**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

**KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG**

A picture containing text, vector graphics, clipart

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

MÔN HỌC: HỆ THỐNG TÌM KIẾM, PHÁT HIỆN

VÀ NGĂN NGỪA XÂM NHẬP

Đề tài:

“Ứng dụng GAN tấn công qua mặt/chống lại ML-based IDPS”

Giảng viên hướng dẫn:

Lớp: NT204.M11.ATCL

Sinh viên thực hiện:

Phạm Nguyễn Việt Tân – 19520262

Phạm Ngọc Thành – 19520958

Nguyễn Khải Đăng – 19521317

Phạm Bảo Hà – 19521457

Tp.Hồ Chí Minh, Tháng 11 năm 2021

# Mục lục

[Mục lục 2](#_Toc88353909)

[Danh sách hình vẽ 4](#_Toc88353910)

[Giới thiệu đề tài 5](#_Toc88353911)

[Phương pháp nghiên cứu 7](#_Toc88353912)

[A. Nội dung 01: Tìm hiểu về IDPS 7](#_Toc88353913)

[B. Nội dung 02: Tìm hiểu về máy học và mạng sinh đối kháng 7](#_Toc88353914)

[C. Nội dung 03: Tìm hiểu về các dataset dùng để training 7](#_Toc88353915)

[D. Nội dung 04: Nghiên cứu phân tích, đánh giá mô hình IDSGAN 7](#_Toc88353916)

[E. Nội dung 05: Thực hiện hóa mô hình IDSGAN trên máy tính, thử nghiệm tấn công các thuật toán máy học dùng làm IDS 7](#_Toc88353917)

[Nội dung thực hiện 8](#_Toc88353918)

[Nội dung 01: Tìm hiểu về IDPS 8](#_Toc88353919)

[Nội dung 02: Tìm hiểu về máy học và mạng sinh đối kháng 8](#_Toc88353920)

[A. Học máy 8](#_Toc88353921)

[1. Cây quyết định (Decision Tree) 8](#_Toc88353922)

[2. Logistic Regression 10](#_Toc88353923)

[3. Random Forest 12](#_Toc88353924)

[4. SupportVectorMachine(SVM) 12](#_Toc88353925)

[B. Học sâu 14](#_Toc88353926)

[1. Multi-layer Perceptron 14](#_Toc88353927)

[C. Mạng sinh đối kháng (GAN) 15](#_Toc88353928)

[Nội dung 03: Tìm hiểu về các dataset dùng để training 17](#_Toc88353929)

[A. DARPA (Lincoln Laboratory 1998-99) 17](#_Toc88353930)

[B. KDD’99 (University of California, Irvine 1998-99): 17](#_Toc88353931)

[C. DEFCON (The Shmoo Group, 2000-2002): 18](#_Toc88353932)

[D. Dataset NSL-KDD 18](#_Toc88353933)

[E. Dataset CICIDS2017 19](#_Toc88353934)

[Nội dung 04: Nghiên cứu phân tích, đánh giá mô hình IDSGAN 19](#_Toc88353935)

[A. Cấu trúc IDSGAN: 19](#_Toc88353936)

[2. Trình tạo (generator): 20](#_Toc88353937)

[3. Trình phân biệt (discriminator): 20](#_Toc88353938)

[4. Thuật toán huấn luyện 21](#_Toc88353939)

[Nội dung 05: Thực hiện hóa mô hình IDSGAN trên máy tính, thử nghiệm tấn công các thuật toán máy học dùng làm IDS 22](#_Toc88353940)

[A. Thực hiện tiền xử lý dữ liệu 22](#_Toc88353941)

[1. Xử lý thô: 23](#_Toc88353942)

[2. Normalize data (Chuẩn hóa dữ liệu) 24](#_Toc88353943)

[3. Separate (Phân loại dữ liệu) 25](#_Toc88353944)

[4. Split (Tách dữ liệu) 25](#_Toc88353945)

[B. Thực hiện huấn luyện Black-box IDS 25](#_Toc88353946)

[C. Thực hiện huấn luyện GAN 26](#_Toc88353947)

[D. Thực hiện tạo mẫu đối kháng từ GAN và lưu vào datasheet 28](#_Toc88353948)

[Bảng phân công công việc 29](#_Toc88353949)

