BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học: CS519 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS519.O11

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

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

Trần Thị Minh Thư - 20521988

Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/MINHTHUKTH/CS519.O11.git
- Link YouTube video: https://www.youtube.com/watch?v=NymfaMx07ps

Trần Thị Minh Thư

Giới thiệu

Thực hiện các cuộc tấn công đối kháng đánh giá khả năng chịu đựng của NIDS dựa trên học máy.

➔ Làm thế nào để mô phỏng tấn công đối kháng có khả năng xảy ra trong thực tế?

➔ Feature-space Blackbox Attack, Traffic-space Blackbox Attack

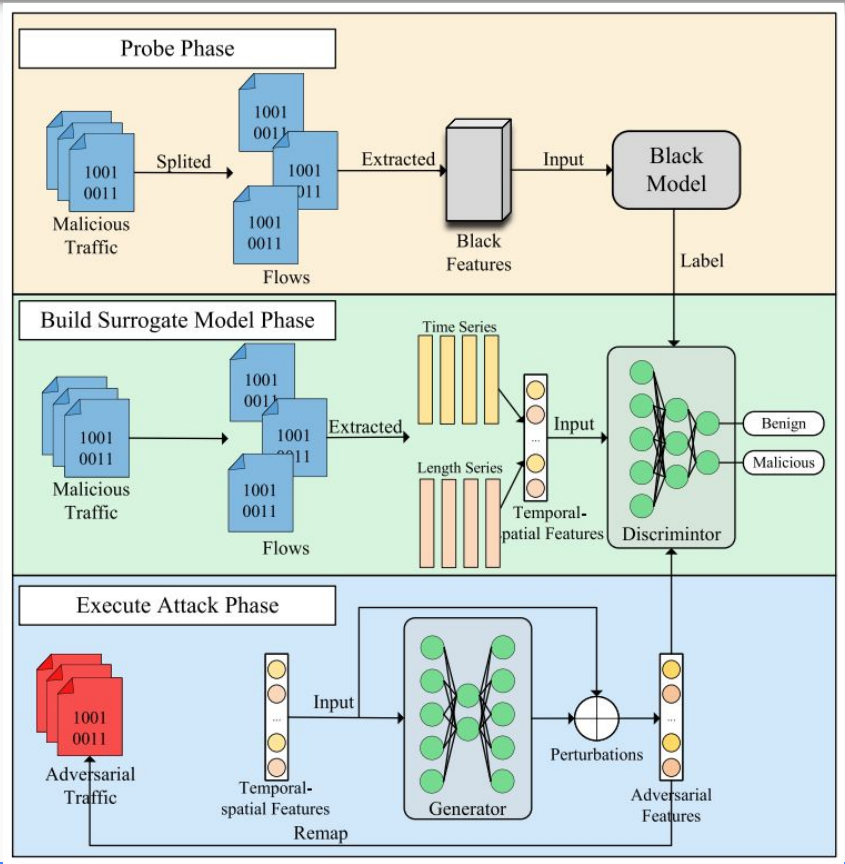
Mô hình GAN được sử dụng trong việc tạo lưu lượng đối kháng nhưng không có tính ổn định tốt trong quá trình đào tạo.

Mục tiêu

- Đào tạo và chỉnh sửa thông số cấu hình cải thiện các mô hình NIDS trên dataset ISCX-Botnet.
- Xây dựng framework GPMT dựa trên WGAN tạo lưu lượng đối kháng thực tế trên dataset CTU-13.
- Sử dụng các mô hình NIDS đánh giá mức độ hiệu quả framework GPMT.

Nội dung và Phương pháp

- 3 giai đoạn trong framework GPMT:
  - Giai đoạn thăm dò
  - Giai đoạn xây dựng mô hình thay thế
  - Giai đoạn thực hiện tấn công



Nội dung và Phương pháp

- Chọn 7 mô hình ML-based làm mô hình hộp đen NIDS, đào tạo trên tập ISCX-Botnet và tìm hiểu thông số cấu hình để cải thiện hiệu suất phân loại.
- Sử dụng 2 feature extractor cho 2 dataset ISCX-Botnet và CTU-13: NFStream và CICFlowmeter.
- Thiết kế hàm mất mát mới (Wasserstein loss) để cải thiện hiệu suất của Generator và Discriminator trong WGAN.
- Giới hạn Time Series, Length Series

Kết quả dự kiến

- Các mô hình NIDS có khả năng phát hiện, phân loại lưu lượng độc hại tốt.
- Framework GPMT tạo ra được lưu lượng đối kháng mạnh mẽ trốn tránh các mô hình NIDS.
- Các lưu lượng đối kháng có thể áp dụng được trong ngữ cảnh thực tế, 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)

Tài liệu tham khảo

[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