THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): https://www.youtube.com/watch?v=NymfaMx07ps
- Link slides (dạng .pdf đặt trên Github của nhóm): https://github.com/MINHTHUKTH/CS519.O11/blob/main/CS519.O11-20521988_Slide.pdf

• Họ và Tên: Trần Thị Minh Thư
• MSSV: 20521988
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 8.5/10
	• Số buổi vắng: 2
	• Số câu hỏi QT cá nhân: 3
	• Số câu hỏi QT của cả nhóm: 6
	• Link Github: https://github.com/MINHTHUKTH/CS519.O11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng đồ án
	  ○ Viết đề cương, làm slide thuyết trình
	  ○ Làm video YouTube, Poster

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHƯƠNG PHÁP PHÁT HIỆN XÂM NHẬP HIỆU QUẢ DỰA TRÊN HỌC TỔNG HỢP KHÁNG MẪU BIẾN THỂ TRỐN TRÁNH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ADVERSARIAL SAMPLE RESISTANCE FOR ENSEMBLE LEARNING-BASED INTRUSION DETECTION SYSTEM

TÓM TẮT

Để đánh giá được khả năng phân loại của NIDS cần mô phỏng các cuộc tấn công đối kháng trong ngữ cảnh thực tế. Nhưng các phương pháp trước đây như Feature-space Blackbox Attack [4], [6] sửa đổi các tính năng trong feature-space nhưng các tính năng đó không dễ ánh xạ để tạo mẫu đối kháng thực tế và Traffic-space Blackbox Attack [3], [5] sửa đổi lưu lượng nhưng không giữ được tính độc hại ban đầu và vi phạm quy tắc giao thức truyền thông. Để giải quyết các khó khăn trong việc thực hiện tấn công đối kháng trong ngữ cảnh thực tế, một phương pháp được đề xuất là xây dựng framework GPMT dựa trên WGAN [2] có khả năng tạo ra lưu lượng đối kháng thực tế mà vẫn giữ được tính độc hại và ít kiến thức về mô hình NIDS để đảm bảo có khả năng xảy ra trong thực tế. Bên cạnh đó, thiết kế hàm mất mát giúp cải thiện khả năng phân loại và tạo ra lưu lượng đối kháng mạnh mẽ hơn trốn tránh các mô hình NIDS.

GIỚI THIỆU

Hệ thống Phát hiện Xâm nhập Mạng (NIDS) là yếu tố quan trọng trong bảo vệ an ninh và tài nguyên mạng của tổ chức ngày nay. Sự tích hợp của học máy (ML), đặc biệt là các kỹ thuật học sâu (DL) đã cải thiện hiệu suất và linh hoạt của NIDS, giúp phát hiện các mối đe dọa mới và tiến triển của tội phạm mạng [1]. Tuy nhiên, các mô hình dựa trên ML/DL có thể bị đánh lừa bởi các cuộc tấn công đối kháng, đặc biệt là trong các hệ thống nhạy cảm về bảo mật. Để đảm bảo hiệu suất của NIDS trước các

cuộc tấn công này, cần thực hiện đánh giá khả năng của hệ thống. Phương pháp phổ biến là thực hiện tấn công đối kháng để kiểm tra khả năng chống lại của NIDS. Tuy nhiên, làm thế nào để mô phỏng cuộc tấn công đối kháng trong ngữ cảnh thực tế? Và làm sao để tạo ra lưu lượng đối kháng thực tế mà không vi phạm quy tắc giao thức mạng và không biết bất cứ thông tin gì về mô hình NIDS?

Các mô hình sinh đối kháng (GAN) cũng được sử dụng để tạo ra lưu lượng đối kháng nhưng không có tính ổn định trong việc tạo ra các lưu lượng đối kháng mạnh mẽ. Trong đề tài này, một phương pháp được đề xuất để giải quyết các vấn đề trên là xây dựng framework GPMT dựa trên WGAN (Wasserstein GAN - một biến thể của GAN) [2], để tạo lưu lượng đối kháng mạnh mẽ. Đồng thời, thiết kế một hàm mất mát mới để tạo ra nhiều mẫu đối kháng mới, có khả năng đánh lừa các mô hình hộp đen NIDS. Điều này mang lại sự độc đáo và hiệu quả trong việc thử nghiệm và cải thiện tính an toàn của NIDS trước các cuộc tấn công đối kháng trong thực tế.