[Tài liệu tham khảo 30](#_Toc88353950)

# Giới thiệu đề tài

Những năm gần đây, đi kèm với sự bùng nổ của cách mạng công nghiệp 4.0, số người sử dụng internet và dịch vụ mạng đang ngày một tăng. Theo [12], cho đến nay đã có hơn 5.1 tỷ người sử dụng internet trên toàn thế giới. Tăng 45% về số lượng người sử dụng internet chỉ trong 5 năm. Bên cạnh đó, các công nghệ như IoT, cloud computing, social networks, cũng đã góp phần tạo nên một khối dữ liệu khổng lồ trên mạng interent. Theo Cisco visual networking index [Cisco,VNI (2017)] trong năm 2022, global IP traffic sẽ đạt tới 4.8 zettabytes mỗi năm, tương ứng với 396 exabytes mỗi tháng. Tăng hơn 224% so với năm 2017. Dẫn đến một lượng lớn dữ liệu lên đến hàng nghìn petabytes sẽ được tạo ra và lưu thông trên internet hàng ngày với tốc độ chóng mặt. Từ đó đã có các công nghệ kĩ thuật hiện đại hơn được sinh ra như đám mây và các hệ thống điện toán hiệu suất cao, để lưu trữ và xử lý các tập dữ liệu khổng lồ. Chúng ta còn có các công cụ phân tích dữ liệu tiên tiến cho phép chúng ta chắt lọc thông tin hữu ích từ dữ liệu sau đó đưa ra dự đoán các xu hướng và sự kiện. Điều này sẽ mở ra một số cơ hội cho các ứng dụng mới sử dụng nhiều dữ liệu trong một số lĩnh vực khác nhau, chẳng hạn như sản xuất và quản lý năng lượng, quản lý chăm sóc sức khỏe và cuộc sống đô thị. Tuy nhiên, song song với sự gia tăng về nhu cầu và lợi ích mà dữ liệu lớn mang lại là rủi ro về bảo mật và an ninh mạng ngày càng cao. Đặc biệt là phương pháp phát hiện, phân biệt chính xác và kịp thời giữa luồng mạng bình thường và các cuộc tấn công mạng với tài nguyên máy tính còn hạn chế.

Với sự phát triển của các mối đe dọa bảo mật trên Internet, hệ thống phát hiện và ngăn ngừa xâm nhập (IDPS) đã trở thành công cụ cần thiết để phát hiện và phòng tránh tấn công mạng được thể hiện dưới dạng lưu lượng độc hại. Mục đích chính của IDPS là phân loại giữa bản ghi mạng bình thường và bất thường thông qua các dữ liệu mà hệ thống có được trước đó, hoặc thông qua các phương pháp dự đoán.

Đối với vấn đề phân loại, các thuật toán máy học đã được áp dụng trong IDPS và đạt được những kết quả khả quan như Support Vector Machine, Cây quyết định, K-Nearest Neighbor,... Trong những năm gần đây, các thuật toán học sâu phát triển nhanh và thúc đẩy sự phát triển trong trong lĩnh vực phát hiện xâm nhập như Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Auto Encoder, v.v. [2]. Các thuật toán này giúp cải thiện độ chính xác và đơn giản hóa bài toán phát hiện xâm nhập [3] trong các hệ thống mạng. Các thuật toán này giúp cải thiện độ chính xác và đơn giản hóa bài toán phát hiện xâm nhập trong các hệ thống mạng.

Tuy nhiên, hệ thống phát hiện xâm nhập dần dần lộ ra lỗ hổng của nó khi gặp phải các mẫu đối kháng: input gần giống với input gốc nhưng được phân loại không chính xác. Và Mạng sinh đối kháng (GAN) là phương pháp tiềm năng cho các cuộc tấn công đối kháng như vậy.

Với sự phát triển nhanh chóng của các thuật toán học máy, việc tạo ra các mẫu đối kháng của các thuật toán học máy đã thu hút các nhà nghiên cứu quan tâm và áp dụng trong nhiều lĩnh vực. Là một lĩnh vực quan trọng và nhạy cảm, bảo mật thông tin đang đối mặt với nhiều thách thức hơn từ các cuộc tấn công đối kháng. Rất nhiều nghiên cứu tập trung vào việc tạo ra các mẫu đối kháng độc hại trong bảo mật.

