**ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH   
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN   
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

**ĐINH BÙI HUY PHƯƠNG  
NGUYỄN THỊ THANH MAI  
LÊ ĐOÀN TRÀ MY  
NGUYỄN PHƯƠNG TRINH**

**ĐỒ ÁN MÔN HỌC**

# ỨNG DỤNG MÁY HỌC VÀO KHỞI TẠO LƯU LƯỢNG MẠNG TRONG CÁC CUỘC TẤN CÔNG MẠNG

**GENERATING PRATICAL MALICIOUS TRAFFIC BASED ON ADVERSARIAL ATTACKS WITH LITTLE PRIOR KNOWLEDGE**

**NGÀNH AN TOÀN THÔNG TIN**

**TP. Hồ Chí Minh, 2022**

**ĐẠI HỌC QUỐC GIA HỒ CHÍ MINH   
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN   
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

**ĐINH BÙI HUY PHƯƠNG - 21520090  
NGUYỄN THỊ THANH MAI - 21521112  
LÊ ĐOÀN TRÀ MY - 21521149  
NGUYỄN PHƯƠNG TRINH - 21521581**

**ĐỒ ÁN MÔN HỌC**

# ỨNG DỤNG MÁY HỌC VÀO KHỞI TẠO LƯU LƯỢNG MẠNG TRONG CÁC CUỘC TẤN CÔNG MẠNG

**GENERATING PRATICAL MALICIOUS TRAFFIC BASED ON ADVERSARIAL ATTACKS WITH LITTLE PRIOR KNOWLEDGE**

**NGÀNH AN TOÀN THÔNG TIN**

**GIẢNG VIÊN HƯỚNG DẪN:**

**ThS. Nghi Hoàng Khoa**

**TP. Hồ Chí Minh, 2023**

**LỜI CẢM ƠN**

Trong quá trình nghiên cứu và hoàn thành đồ án môn học, nhóm đã nhận được sự định hướng, giúp đỡ và ý kiến đóng góp quý báu và lời động viên của giảng viên hướng dẫn bộ môn. Nhóm xin bày tỏ lời cảm ơn tới thầy Nghi Hoàng Khoa đã tận tình hướng dẫn, giúp đỡ trong quá trình thực hiện đồ án.

**Nhóm thực hiện**

**MỤC LỤC**

**DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT**

**DANH MỤC CÁC HÌNH VẼ**

**DANH MỤC CÁC BẢNG BIỂU**

**TÓM TẮT ĐỒ ÁN**

Trong thời đại của công nghệ thông tin, an ninh mạng đóng vai trò quan trọng trong việc bảo vệ hệ thống và dữ liệu khỏi các cuộc tấn công. Một trong những cuộc tấn công phổ biến là tấn công mạng, trong đó kẻ tấn công gửi lưu lượng mạng giả mạo, độc hại vào hệ thống mạng để gây ra hỏng hóc hoặc làm suy giảm hiệu suất. Việc phát hiện và ngăn chặn các hình thức tấn công mạng là một thách thức lớn. Một phương pháp hiệu quả để nghiên cứu và phát triển các biện pháp phòng ngừa là sử dụng các công cụ và kỹ thuật tạo lưu lượng mạng giả mạo, độc hại để kiểm tra sự bảo mật và khả năng chống chịu của hệ thống. Trong đồ án này, nhóm thực hiện giới thiệu một phương pháp mới để tạo ra lưu lượng mạng giả mạo, độc hại có tính thực tiễn, dựa trên các cuộc tấn công mạng với ít kiến thức về hệ thống mục tiêu được cho biết trước. Bài báo đề xuất một framework tấn công mới là GPMT (Generating Practical Malicious Traffic), nhằm tạo ra lưu lượng gian lận độc hại thực tế để đánh lừa các phương pháp phát hiện dựa trên Machine Learning, cụ thể là sử dụng mô hình Wasserstein GAN (WGAN) để thực hiện các cuộc tấn công và thiết kế một hàm mất mát mới, tạo ra các lưu lượng giả mạo độc hại nói trên. Qua việc tạo ra lưu lượng mạng giả mạo, độc hại, có thể thử nghiệm hiệu suất của phương pháp phát hiện và phòng ngừa hiện có, đồng thời đánh giá khả năng chống chịu và ứng phó với các cuộc tấn công mạng thực tế. Điều này giúp cải thiện hệ thống bảo mật mạng và đảm bảo sự an toàn cho dữ liệu và thông tin quan trọng.

Trong các phần tiếp theo của đồ án này, nhóm thực hiện sẽ mô tả chi tiết về phương pháp tạo lưu lượng mạng giả mạo, độc hại dựa trên bài báo, các kỹ thuật và công cụ được sử dụng, cũng như kết quả thử nghiệm và đánh giá hiệu suất. Nhóm hy vọng rằng kết quả đồ án sẽ đóng góp vào việc nâng cao hiệu quả của an ninh mạng và đảm bảo sự ổn định và an toàn cho các hệ thống mạng trong tương lai tới.

**CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN**

Chương này giới thiệu về vấn đề nghiên cứu và một số công trình liên quan. Đồng thời, trong chương này nhóm cũng trình bày phạm vi và cấu trúc của Đồ án.