Input: lưu lượng độc hại ban đầu.

Output: lưu lượng đối kháng có khả năng trốn tránh mô hình hộp đen NIDS.

MỤC TIÊU

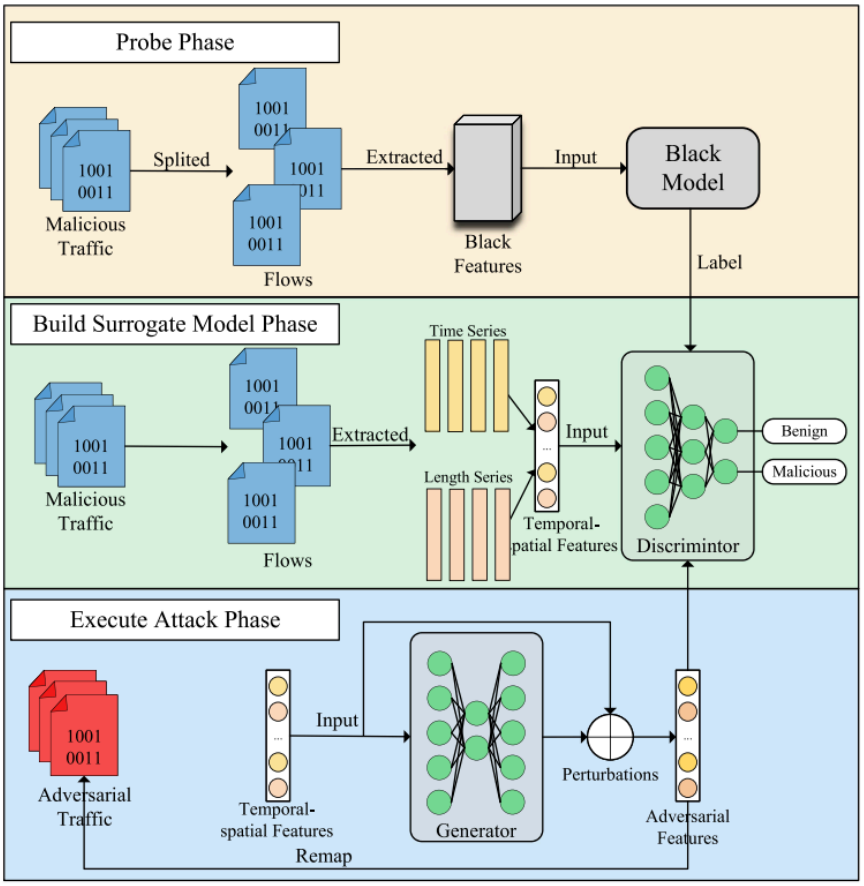
- Đào tạo 7 mô hình NIDS để đánh giá khả năng phân loại trên tập dữ liệu ISCX-Botnet, chỉnh sửa thông số cấu hình các mô hình để cải thiện khả năng phân loại.
- Xây dựng framework GPMT có khả năng tạo ra lưu lượng đối kháng từ lưu lượng độc hại ban đầu trên tập dữ liệu CTU-13. Thiết kế hàm mất mát mới giúp cải thiện hiệu suất GAN.
- Sử dụng 7 mô hình đã được đào tạo để kiểm tra tính linh hoạt của GPMT.

NỘI DUNG VÀ PHƯƠNG PHÁP

a. NỘI DUNG

- Kiến trúc framework GPMT dựa trên WGAN sẽ bao gồm 3 giai đoạn chính được thể hiện trong hình bên dưới: (1) Gửi lưu lượng độc hại ban đầu từ tập dữ liệu CTU-13 đến mô hình NIDS và có thể gán nhãn độc hại cho lưu lượng dựa