Ở đề tài này, chúng em khảo sát một số mô hình tiếp cận GAN, các giải pháp sử dụng cách tiếp cận đó, các dataset IDS; phân tích, so sánh ưu nhược điểm và thực nghiệm chạy một mô hình tấn công qua mặt IDS sử dụng các thuật toán machine learning bằng GAN, được gọi là IDSGAN, trên máy tính.

# Phương pháp nghiên cứu

Gồm 5 nội dung chính: Tìm hiểu về IDPS; tìm hiểu về máy học và mạng sinh đối kháng; tìm hiểu về các dataset dùng để training; nghiên cứu phân tích, đánh giá mô hình IDSGAN; thực hiện hóa mô hình IDSGAN trên máy tính, thử nghiệm tấn công các thuật toán máy học dùng làm IDS.

## A. Nội dung 01: Tìm hiểu về IDPS

• Mục tiêu: Nắm được kiến trúc, nguyên tắc hoạt động của IDPS.

• Phương pháp: Ôn lại kiến thức đã được học trong môn học “HTPH”.

## B. Nội dung 02: Tìm hiểu về máy học và mạng sinh đối kháng

• Mục tiêu: Nắm được các khái niêm cơ bản trong máy học, học sâu; Tìm hiểu các thuật toán ML/DL và tìm hiểu cơ chế hoạt động của mạng GAN, Wasserstein GAN.

• Phương pháp: Nghiên cứu tài liệu về máy học, mạng GAN, WGAN..

## C. Nội dung 03: Tìm hiểu về các dataset dùng để training

• Mục tiêu: Nắm được cách tạo ra dataset dùng để training, tìm hiểu sơ về các dataset như NSL-KDD, CICIDS2017, ưu điểm và nhược điểm của các dataset này, và sẽ tìm hiểu sâu hơn vào NSL-KDD, dataset được dùng cho mô hình IDSGAN .

• Phương pháp: Nghiên cứu tài liệu về dataset, các review paper.

## D. Nội dung 04: Nghiên cứu phân tích, đánh giá mô hình IDSGAN

• Mục tiêu: Tìm hiểu về nghiên cứu sử dụng Wasserstein GAN tấn công qua mặt IDS thông qua mô hình IDSGAN.

• Phương pháp: Nghiên cứu tài liệu về IDSGAN.

## E. Nội dung 05: Thực hiện hóa mô hình IDSGAN trên máy tính, thử nghiệm tấn công các thuật toán máy học dùng làm IDS

• Mục tiêu: Triển khai mô hình IDSGAN tạo ra các mẫu lưu lượng độc hại; Thử nghiệm tấn công IDS sử dụng lưu lượng độc hại

• Phương pháp: Xây dựng IDSGAN từ code có sẵn, thử nghiệm tấn công IDS bằng lưu lượng đối kháng từ IDSGAN, thống kê kết quả.

# Nội dung thực hiện

## Nội dung 01: Tìm hiểu về IDPS

### Giới thiệu:

Trong những thập kỷ qua, Internet và hệ thống máy tính đã nâng cao nhiều vấn đề về bảo mật do sự bủng nổ của việc sử dụng mạng. CERT báo cáo rằng số lượng của các cuộc tấn công càng ngày càng tăng cao qua từng năm. Bất kỳ sự câm nhập hay tấn công lổ hổng mạng, máy tính hoặc hệ thống thông tin đều có thể gây ra thảm hoạ nghiêm trọng, và vi phạm chính sách bảo mật máy tính: Tính bảo mật (Confidentiality), Tính toàn vẹn (Integrity) và Tính khả dụng (Availability) – CIA. Cho đến nay, các mối đe doạ trong mạng và bảo mật thông tin vẫn còn là vấn đề cần nghiên cứu. Mặc dù một số tài liệu khảo sát và phân loại IDS hiện có (Denning, 1987; Lunt, 1993; Mukherjee et al., 1994; Debar et al., 1999; Axelsson, 2000; Mishra et al., 2004; Krugel and Toth, 2000; Jones and Sielken, 2000; Debar et al., 2000; Mukkamala and Sung, 2003; Estevez-Tapiador et al., 2004; Delgado et al., 2004; Kabiri and Ghorbani, 2005; Anantvalee and Wu, 2007; Patcha and Park, 2007; Tucker et al., 2007; Mandala et al., 2008; Garcia- Teodoro et al., 2009; Amer and Hamilton, 2010; Xie et al., 2011), nhóm cố gắng đưa ra hệ thống hoá, kiến trúc và hình ảnh đương đại hơn để đánh giá toàn diện.