**1.1. Đặt vấn đề**

Trong thời đại của công nghệ thông tin, an ninh mạng đóng vai trò quan trọng trong việc bảo vệ hệ thống và dữ liệu khỏi các cuộc tấn công. Một trong những cuộc tấn công phổ biến là tấn công mạng, trong đó kẻ tấn công gửi lưu lượng mạng giả mạo, độc hại vào hệ thống mạng để gây ra hỏng hóc hoặc làm suy giảm hiệu suất. Việc phát hiện và ngăn chặn các hình thức tấn công mạng là một thách thức lớn. Một phương pháp hiệu quả để nghiên cứu và phát triển các biện pháp phòng ngừa là sử dụng các công cụ và kỹ thuật tạo lưu lượng mạng giả mạo, độc hại để kiểm tra sự bảo mật và khả năng chống chịu của hệ thống dựa trên Machine Learning. Mặc dù machine learning (ML) đã được sử dụng ngày càng nhiều trong việc phát hiện lưu lượng mạng độc hại và đã được chứng minh là hiệu quả, nhưng việc sử dụng ML cũng có nguy cơ bị lừa bằng các ví dụ đối kháng. Một số nghiên cứu đã nghiên cứu về cuộc tấn công đối kháng đối với các phương pháp dựa trên ML, nhưng hầu hết trong số đó đều diễn ra trong các tình huống không thực tế. Điều này chủ yếu bao gồm hai khía cạnh: (i) các cuộc tấn công phản tác động yêu cầu kiến thức trước về các mô hình dựa trên ML, chẳng hạn như tập dữ liệu và đặc trưng được sử dụng bởi mô hình, điều này thường không có sẵn trong thực tế; (ii) các cuộc tấn công phản tác động tạo ra các ví dụ không thực tế, tức là các đặc trưng lưu lượng mạng hoặc lưu lượng mạng không tuân thủ các quy tắc giao thức truyền thông.

Trong bài báo này, nhóm tác giả đề xuất một framework tấn công các ví dụ đối kháng có tên là GPMT (Generating Practical Malicious Traffic), nhằm tạo ra lưu lượng gian lận độc hại thực tế để đánh lừa các phương pháp phát hiện dựa trên ML. So với các công trình trước đây, GPMT có những ưu điểm: (i) yêu cầu ít kiến thức trước: giới hạn kiến thức trước có sẵn để mô phỏng các cuộc tấn công hộp đen cho các tình huống thực tế; (ii) ví dụ đối kháng thực tế hơn: chúng tôi sử dụng mô hình Wasserstein GAN (WGAN) để thực hiện các cuộc tấn công và thiết kế một hàm mất mát mới, tạo ra các ví dụ đối kháng thực tế, có khả năng lừa các phương pháp phát hiện. Tiến hành tấn công chín mô hình dựa trên ML trong tập dữ liệu CTU-13 để chứng minh tính hiệu quả của khung công cụ. Kết quả thử nghiệm cho thấy GPMT hiệu quả và linh hoạt hơn so với các phương pháp khác.

**1.2. Công trình liên quan**

Các cuộc tấn công adversarial cho việc phát hiện thông tin độc hại dựa trên máy học có thể được chia thành hai loại: white-box model và black-box model.

**1.2.1. Tấn công White-box**

Trong white-box attack, kẻ tấn công biết tất cả chi tiết về mô hình và có thể thay đổi đặc trưng giao thông để đánh lừa mô hình.

***a, Feature-space white-box attack (FWA) – Tấn công hộp trắng không gian đặc trưng***

Trong tấn công này, kẻ tấn công biết thông tin đầy đủ về mô hình mục tiêu và chỉ thay đổi các đặc trưng lưu lượng để lừa các mô hình dựa trên máy học. Aiken và Scott-Hayward (2019) đề xuất một phương pháp để né tránh việc các thuật toán máy học trong cuộc tấn công phủ định phân tán (DDoS). Họ lựa chọn ba đặc trưng để thêm sự biến đổi và thành công giảm tỷ lệ phát hiện của các mô hình từ 100% xuống 0%. Tuy nhiên, phương pháp này chỉ tập trung vào cuộc tấn công DDoS mà không thảo luận về các lớp tấn công khác, và các đặc trưng được lựa chọn có phạm vi tương đối hẹp.

Huang và cộng sự (2018) tiến hành các thử nghiệm trên ba thuật toán học sâu. Họ chủ yếu sử dụng ba biện pháp: Fast gradient sign method (FGSM), jacobian-based saliency map attack (JSMA), và JSMA đảo ngược (JSMA-RE) Bằng cách thay đổi một số đặc trưng, họ đã đáng kể làm suy yếu hiệu quả của việc phát hiện mô hình.

FWA là cuộc tấn công dễ nhất để thực hiện.nhưng FWA chỉ có thể chứng minh lý thuyết khả thi của cuộc tấn công adversarial trong các tình huống đặc biệt.

***b, Traffic-space white-box attack (TWA) – Tấn công hộp trắng lưu lượng đặc trưng***

Trong tấn công này, kẻ tấn công biết tất cả các chi tiết về mô hình mục tiêu và ánh xạ các đặc trưng được tạo ra vào không gian lưu lượng thực. Hashemi và đồng nghiệp (2019) tạo ra các đặc trưng đối kháng và sau đó chuyển đổi chúng thành lưu lượng ban đầu.

Sadeghzadeh và đồng nghiệp (2021) đề xuất ba cuộc tấn công mới để đưa những biến đổi đối kháng vào lưu lượng mạng. Kết quả là đã giảm đáng kể hiệu suất của các bộ phân loại lưu lượng mạng dựa trên học sâu.

Chernikova và Oprea (2022) đề xuất một framework dựa trên gradient lặp chung để tạo ra các cuộc tấn công trốn tránh. Họ cho thấy với sự cố gắng tối thiểu (ví dụ: tạo ra 12 kết nối mạng bổ sung) có thể thay đổi các dự đoán của mô hình.

**1.2.2. Tấn công Black-box**

Trong black-box attack, kẻ tấn công không biết về mô hình và cố gắng tấn công mô hình mà không có thông tin trước.

