**1. Giới thiệu tổng quan**

\* Ngữ cảnh:

- Các cuộc nghiên cứu tấn công đối kháng dựa trên ML trước không thực tế: biết trước về các tham số, thông tin của hệ thống; lưu lượng mạng không thực tế (tính năng lưu lượng/lưu lượng không tuân thủ quy tắc giao thức truyền thông)

- Tạo 1 framework mới GMPT, tạo lưu lượng mạng mới vừa thực tế, vừa bảo đảm đặc điểm gây hại của cuộc tấn công​ (đánh lừa phát hiện dựa trên ML)

- Ưu điểm :

+ little prior knowledge (ít kiến thức có sẵn): mô phỏng tấn công black-box

+ more adversarial and practical examples (nhiều ví dụ đối kháng và thực tế hơn): sử dụng WGAN thực hiện tấn công và tạo hàm loss mới?

+ 9 mô hình dựa trên ML trong bộ dữ liệu CTU-13

+ Kết quả: EIR có thể đạt 65, 53%, cao hơn 16,48% so với DIGFuPAS (tốt nhất trong 9)

**2. Phương pháp thực hiện**

- Quy trình tấn công đối kháng để phát hiện lưu lượng truy cập độc hại dựa trên ML:

+ GĐ Thăm dò: Attacker tấn công vào mục tiêu và tạo các lưu lượng độc hại

+ GĐ 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.

+ GĐ 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 (GĐ thăm dò) cho đến khi vượt qua model phát hiện

+ GĐ ByPass (vượt qua): Lưu lượng vượt qua model phát hiện thành công  
A diagram of a security system

Description automatically generated

- 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.

\* Kiến trúc mô hình: dựa trên WGAN

A diagram of a black model

Description automatically generated

- 3 giai đoạn:

+ Probe phase (GĐ 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 (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: 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

2.1. Giai đoạn thăm dò (Probe Phase)

A diagram of a block diagram

Description automatically generated

A white background with black text

Description automatically generated

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. Attacker không biết các (features) được black-box model sử dụng, lưu lượng được trích xuất dưới dạng black features; - Attacker 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 là 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 hát 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

2.2 Giai đoạn xây dựng model thay thế (build surrogate model phase): xây dựng từ lưu lương và nhãn

- Attacker mô phỏng các tính năng của model mục tiêu để mô phỏng ranh giới quyết định của model mục tiêu

- feature về không gian-thời gian: Chọn chuỗi thời gian và chuỗi độ dài gói tin làm đặc trưng thời gian-không gian ( vì 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 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).

A diagram of a process

Description automatically generated

A white paper with black text

Description automatically generated

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 dụng trích xuất đặc trưng FE.

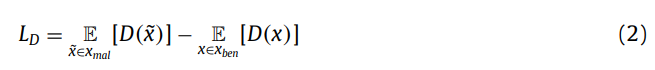
+ Thêm đặc trưng tại thời điểm t fi vào danh sách đặc trưng lưu lượng TF.

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 tf(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.

2.3. Execute attack phase: “trận đấu” giữa bộ khởi tạo G và bộ phân biệt D của WGAN

A diagram of a machine

Description automatically generated

A close up of a math equation

Description automatically generated with medium confidence  
A screenshot of a computer program

Description automatically generated

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).

1. Bộ khởi tạo (Generator) G, ncritic (số lần lặp của bộ phân biệt trong mỗi lần lặp của G).

2. Với một số lần lặp trong quá trình huấn luyện:

+Lấy mẫu một lô gồm các đặc trưng tf(j) từ danh sách đặc trưng.

+ Với mỗi đặc trưng tf(j) trong lô:

+ Tạo đặc trưng đối kháng tại thời điểm t 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 (3): 

+ Cập nhật bộ phân biệt D dựa trên công thức (4)  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 điểm/đặc trưng không gian-thời gian đối kháng/đối kháng. Đầu vào của G là các đặc điểm độc hại, đầu ra là các nhiễu loạn được thêm vào các đặc điểm độc hại → đặc điểm/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 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.



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



→ 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.

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 

- 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.

\* Xử lý dữ liệu:

- Sử dụng NFStream2 – framework Python để trích xuất tính năng của các mô hình mục tiêu.

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).

- Tiền xử lý dữ liệu: Sử dụng phương pháp tăng cường dữ liệu SMOTE để giải quyết vấn đề mất cân bằng dữ liệu. Dữ liệu cũng được chuẩn hóa bằng phương pháp Max-Min normalization. Dữ liệu chia 50%train:50%test, epoch size 50, batch size 128, learning rate of D and G is 0.0001, cutting weight threshold of D when training is 0.01.  
- Mô hình surrogate được kiểm tra mỗi 10 epoches và model gây khó khăn cho mô hình surrogate sẽ được chọn làm mô hình sinh.

- Lựa chọn mô hình black-box: sử dụng 9 mô hình máy học khác nhau để so sánh, gồm: LR, SVM, NB, DT, RF, KNN, MLP, CNN, RNN

- Sử dụng dataset ISCX-Botnet để train surrogate model, ISCX-Botnet (8 labels: 1 benign, 7 attack)  
- Sử dụng dataset CTU-13 để tấn công surrogate model (CTU-13 (2 label: 2 attack)