Đầu tiên, nhóm tạo sự phân biệt rõ ràng giữa xâm nhập, phát hiện xâm nhập, hệ thống phát hiện xâm nhập (IDS) và hệ thống ngăn chặn xâm nhập (IPS). NIST (Bace and Mell, 2001) mô tả sự xâm nhập là một nỗ lực thoả hiệp với CIA, hoặc vượt qua cơ chế bảo mật của máy tính hoặc mạng, phát hiện xâm nhập là quá trình theo dõi các sự kiện xảy ra trong hệ thống máy tính hoặc mạng, và phân tích dấu hiệu của xâm nhập. Đặc biệt, gần đây mạng không dây được triển khai rộng rãi, và chúng rất dễ bị tấn công hơn bất cứ mạng dây. Trong các nghiên cứu gần đây (Pelechriní et al., 2011; Tan et al., 2011), nhiều kiểu tấn công từ chối dịch vụ không dây (WdoS) đã được phân tích. Do đó, nhóm phân loại IDS dựa trên công nghệ không dây và các loại công nghệ khác. Hệ thống phát hiện và ngăn ngừa xâm nhập là hệ thống phần mềm hoặc phần cứng tự động hoá quá trình phát hiện xâm nhập (Bace and Mell, 2001; Stavroulakis and Stamp, 2010). Hơn thế nữa, hệ thống ngăn chặn xâm nhập là hệ thống có tất cả khả năng của IDS, và có thể cố gắng ngăn chặn các sự cố có thể xảy ra (Stavroulakis and Stamp, 2010). Trong một số bài viết, thuật ngữ hệ thống phát hiện và ngăn chăn xâm nhập (IDPS) và IPS là từ đồng nghĩa, tuy nhiên thuật ngữ IDPS hiếm khi được sử dụng trong cộng đồng bảo mật. Trong bài báo này, nhóm tập trung vào khảo sát và phân loại các kỹ thuật liên quan đến IDS, và đưa ra so sánh ngắn gọn giữa chúng.

Mặt khác, điện toán đám mây thúc đẩy các công nghệ hiện nay, như là ảo hoá và điện toán phân tán, và gần đây nổi lên như một mô hình mới để lưu trữ và cung cấp dịch vụ qua Internet. Ảo hóa là một công nghệ trừu tượng hóa các chi tiết của phần cứng vật lý và cung cấp khả năng tổng hợp các tài nguyên tính toán từ các cụm dịch vụ, kho lưu trữ và mạng cho các ứng dụng cấp cao. Nền tảng đám mây tận dụng công nghệ ảo hóa để đạt được mục tiêu cung cấp tài nguyên máy tính như một tiện ích. Do đó, chúng tôi cũng nghiên cứu các vấn đề bảo mật trên Máy ảo (VM).

Nhóm chia các phần như sau: mô tả các phương pháp luận IDS ở phần B, và phân loại các phương pháp tiếp cận IDS ở phần C. Phần D giới thiệu bốn nhóm công nghệ IDS.

### Phương pháp luận phát hiện:

Các phương pháp phát hiện xâm nhập được phân thành 3 loại chính: phát hiện xâm nhập dựa trên dấu hiệu (SD), phát hiện xâm nhập dựa trên hành vi bất thường (AD), và phân tích trạng thái giao thức (SPA). Khái niệm phát hiện xâm nhập dựa trên dấu hiệu (SD) được mô tả như sau – Signature là một mẫu hoặc chuỗi tương ứng với một cuộc tấn công hoặc mối đe doạ đã biết trước. SD là một quá trình so sánh các signature với các sự kiện quan sát được để xác định, nhận ra các hành vi xâm nhập có thể xảy ra. Vì SD sử dụng những kiến thức tích luỹ được từ các cuộc tấn công và lỗ hổng hệ thống, SD còn được gọi là knowledge-based hoặc Misue Detection (phát hiện lạm dụng). Phát hiện xâm nhập dựa trên hành vi bất thường (AD) – Một bất thường là sự sai lệch với hành vi đã biết, và SD tạo ra 1 profile đại diện cho các hành vi bình thường/ dự kiến trong mạng, hầu hết được tạo ra thông qua theo dõi các hoạt động thường xuyên, kết nối mạng, máy chủ hoặc người dùng trong 1 khoảng thời gian. Profile có thể là tĩnh hoặc động, và phát triển cho nhiều thuộc tính, ví dụ: các lần đăng nhập thất bại, sử dụng bộ xử lý, số lượng email,… Sau đó, AD so sánh profile bình thường với các sự kiện được quan sát để nhận ra các cuộc tấn công. AD còn được gọi là Behavior-based Detection trong 1 số bài báo. Một số ví dụ của AD: cố gắng đột nhập, giả mạo, xâm nhập bởi người dùng hợp pháp, từ chối dịch vụ (DOS),…