***a, Feature-space black-box attack (FBA) – Tấn công hộp đen không gian đặc trưng***

Trong tấn công này, kẻ tấn công không biết về mô hình và các thông số của mô hình mục tiêu, và kẻ tấn công có thể tạo ra các đặc trưng phản đối để thoát khỏi mô hình dựa trên học máy.

Lin và đồng nghiệp (2022) đề xuất một mô hình phản đối dựa trên GAN (Goodfellow et al., 2020) để tấn công các mô hình phát hiện lưu lượng, gọi là IDSGAN. IDSGAN được chia thành ba phần: một hệ thống phát hiện xâm nhập (IDS) là mô hình black-box bị tấn công, một bộ phân loại là mô hình phân loại để học mô hình black-box, và một bộ tạo là mô hình tạo ra các ví dụ phản đối độc hại. Thực nghiệm cho thấy các ví dụ phản đối được tạo ra bởi IDSGAN có thể giảm đáng kể tỷ lệ phát hiện của các mô hình.

Usama và đồng nghiệp (2019) cải tiến IDSGAN và đảm bảo rằng tất cả các đặc trưng được sửa đổi không có chức năng và các đặc trưng chức năng vẫn giữ nguyên.

Duy và đồng nghiệp (2021) đề xuất một khung thức cải tiến dựa trên WGAN, gọi là DIGFuPAS, và thực hiện các thí nghiệm trên nhiều bộ dữ liệu để chứng minh tính linh hoạt và tổng quát của DIGFuPAS.

Jmila và đồng nghiệp (2022) đánh giá các phương pháp phát hiện dựa trên học máy thông qua nhiều phương pháp tấn công đối kháng. Kết quả cho thấy các cuộc tấn công không ảnh hưởng đồng đều đến tất cả các bộ phân loại.

Zolbayar và đồng nghiệp (2022) sửa đổi các đặc trưng mạng để tấn công dưới ràng buộc của miền. Họ tạo ra các luồng lưu lượng đối kháng cực kỳ chân thực để lừa các hệ thống phát hiện dựa trên học máy.

Tuy nhiên, FBA vẫn chỉ là mô phỏng lý thuyết của cuộc tấn công đối kháng. Các đặc trưng được sửa đổi trực tiếp vào mô hình dựa trên học máy. Mặc dù tỷ lệ phát hiện của mô hình giảm đáng kể, nhưng không xem xét các đặc trưng có thể được ánh xạ vào không gian lưu lượng thực tế hay không.

***b, Traffic-space black-box attack (TBA) – Tấn công hộp đen lưu lượng đặc trung***

Các phương pháp TBA đã được đề xuất nhằm tạo ra lưu lượng giao thông phản đối thực tế trong các tình huống black-box.

Han và đồng nghiệp (2021) sử dụng các kỹ thuật học máy để tìm kiếm các đặc trưng nằm tại ranh giới quyết định của mô hình học máy.

Cheng và đồng nghiệp (2021) đề xuất Attack-GAN để tạo ra lưu lượng mạng giao thông phản đối ở mức gói tin và tuân thủ các ràng buộc miền.

Nasr và đồng nghiệp (2021) áp dụng một thuật toán biến đổi phản đối mù cho các bộ phân loại phân tích lưu lượng. Thuật toán biến đổi phản đối được áp dụng trực tiếp trên lưu lượng thời gian thực. Để duy trì các đặc trưng của lưu lượng, họ thiết kế một hàm ánh xạ lại. Phương pháp này có thể áp dụng cho nhiều bộ phân loại lưu lượng khác nhau.

Tuy nhiên, các phương pháp TBA hiện tại vẫn chỉ là kịch bản. Không chỉ mô hình và các thông số phải không hiển thị cho kẻ tấn công mà các đặc trưng và tập dữ liệu cũng phải không hiển thị. Lý thuyết, các ranh giới quyết định kết quả sẽ tương tự khi các mô hình black-box khác nhau được huấn luyện trên cùng các đặc trưng và dữ liệu huấn luyện. Điều này có nghĩa là kẻ tấn công sẽ thu được nhiều kiến thức kinh nghiệm về mô hình black-box.

**1.3. Cấu trúc đồ án môn học**

Nhóm xin trình bày nội dung đồ án theo cấu trúc như sau:

Chương 1: Giới thiệu tổng quan về đề tài.

Chương 2: Sơ lược kiến thức nền tảng liên quan.

Chương 3: Phương pháp và thiết kế hệ thống của mô hình.

Chương 4: Hiện thực và đánh giá kết quả.

Chương 5: Kết luận và hướng phát triển của đề tài.

**CHƯƠNG 2. KIẾN THỨC NỀN TẢNG**

Chương này trình bày sơ lược cơ sở lý thuyết của nghiên cứu, bao gồm: Hệ thống phát hiện xâm nhập, mô hình học máy, mô hình mạng sinh đối kháng.

**2.1. Hệ thống phát hiện xâm nhập – IDS (Intrusion Detection Systems)**

Hệ thống phát hiện xâm nhập - IDS (Intrusion Detection Systems) là một phần mềm hoặc công cụ được sử dụng để bảo vệ hệ thống và cảnh báo về các hành vi nghi ngờ xâm nhập vào hệ thống. Mục tiêu chính của IDS là ngăn chặn và phát hiện các hành vi gây hại đối với tính bảo mật của hệ thống, chẳng hạn như các cuộc tấn công quét cổng hoặc dò tìm.