trên phản hồi của các mô hình NIDS. (2) Sử dụng lưu lượng ban đầu cùng với nhãn thu được tiến hành huấn luyện cục bộ mô hình thay thế bắt chước ranh giới quyết định của các mô hình NIDS. Mô hình thay thế ở đây sẽ là Discriminator của WGAN để có thể đào tạo lại nhiều lần. (3) Sử dụng Generator của WGAN thêm nhiễu vào lưu lượng độc hại ban đầu, cụ thể là các tính năng temporal-spatial vì các tính năng này được sử dụng phổ biến trong đào tạo mô hình phân loại và dễ ánh xạ vào không gian lưu lượng để tạo ra lưu lượng đối kháng, bằng cách trì hoãn thời gian gửi các gói lưu lượng và tăng chiều dài gói tin. Sau đó đưa vào Discriminator để tiến hành phân loại mẫu giả được tạo ra từ Generator và mẫu thật từ tập dữ liệu CTU-13. Nếu Discriminator phân biệt đúng, các lưu lượng này được đưa trở lại Generator để tiến hành chỉnh sửa tạo lưu lượng đối kháng mạnh mẽ hơn và đưa lại Discriminator để kiểm tra. Quá trình này là vòng lặp cho đến khi Discriminator bị đánh lừa bởi các lưu lượng thì lưu lượng đối kháng được tạo thành công.



Hình 1. Kiến trúc framework GPMT dựa trên WGAN

b. PHƯƠNG PHÁP

- Chọn ra 7 mô hình dựa trên học máy chính thống khác nhau làm mô hình hộp đen NIDS đào tạo trên tập dữ liệu ISCX-Botnet.
- Tìm tập dữ liệu CTU-13 ở nguồn bổ sung khác, không có giao điểm với tập ISCX-Botnet.
- Sử dụng 2 Feature Extractor khác nhau cho ISCX-Botnet và CTU-13: NFStream và CICFlowmeter.
- Tìm hiểu hạn chế của các phương pháp mô phỏng tấn công đối kháng vào mô hình hộp đen đã được nghiên cứu trước đó: Feature-space Black-box Attack ([4], [6]), Traffic-space Black-box Attack ([3], [5]).
- Giới hạn Time Series và Length Series để đảm bảo chúng thực tế khi ánh xạ vào không gian lưu lượng.
- Tìm hiểu, nghiên cứu và kiểm tra giá trị hàm mất mát mới (Wasserstein loss) trong quá trình đào tạo GPMT.
- Sử dụng các thông số đánh giá cho mô hình NIDS và framework GPMT.

KẾT QUẢ MONG ĐỢI

- Khả năng phân loại của 7 mô hình NIDS tốt trên tập ISCX-Botnet.
- GPMT đạt được hiệu quả khá tốt trong việc tạo lưu lượng đối kháng trốn tránh mô hình NIDS và có tính linh hoạt khi hiệu quả ở nhiều mô hình.
- Có thể sử dụng trong ngữ cảnh thực tế, không phá vỡ các nguyên tắc giao thức của lưu lượng truy cập mà vẫn giữ được tính độc hại ban đầu.

TÀI LIỆU THAM KHẢO

[1]. Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, Kevin A. Roundy: "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. CoRR abs/2212.14315 (2022)

[2]. Peishuai Sun, Shuhao Li, Jiang Xie, Hongbo Xu, Zhenyu Cheng, Rong Yang:

GPMT: Generating practical malicious traffic based on adversarial attacks with little prior knowledge. Comput. Secur. 130: 103257 (2023)

[3]. Dongqi Han, Zhiliang Wang, Ying Zhong, Wenqi Chen, Jiahai Yang, Shuqiang Lu, Xingang Shi, Xia Yin: Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors. IEEE J. Sel. Areas Commun. 39(8): 2632-2647 (2021)

[4]. Phan The Duy, Le Khac Tien, Nghi Hoang Khoa, Do Thi Thu Hien, Anh Gia-Tuan Nguyen, Van-Hau Pham: DIGFuPAS: Deceive IDS with GAN and function-preserving on adversarial samples in SDN-enabled networks. Comput. Secur. 109: 102367 (2021)

[5]. Milad Nasr, Alireza Bahramali, Amir Houmansadr: Defeating DNN-Based Traffic Analysis Systems in Real-Time With Blind Adversarial Perturbations. USENIX Security Symposium 2021: 2705-2722

[6]. Zilong Lin, Yong Shi, Zhi Xue: IDSGAN: Generative Adversarial Networks for Attack Generation Against Intrusion Detection. PAKDD (3) 2022: 79-91