Hơn nữa, phân tích giao thức trạng thái (SPA) – Trạng thái trong SPA chỉ ra rằng IDS có thể biết và theo dõi các trạng thái giao thức (ví dụ: ghép các yêu cầu với câu trả lời). Tưởng rằng tiến trình SPA giống ADs, nhưng về cơ bản chúng khác nhau. AD sử dụng mạng preload hoặc cấu hình máy chủ cụ thể, còn SPA phụ thuộc vào profile tổng quát cung cấp bởi nhà sản xuất quy định 1 giao thức nên làm gì và không nên làm gì. Thông thường, các mô hình giao thức mạng dựa trên giao thức tiêu chuẩn tổ chức tiêu chuẩn quốc tế như IETF. SPA còn được gọi là Specification-based Detection. Hybrid – Nhiều IDSs kết hợp nhiều phương pháp để tăng khả năng phát hiện và chính xác hơn. Ví dụ, SD và AD là phương pháp bổ sung cho nhau, vì SD phát hiện các cuộc tấn công/ mối đe doạ đã biết, còn AD tập trung vào các cuộc tấn công chưa biết.

### Tiếp cận xâm nhập

Theo truyền thống, người ta nghiên cứu các phương pháp tiếp cận xâm nhập từ 2 quan điểm chính, phát hiện bất thường (AD) và phát hiện lạm dụng (MD), nhưng sự khác biệt giữa chúng là không đáng kể. Stavroulakis and Stamp (2010) đã đề xuất phân loại để chia nhỏ các cách tiếp cận này thành 3 mục nhỏ gồm: tiếp cận dựa vào tính toán, trí tuệ nhân tạo, và các khái niệm sinh học. Tuy nhiên, việc phân loại như vậy quá khó để có thể thấy được toàn bộ mối quan hệ của tiếp cận xâm nhập. Bởi vì thiếu cái nhìn chi tiết của tiếp cận xâm nhập, nhóm trình bày phân loại của 5 nhóm con với góc nhìn chuyên sâu về đặc trưng của chúng: dựa trên thống kê, dựa trên khuôn mẫu, dựa trên quy tắc, dựa trên trạng thái và dựa trên Heuristic.

Tiếp cận dựa trên thống kê chủ yếu dựa trên ngưỡng được xác định trước, giá trị trung bình và độ lệch chuẩn, xác suất để xác định các cuộc xâm nhập. Phát hiện dựa trên khuôn mẫu tập trung vào các cuộc tấn công đã biết thông qua sự so khớp dữ liệu. Hơn thế nữa, các quy tắc If – Then hay If – Then – Else được áp dụng trong kỹ thuật dựa trên quy tắc, để xây dựng mô hình và hồ sơ của các lần xâm nhập đã biết. Đặc biệt, phương pháp dựa trên trạng thái khai thác hữu hạn trạng thái của máy bắt nguồn từ các hành vi mạng để xác định các cuộc tấn công. Cuối cùng là phương pháp tiếp cận dựa trên heuristic, lấy cảm hứng từ các khải niệm sinh học và trí tuệ nhân tạo. Các công trình gần đây (Fragkiadakis et al.,2012; Mar et al., 2012; Kartit et al., 2012; Farooqi et al.,2012; Modi et al., 2012; Wang et al., 2011; Couture, 2012; Li et al., 2012) tích hợp một số phương pháp phát hiện của 5 lớp con thánh một phương pháp tinh vi để mang lại hiệu quả tốt hơn và tỷ lệ cảnh báo sai thấp hơn so với các phương pháp tiếp cận riêng lẻ.