Có nhiều loại IDS khác nhau, mỗi loại có chức năng và nhiệm vụ riêng. Dưới đây là một số loại IDS phổ biến:

- NIDS (Network Intrusion Detection Systems): Thường được triển khai tại các điểm dễ bị tấn công trong hệ thống mạng. NIDS giám sát lưu lượng mạng đến (inbound) và đi (outbound) từ tất cả các thiết bị trong mạng. Ưu điểm lớn nhất của NIDS là có thể quét tất cả lưu lượng mạng, nhưng điều này có thể làm giảm tốc độ mạng tổng thể.

- HIDS (Host Intrusion Detection Systems): Thường được triển khai trên các thiết bị trong hệ thống có khả năng kết nối Internet. HIDS chỉ giám sát các gói dữ liệu vào (inbound) và ra (outbound) từ thiết bị hoặc các hành vi đáng ngờ với truy cập nội bộ.

- IDS dựa trên chữ ký (Signature-Based IDS): Hoạt động dựa trên các chữ ký đã được học, giám sát các gói tin trên mạng như cách phần mềm diệt virus hoạt động. Tuy nhiên, phương pháp này có thể không phát hiện được các mối đe dọa mới mà chưa có chữ ký tương ứng.

- IDS dựa trên bất thường (Anomaly-Based IDS): Hoạt động dựa trên việc phát hiện sự bất thường. IDS này giám sát lưu lượng mạng và so sánh với một mô hình cơ sở đã được thiết lập trước đó. Mô hình cơ sở xác định các hoạt động bình thường của mạng và cảnh báo cho quản trị viên hoặc người dùng khi phát hiện lưu lượng truy cập không bình thường hoặc khác biệt so với mô hình cơ sở.

**2.2. Mô hình máy học Machine Learning**

Học máy hay máy học (Machine Learning) là một lĩnh vực thuộc trí tuệ nhân tạo (AI) mà trong đó các ứng dụng phần mềm có khả năng tự động học và cải thiện từ dữ liệu mà không cần được lập trình cụ thể. Học máy sử dụng các thuật toán và mô hình để phân tích dữ liệu lịch sử và dự đoán kết quả mới.

Có nhiều ứng dụng phổ biến của học máy, bao gồm:

- Hệ thống đề xuất (Recommendation Systems): Hệ thống đề xuất sử dụng học máy để đưa ra các gợi ý hoặc khuyến nghị cho người dùng, ví dụ như gợi ý sản phẩm khi mua sắm trực tuyến hoặc khuyến nghị nội dung trên các nền tảng truyền thông xã hội.

- Phát hiện bất thường (Anomaly Detection): Học máy có thể được sử dụng để phát hiện các điểm dữ liệu không bình thường hoặc ngoại lệ trong một tập dữ liệu. Điều này hữu ích trong việc phát hiện gian lận tín dụng, phát hiện hành vi đáng ngờ trong mạng, hoặc giám sát các hệ thống để phát hiện các sự cố.

- Phát hiện xâm nhập (Intrusion Detection): Học máy có thể giúp xác định các hành vi xâm nhập trong hệ thống máy tính hoặc mạng. Các mô hình học máy có thể học từ các mẫu dữ liệu xâm nhập đã biết và phát hiện các hành vi tương tự trong thời gian thực.

- Phân loại phần mềm độc hại (Malware Classification): Học máy có thể được sử dụng để phân loại và nhận diện phần mềm độc hại, giúp bảo vệ hệ thống khỏi các mối đe dọa của virus, mã độc và phần mềm độc hại khác.

- Lọc thư rác (Spam Filtering): Học máy có thể được áp dụng để phân loại và lọc ra các email rác (spam) từ hộp thư đến của người dùng, giúp cải thiện trải nghiệm người dùng và bảo vệ khỏi các cuộc tấn công qua email.

Có bốn cách tiếp cận cơ bản trong học máy:

- Học có giám sát (Supervised Learning): Mô hình được huấn luyện trên dữ liệu huấn luyện đã được gắn nhãn, trong đó đầu vào và đầu ra đã được xác định. Mô hình học từ các mẫu dữ liệu đã biết và sau đó có thể dự đoán kết quả cho các dữ liệu mới.

- Học không giám sát (Unsupervised Learning): Mô hình học từ dữ liệu không gắn nhãn và tìm kiếm các mẫu, cấu trúc hoặc nhóm tự nhiên trong dữ liệu. Mô hình không được cung cấp đầu ra đúng hoặc gắn nhãn trước.

- Học bán giám sát (Semi-Supervised Learning): Kết hợp cả học có giám sát và học không giám sát, trong đó một phần dữ liệu huấn luyện được gắn nhãn và một phần không được gắn nhãn. Mô hình sử dụng thông tin từ cả hai loại dữ liệu để học và dự đoán kết quả.

- Học tăng cường (Reinforcement Learning): Trong học tăng cường, một tác tử học cách đạt được mục tiêu trong một môi trường. Tác tử tương tác với môi trường thông qua các hành động và nhận được phần thưởng hoặc hình phạt dựa trên hành động đó. Mục tiêu của tác tử là tối đa hóa phần thưởng thu được thông qua việc tìm hiểu và tối ưu hóa chiến lược hành động.

**2.3. Mô hình mạng sinh đối kháng (GAN/WGAN)**

**2.3.1. GAN là gì?**

Mô hình mạng sinh GAN - Generative Adversarial Networks, là mô hình học máy đặc biệt được đề xuất bởi Ian Goodfellow vào năm 2014. GAN bao gồm hai thành phần chính: một mạng sinh (generator) và một mạng phân biệt (discriminator). Mạng sinh cố gắng tạo ra những mẫu dữ liệu mới từ một không gian ngẫu nhiên, trong khi mạng phân biệt cố gắng phân biệt giữa các mẫu thực và mẫu được tạo ra bởi mạng sinh. Hai mạng này cạnh tranh và cải thiện lẫn nhau thông qua quá trình huấn luyện, tạo ra những mẫu dữ liệu mới có chất lượng ngày càng tốt.

