversarial robustness of network intrusion detection systems. Comput. Secur. 139: 103700 (2024) *(Giải thích: Tham khảo thêm về phương pháp nâng cao tính Bền vững - Robustness, bổ trợ cho mục tiêu số 2 của đề tài)*

[3]. Vivek Kumar, Kamal Kumar, Maheep Singh, Neeraj Kumar: NIDS-DA: Detecting functionally preserved adversarial examples for network intrusion detection system using deep autoencoders. Expert Syst. Appl. 270: 126513 (2025) *(Giải thích: Tham khảo về cách phát hiện mẫu đối kháng, so sánh hiệu quả giữa Autoencoder (của bài này) và Bi-LSTM (của bạn))*

[4]. Mohamed Amine Ferrag, Leandros A. Maglaras: Constraining Adversarial Attacks on Network Intrusion Detection Systems: Transferability and Defense Analysis. IEEE Access 12: 24500-24515 (2024) *(Giải thích: Cung cấp kiến thức nền tảng về cơ chế tấn công và phòng thủ mới nhất để viết phần Giới thiệu)*

[5]. S. M. Udin, M. A. Al-Garadi: SiamIDS: A Novel Cloud-Centric Siamese Bi-LSTM Framework for Interpretable Intrusion Detection in Large-Scale IoT Networks. IEEE Internet of Things Journal (Early Access) (2025) *(Giải thích: Tham khảo kiến trúc Bi-LSTM hiện đại nhất năm 2025 để tối ưu hóa mô hình trong phần Phương pháp)*