### Các loại công nghệ

Hiện nay, có rất nhiều loại công nghệ IDS. Chúng tôi phân loại công nghệ thành 4 nhóm tuỳ theo nơi chúng được đặt để kiểm tra các hoạt động đáng ngờ, và loại sự kiện chúng có thể nhận biết (Mukherjee et al., 1994; Stavroulakis and Stamp, 2010; Sabahi and Movaghar, 2008; Modi et al., 2012). 4 nhóm công nghệ: IDS dựa trên máy chủ (HIDS), IDS dựa trên mạng (NIDS), IDS dựa trên không dây (WIDS), phân tích hành vi mạng (NBA) và IDS hỗn hợp (MIDS). HIDS giám sát và thu thập các đặc điểm của máy chủ chứa các thông tin nhạy cảm, máy chủ chạy các dịch vụ công cộng và các hoạt động đáng ngờ. NIDS giám sát và thu thập lưu lượng mạng tại phân đoạn mạng cụ thể thông qua cảm biến, và sau đó phân tích hoạt động của các ứng dụng và giao thức để nhận ra sự cố đáng ngờ. WIDS tương tự như NIDS, nhưng nó bắt lưu lượng mạng không dây, như mạng ad hoc, cảm biến không dây, và mạng lưới không dây. Bên cạnh đó, hệ thống NBA kiểm tra lưu lượng mạng để nhận ra các cuộc tấn công với luồng dữ liệu bất ngờ. Việc áp dụng nhiều công nghệ như MIDS có thể đạt được mục tiêu phát hiện đầy đủ và chính xác hơn.

Các thành phần trong IDS gồm sensor (cảm biến) và agent, sensor thường được sử dụng cho các hệ thống NIDS, WIDS và NBA để giám sát các hoạt động mạng, và HIDS dùng agent để giám sát và phân tích các hoạt động. Cả sensor và agent đều có thể cung cấp dữ liệu đến Máy chủ quản lý (Management Server – MS) và Máy chủ cơ sở dữ liệu (Data Server – DS), MS là thiết bị tập trung để xử lý các sự cố đã ghi và DS là một kho lưu trữ thông tin, sự kiện. Hơn nữa, có 2 loại kiến trúc mạng. Một là mạng được quản lý (MN), một mạng riêng biệt được triển khai để quản lý phần mềm bảo mật nhằm che giấu thông tin IDS khỏi những kẻ xâm nhập. MN làm tăng thêm chi phí phần cứng và mang lại những bất lợi nhất định cho quản trị viên. Một mạng khác là Mạng tiêu chuẩn (SN), là một mạng công cộng không có bảo vệ. Cách để cải thiện tính bảo mật của SN là xây dựng một mạng ảo cô lập bằng cách định cấu hình mạng cục bộ ảo. Mặt khác, hầu hết các công nghệ IDS cung cấp bốn khả năng chung để giữ bảo mật, bao gồm thu thập thông tin, ghi nhật ký, phát hiện và ngăn chặn. Thu thập thông tin thu thập thông tin trên máy chủ/ mạng từ các hoạt động được quan sát. Ghi nhật ký, dữ liệu ghi liên quan cho các sự kiện được phát hiện, có thể được sử dụng để xác nhận các cảnh báo và các sự cố đã điều tra. Các phương pháp phát hiện trong hầu hết các IDS thường cần sự điều chỉnh phức tạp để nhận được độ chính xác cao hơn. Đối với vấn đề phòng chống, chúng tôi đề nghị người đọc tham khảo tài liệu khảo sát (Stavroulakis and Stamp, 2010).