**2.3.2. WGAN là gì?**

WGAN (Wasserstein GAN) là một biến thể của GAN được đề xuất bởi Martin Arjovsky, Soumith Chintala và Léon Bottou vào năm 2017 để giải quyết một số vấn đề gặp phải trong quá trình huấn luyện các mô hình GAN truyền thống. WGAN sử dụng một hàm mất mát dựa trên khoảng cách Wasserstein thay vì sử dụng hàm mất mát thông thường như cross-entropy. Kỹ thuật này giúp giảm hiện tượng "vanishing gradient" và cải thiện sự ổn định và chất lượng của quá trình huấn luyện.

*\* Khoảng cách Wasserstein (Wasserstein distance), còn được gọi là khoảng cách Earth Mover (Earth Mover's distance), là một phép đo khoảng cách giữa hai phân phối xác suất. Nó được sử dụng trong mô hình WGAN để đo lường sự khác biệt giữa phân phối dữ liệu thực và phân phối dữ liệu được tạo ra bởi mạng sinh.*

*\* Trong học máy, "vanishing gradient" là hiện tượng xảy ra khi các đạo hàm của hàm mất mát giảm đáng kể khi lan truyền ngược qua các tầng ẩn của mạng nơ-ron. Điều này đồng nghĩa với việc gradient truyền đi trở nên rất nhỏ, dẫn đến tốc độ học chậm và khó khăn trong việc cập nhật các trọng số của mạng.*

**2.3.3. Cấu trúc mạng**

Mô hình mạng GAN được cấu tạo bởi 2 mạng nơ-ron luôn hoạt động đối nghịch nhau: bộ sinh (Generator) và bộ phân biệt (Discriminator).

- Generator: Học cách sinh ra dữ liệu giả để lừa mô hình Discriminator. Để có thể đánh lừa được Discriminator thì đòi hỏi mô hình sinh ra output phải thực sự tốt. Do đó chất lượng dữ liệu phải càng như thật càng tốt.

- Discriminator: Học cách phân biệt giữa dữ liệu giả được sinh từ mô hình Generator với dữ liệu thật. Discriminator như một giáo viên chấm điểm cho Generator biết cách nó sinh dữ liệu đã đủ tinh xảo để qua mặt được Discriminator chưa và nếu chưa thì Generator cần tiếp tục phải học để tạo ra dữ liệu thật hơn. Đồng thời Discriminator cũng phải cải thiện khả năng phân biệt của mình vì chất lượng dữ liệu được tạo ra từ Generator càng ngày càng giống thật hơn. Thông qua quá trình huấn luyện thì cả Generator và Discriminator cùng cải thiện được khả năng của mình.

**CHƯƠNG 3. PHƯƠNG PHÁP & THIẾT KẾ HỆ THỐNG CỦA MÔ HÌNH**

**3.1. Quy trình tấn công đối kháng để phát hiện lưu lượng truy cập độc hại dựa trên ML**

A diagram of a security system

Description automatically generated

*Hình . Quy trình tấn công đối kháng để phát hiện lưu lượng truy cập độc hại dựa trên ML*

Quy trình tấn công đối kháng để phát hiện lưu lượng truy cập độc hại dựa trên ML gồm các giai đoạn:

- Giai đoạn Thăm dò: Kẻ tấn công (Attacker) tấn công vào mục tiêu và tạo các lưu lượng độc hại.

- Giai đoạn Phản hồi: Model phát hiện (detection) tìm ra lưu lượng độc hại và đưa ra phản hồi để chặn chúng. Attacker gán nhãn lưu lượng dựa trên việc nó đã bị chặn hay không.

- Giai đoạn Tấn công: Attacker tìm ra lưu lượng có các đặc trưng vượt qua được model phát hiện dựa trên nhãn. Attacker huấn luyện attack model dựa trên lưu lượng và nhãn, dùng nó để sửa đổi các lưu lượng độc hại (ở giai đoạn thăm dò) cho đến khi vượt qua model phát hiện.

- Giai đoạn ByPass (vượt qua): Lưu lượng vượt qua model phát hiện thành công.

Trong bài báo, thực hiện sửa đổi các đặc điểm/đặc trưng không gian, thời gian của lưu lượng bằng cách trì hoãn thời gian gửi gói lưu lượng và tăng độ dài của gói để đánh lừa mô hình phòng thủ.

**3.2. Kiến trúc mô hình**

Kiến trúc mô hình này dựa trên WGAN:

A diagram of a black model

Description automatically generated

*Hình . Tổng quan kiến trúc mô hình*

Mô hình này gồm 3 giai đoạn:

- Probe phase (Giai đoạn thăm dò): Attacker gửi lưu lượng mạng độc hại tới black-box model và đánh label dựa trên phản hồi của black-box model

- Build surrogate model phase (Giai đoạn xây dựng model thay thế): Từ label thu được từ Probe phase, xây dựng mô hình bắt chước black-box model. Discriminator (bộ phân biệt D) của mô hình WGAN được chọn để làm surrogate model.

- Execute attack phase (Giai đoạn thực hiện tấn công): Đây là quá trình tranh đấu giữa generator G và discriminator D của WGAN. G được dùng để sinh lưu lượng mạng có hại còn D sẽ cố gắng phân biệt lưu lượng mạng từ D là benign hay malicious. Các kết quả sau đó được đưa vào lại D và G để tiếp tục huấn luyện

Do các kết quả thuộc tính được sinh ra là các thuộc tính về thời gian và kích thước lưu lượng mạng nên dễ dàng chuyển đổi lưu lượng mạng để giống với thực tế mà vẫn có thể gây hại.

**3.2.1. Giai đoạn thăm dò (Probe Phase)**

A diagram of a block diagram

Description automatically generated

*Hình. Mô hình của giai đoạn thăm dò*

|  |  |
| --- | --- |
| A white background with black text  Description automatically generated  *Hình. Thuật toán giai đoạn thăm dò* | 1. Khởi tạo mô hình phát hiện hộp đen đã được huấn luyện và danh sách nhãn rỗng.  2. Với mỗi lưu lượng đầu vào trong danh sách flows từ f1 đến fn:  + Gán nhãn li bằng cách đưa lưu lượng hiện tại fi qua model phát hiện M.  + Thêm nhãn li vào danh sách L.  3. Kết thúc vòng lặp.  → Kết quả của quá trình này là danh sách nhãn L, chứa các nhãn tương ứng với từng lưu lượng trong danh sách flows ban đầu. |

- Attacker gửi lưu lượng độc hại đến mục tiêu hộp đen nhưng không biết các đặc trưng (features) được black-box model sử dụng, lưu lượng được trích xuất dưới dạng black features; đồng thời attacker cũng không biết tập huấn luyện được sử dụng mởi model của mục tiêu.

→ Không thể tấn công mục tiêu cụ thể, lưu lượng độc hại và tập huấn luyện khác nhau.

- Thông tin mà Attacker thu được từ black-box mục tiêu là nhãn (label) từ mỗi luồng.

- Tường lửa làm gián đoạn các kết nối độc hại đến máy chủ của attacker và phát hiện các lưu lượng độc hại, những lưu lượng bị phát hiện này được gắn nhãn. Attacker lấy lưu lượng độc hại và nhãn tương ứng tiến hành tấn công đối kháng.

**3.2.2. Giai đoạn xây dựng model thay thế (build surrogate model phase)**

A diagram of a process

Description automatically generated

*Hình. Mô hình của giai đoạn xây dựng model thay thế*

|  |  |
| --- | --- |
| *Hình. Thuật toán giai đoạn xây dựng model thay thế* | 1. Đặt trích xuất đặc trưng FE và danh sách đặc trưng lưu lượng TF rỗng.  2. Đối với mỗi lưu lượng đầu vào trong danh sách flows từ f1 đến fn:  + Trích xuất đặc trưng tại thời điểm t từ lưu lượng hiện tại fi bằng cách áp dụngtrích xuất đặc trưng FE.  + Thêm đặc trưng tại thời điểm t fi vào danh sách đặc trưng lưu lượng TF.  3. Khởi tạo bộ phân biệt D.  4. Với một số lần lặp trong quá trình huấn luyện:  + Lấy mẫu một lô gồm các đặc trưng và nhãn tương ứng {tf(j), l(j)} từ danh sách đặc trưng và nhãn.  + Tính toán giá trị dự đoán trước đó pre(j) bằng cách đưa đặc trưng tại thời điểm t f(j) qua bộ phân biệt D.  + Cập nhật bộ phân biệt D dựa trên công thức  5. Kết thúc vòng lặp.  Kết quả của quá trình này là mô hình D đã được huấn luyện, được sử dụng trong các giai đoạn huấn luyện đối kháng tiếp theo. |

- Attacker mô phỏng các đặc trưng của model mục tiêu để mô phỏng ranh giới quyết định của model mục tiêu.

- Đặc trưng (feature) về không gian - thời gian: Chọn chuỗi thời gian và chuỗi độ dài gói tin làm đặc trưng thời gian - không gian. (*Sử dụng đặc trưng về không gian - thời gian vì chúng được sử dụng bởi hầu hết các mô hình phát hiện lưu lượng dựa trên học máy (ML)→ mô phỏng mô hình hộp đen tốt hơn ; chuỗi thời gian và chuỗi không gian có thể được ánh xạ dễ dàng hơn vào không gian lưu lượng thực tế*).

- Mô hình thay thế: Discriminator (bộ phân biệt D) của mô hình WGAN được chọn để làm surrogate model vì có thể huấn luyện lại trong các quá trình huấn luyện tiếp theo, tạo nhiều ví dụ/mẫu đối kháng hơn. Đầu vào của D là đặc trưng thời gian-không gian của luồng dữ liệu và đầu ra là phân loại nhị phân, kết quả đầu ra được dùng như nhãn.

**3.2.3. Giai đoạn thực hiện tấn công (Execute attack phase): “trận đấu” giữa bộ sinh G và bộ phân biệt D của WGAN**

A diagram of a machine

Description automatically generated

*Hình. Mô hình của giai đoạn thực hiện tấn công*

Là một quá trình lặp để thực hiện giai đoạn tấn công bằng cách sử dụng mô hình sinh/khởi tạo (G) và mô hình phân biệt (D). Mục tiêu của cuộc tấn công là tạo ra các đặc trưng đối kháng (TF\*) dựa trên các đặc trưng ban đầu (TF).

|  |  |
| --- | --- |
| *Hình. Thuật toán giai đoạn tấn công* | 1. Bộ khởi tạo (Generator) G, ncritic (số lần lặp của D trong mỗi lần lặp của G).  2. Với số lần lặp trong huấn luyện:  +Lấy mẫu một lô gồm các đặc trưng tf(j) từ danh sách đặc trưng.  + Với mỗi đặc trưng tf(j) trong đó:  ++ Tạo đặc trưng đối kháng tại thời điểm tf(j) = cách áp dụng G vào đặc trưng tf(j) và cộng với tf(j).  ++ Tính toán giá trị dự đoán mới pre\*(j) bằng cách đưa đặc trưng đối kháng tại thời điểm t tf\*(j) qua bộ phân biệt D.  ++ Cập nhật G dựa trên công thức:  ++ Cập nhật bộ phân biệt D dựa trên công thức:  sử dụng giá trị dự đoán mới pre\*(j) và pre(j).  3 Kết thúc vòng lặp.  Kết quả của quá trình này là danh sách đặc trưng phản đối TF\* đã được tạo ra từ danh sách đặc trưng ban đầu TF. |

- Bộ sinh G được sử dụng để tạo các đặc trưng không gian - thời gian đối kháng. Đầu vào của G là các đặc điểm độc hại, đầu ra là các nhiễu loạn được thêm vào các đặc trưng độc hại → tạo ra các đặc trưng đối kháng.

- Bộ phân biệt D để xác định xem lưu lượng là độc hại (giả) hay lành tính. Đầu vào của D là các đặc điểm không gian thời gian đối kháng và lành tính, còn đầu ra là phân loại nhị phân.

- Quá trình huấn luyện khác biệt so với giai đoạn tiền huấn luyện là sử dụng các đặc trưng đối kháng thay thế cho các đặc trưng độc hại. Do quá trình tiền huấn luyện, D có thể hội tụ nhanh hơn và nhạy hơn với các ví dụ/mẫu đối kháng. Sau đó, thông tin từ D được phản hồi để huấn luyện G và D phù hợp.

- Chiến lược huấn luyện: WGAN dựa chủ yếu vào hàm mất mát LG và LD của mô hình sinh G và mô hình phân biệt D để thực hiện huấn luyện lặp lại. Trong quá trình backpropagation của LG, các tham số của D được đóng băng, trong khi trong quá trình backpropagation của LD, các tham số của G được đóng băng.

+ G được huấn luyện bằng cách cập nhật trọng số tương ứng với thông tin gradient từ điểm số của D theo công thức:



+ Trong huấn luyện D, các tính năng/đặc trưng độc hại được thay thế bằng các tính năng đối kháng theo công thức:



→ D là một mô hình giả mạo mô hình hộp đen, tấn công mô hình D giúp ta hiểu được cách mô hình hộp đen hoạt động và tìm ra các điểm yếu của nó; liên tục cập nhật mô hình phân biệt D giúp làm cho mô hình sinh G trở nên đối kháng hơn. Nếu mô hình D không được cập nhật, mô hình sinh G sẽ nhanh chóng rơi vào một giải pháp cục bộ tối ưu và không đạt được hiệu suất tốt trong việc tạo ra các đặc trưng đối kháng.

**3.2.4. Lập bản đồ tới không gian lưu lượng**

- Vấn đề với mô hình tấn công đối kháng hiện tại: Nếu chỉ sử dụng mô hình tấn công đối kháng để tạo ra đặc trưng độ dài và chuỗi thời gian, công việc này không thể được sử dụng trong thực tế.

- Thiết kế phương pháp ánh xạ đặc trưng trong không gian dữ liệu thực tế: thiết kế một tập hợp các phương pháp ánh xạ đặc trưng trong không gian dữ liệu lưu lượng thực tế để đảm bảo tính thực tế của dữ liệu, bằng cách áp dụng giới hạn cho cả chiều dài và chuỗi thời gian, như được hiển thị trong phương trình:



*Trong công thức trên, chúng ta có các thành phần sau:*

*X̃\*: Biểu diễn cho kết quả cuối cùng sau quá trình ánh xạ.*

*x: Đầu vào ban đầu, có thể là chuỗi dữ liệu về độ dài hoặc chuỗi thời gian.*

*G(x): Hàm generator (mạng sinh) được sử dụng trong mô hình GAN để tạo ra dữ liệu mới từ đầu vào x.*

*ε: Một giá trị nhỏ dương, được sử dụng để giới hạn sự thay đổi tối đa của dữ liệu đầu vào.*

*Min(G(x), ε): Lấy giá trị nhỏ nhất giữa đầu ra của generator G(x) và ε. Điều này đảm bảo rằng sự thay đổi của dữ liệu không vượt quá ε.*

*(x + Min(G(x), ε)): Tổng của đầu vào ban đầu và giá trị nhỏ nhất giữa đầu ra của generator và ε. Điều này tạo ra một dữ liệu mới kết hợp giữa dữ liệu ban đầu và biến đổi được tạo ra bởi generator.*

*σ: Giới hạn trên cho dữ liệu đầu ra, đảm bảo rằng kết quả cuối cùng không vượt quá giới hạn này.*

*Tổng quát, công thức này áp dụng một phép biến đổi có giới hạn lên dữ liệu ban đầu, kết hợp với đầu ra của generator G(x). Điều này nhằm đảm bảo rằng dữ liệu mới đáng tin cậy và thực tế, không vượt quá giới hạn đã định trước.*

- Cân bằng giữa mô hình sinh G và mô hình phân biệt D: Trong quá trình đối kháng, mô hình phân biệt D bắt đầu bằng cách sao chép ranh giới quyết định của mô hình hộp đen thông qua quá trình huấn luyện trước. Mô hình sinh G cố gắng lừa mô hình phân biệt D, mô hình này có khả năng phát hiện tương tự mô hình hộp đen. Điều này giúp thực hiện tấn công đối kháng gián tiếp đối với mô hình hộp đen.

**CHƯƠNG 4. HIỆN THỰC VÀ ĐÁNH GIÁ KẾT QUẢ**

**4.1. Xây dựng mô hình phát triển lưu lượng độc hại**

- Dataset: Sử dụng dataset ISCX để train và test model, ISCX-Botnet gồm 8 labels (1 benign, 7 attack).

- Lựa chọn model black-box (9 model) : LR, SVM, NB, DT, RF, KNN, MLP, CNN, RNN.

- Lựa chọn features cho model: Sử dụng thư viện NFStream để trích xuất thông tin thời gian, kích thước từ tập dataset.

*Trích xuất các đặc trưng thống kê thời gian (như giá trị nhỏ nhất, trung bình, độ lệch chuẩn và giá trị lớn nhất của khoảng thời gian giữa các gói tin) và đặc trưng thống kê không gian (như giá trị nhỏ nhất, trung bình, độ lệch chuẩn và giá trị lớn nhất của khoảng thời gian giữa các kích thước gói tin).*

- Xử lý dữ liệu: Sử dụng SMOTE để xử lý bất cân bằng dữ liệu và dùng Min-Max normalization để chuẩn hóa dữ liệu.

*Dữ liệu chia 50%train:50%test, epoch size: 50, batch size: 128, tốc độ học của D và G: 0.0001, ngưỡng cắt trọng số của D khi train: 0.01.*

- Kết quả:

*Bảng. Kết quả xử lý dữ liệu*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Precision | Model | Precision | Model | Precision |
| DT | 98.2% | BYS | 83.77% | RFC | 98.04% |
| SVM | 82.41% | LR | 80.43% | KNN | 96.55% |
| MLP | 93.55% | CNN | 84.34% | RNN | 93.80% |

**A screenshot of a computer

Description automatically generated**

*Hình. Kết quả thực hiện xử lý dữ liệu*

**4.2. Giai đoạn thu thập nhãn cho malicious traffic**

- Dataset: Sử dụng dataset CTU-13, cụ thể là dataset về botnet Neris và Rbot. Dataset này chỉ chứa các malicious traffic; CTU-13 gồm 2 label (2 attack).

- Trích xuất: Sử dụng thư viện pcap-splitter để tách traffic thành từng flow, mỗi flow là luồng gói tin ứng với 4 thông tin (src\_ip, dst\_ip, src\_port, dst\_port).

- Thực hiện gửi mỗi flow tới black-box model ở bước trước và thu thập output của model. Do sử dụng 9 model khác nhau nên kết quả cuối cùng là kết quả có số lượng nhiều nhất.

- Kết quả:

+ Lượng flow gửi: 32,410.

+ Lượng flow có output là malicious: 24,717.

→ Precision = 76.26%

- Mỗi flow được chia vào 2 folder Attack và Normal ứng với output từ black-box model.

A screenshot of a computer

Description automatically generated

*Hình. Folder Attack và Normal*

**4.3. Giai đoạn xây dựng mô hình giả**

- Dataset: Sử dụng dữ liệu thu được từ bước trước.

- Model: Sử dụng 7 model LR, SVM, NB, DT, RF, KNN, MLP.

- Lựa chọn feature cho model: Sử dụng thông tin về thời gian giữa các gói tin và kích thước các gói tin.

- Kết quả:

*Bảng.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Precision | Model | Precision | Model | Precision |
| DT | 94.24% | BYS | 79.41% | RFC | 94.39% |
| SVM | 96.01% | LR | 76.93% | KNN | 93.91% |
| MLP | 98.69% |  |  |  |  |

A screenshot of a computer screen

Description automatically generated

*Hình. Kết quả xây dựng mô hình giả*

**4.4. Giai đoạn huấn luyện WGAN**

- Discriminator là nhóm các model giả vừa được huấn luyện ở bước trước.

- Quá trình huấn luyện WGAN xảy ra giữa G() và D() với hàm loss tương ứng.

A black and white text

Description automatically generated



- G() muốn giá trị LG là tối thiểu ↔ Dữ liệu được tạo ra giống thật.

- D() muốn giá trị LD là tối thiểu ↔ Có khả năng đánh giá thật giả tốt.

- Kết quả: Mô hình G() có khả năng tạo dữ liệu giả giống thật.

A screenshot of a computer

Description automatically generated

*Hình.*

**CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI**

Trong đồ án này, nhóm tập trung vào kiến trúc mô hình và thực hiện mô phỏng lại cách tạo ra lưu lượng đối kháng thực tế để đánh lừa các hệ thống phát hiện dựa trên máy học mà không cần biết thông tin chi tiết về mô hình phát hiện bằng GPMT. Mô hình này được xây dựng trên mô hình GAN để tạo ra các chuỗi dữ liệu về không gian và thời gian trong không gian lưu lượng. Đồng thời, trong bài báo nhóm tác giả cũng đã thiết kế một phương pháp ánh xạ đặc trưng để đảm bảo tính thực tế của dữ liệu được tạo ra. Kết quả thực nghiệm cho thấy khả năng của phương pháp trong tạo ra các chuỗi dữ liệu tấn công phản đối hiệu quả, mô phỏng ranh giới quyết định của mô hình hộp đen. Các thí nghiệm đã chứng minh rằng GPMT có thể ảnh hưởng mạnh mẽ đến mô hình mục tiêu. Hơn nữa, chúng tôi khám phá tác động của GPMT lên các tập dữ liệu khác nhau một cách riêng biệt và chứng minh thực nghiệm rằng GPMT là đa năng.

Tuy nhiên, hiện chỉ mới thực hiện Binary Classification​ trong tương lai, có thể mở rộng nghiên cứu sang Multi-classification. Đồng thời, có thể tìm thêm và sửa đổi nhiều đặc trưng hơn để đảm bảo rằng tấn công phản đối phù hợp để tấn công một loạt các mô hình toàn diện hơn hay những yếu tố có thể thay đổi mà không ảnh hưởng đặc tính của lưu lượng mạng​.