Một nhược điểm chung của các công nghệ IDS là chúng không thể cung cấp khả năng phát hiện chính xác tuyệt đối. Dương tính giả (FP) và âm tính giả (FN) là hai chỉ số để đánh giá mức độ chính xác. Dương tính giả xảy ra khi IDS xác định không chính xác hoạt động lành tính là độc hại, còn âm tính giả xảy ra nếu IDS không xác định được hoạt động độc hại. Trong trường hợp chỉ có thể chọn 1 trong 1, nhiều quản trị viên bảo mật thích giảm FN hơn để tăng FP do tính bảo mật cao. Gần đây, Ho et al. (2012) thu thập các trường hợp FP và FN từ lưu lượng trong thế giới thực, phân tích thống kê các trường hợp này và đề xuất ba phát hiện. Đầu tiên, phần lớn các trường hợp sai là FN, bởi vì hầu hết các hành vi ứng dụng và định dạng nội dung của nó là tự xác định, không tuân theo các đặc tả RFC. Thứ hai, hầu hết các cảnh báo FP không liên quan đến các vấn đề an ninh, mà liên quan đến chính sách quản lý. Cuối cùng, có một tỷ lệ FN cực kỳ cao đối với các cuộc tấn công cũ, bao gồm tràn bộ đệm, các cuộc tấn công máy chủ SQL và các cuộc tấn công vào máy chủ.

Hơn nữa, chúng tôi tóm tắt và sàng lọc nhiều cuộc khảo sát trước đây (Debar et al., 1999, , 2000; Axelsson, 2000; Estevez-Tapiador et al., 2004; Amer and Hamilton, 2010; Bace and Mell, 2001; Sabahi and Movaghar, 2008; Lazarevic et al., 2005; Xenakis et al., 2011) để đưa ra 1 quan điểm mới về phân loại IDSs. Hình 1 giới thiệu bốn khía cạnh để phân loại IDS và phần sau mô tả ngắn gọn theo trình tự. Ở nhánh System Deployment, Kiến trúc mạng “tập trung” thu thập và phân tích thông tin từ một hệ thống được giám sát duy nhất. Kiến trúc mạng “phân tán” thu thập dữ liệu từ nhiều hệ thống được giám sát, để phát hiện toàn bộ, phân tán và các cuộc tấn công hợp tác hoặc “kết hợp” cả hai. Với tính năng hiện đại, cấu hình phân tán nên được song song hóa, dựa trên lưới hoặc dựa trên đám mây. Networking Type chỉ ra sự kết nối của IDS với hệ thống, được giảm sát thông qua “có dây”, “không dây” hoặc “hỗn hợp”. Đặc biệt, IDS không dây đạt được các yêu cầu bùng nổ, thiết lập trong môi trường độc lập, hợp tác hoặc phân cấp. Mục quan trọng nhất là Loại Công nghệ, đã được trình bày trong phần trước.

## Nội dung 02: Tìm hiểu về máy học và mạng sinh đối kháng

### Học máy

(Machine Learnning - ML) là kỹ thuật thiết kế và phát triển các thuật toán cho phép máy tính đánh giá hành vi dựa trên dữ liệu, chẳng hạn như dữ liệu cảm biến hoặc cơ sở dữ liệu. Trọng tâm chính của nghiên cứu học máy là tự động học cách nhận ra các mẫu phức tạp và đưa ra quyết định thông minh dựa trên dữ liệu. Học máy có thể được chia thành các nhánh như sau: học có giám sát, học không giám sát, học nửa giám sát và học củng cố.

#### Cây quyết định (Decision Tree)

Cây quyết định là cây mà mỗi nút biểu diễn một đặc trưng (tính chất), mỗi nhánh (branch) biểu diễn một quy luật (rule) và mỗi lá biểu biễn một kết quả (giá trị cụ thể hay một nhánh tiếp tục).

Có một vài thuật toán để tạo một cây quyết định, nhưng sẽ có 2 thuật toán nổi bật trong số đó:

1. CART (Classification and Regression Trees) → dùng Gini Index(Classification) để kiểm tra.

2. ID3 (Iterative Dichotomiser 3) → dùng Entropy function và Information gain để kiểm tra.

**Tiêu chuẩn dừng**

Trong các thuật toán Decision tree, ta sẽ chia mãi các node nếu nó chưa tinh khiết. Như vậy, ta sẽ thu được một tree mà mọi điểm trong tập huấn luyện đều được dự đoán đúng (giả sử rằng không có hai input giống nhau nào cho output khác nhau). Khi đó, cây có thể sẽ rất phức tạp (nhiều node) với nhiều leaf node chỉ có một vài điểm dữ liệu. Như vậy, nhiều khả năng overfitting sẽ xảy ra.

• Nếu node đó có entropy bằng 0, tức mọi điểm trong node đều thuộc một class.

• Nếu node đó có số phần tử nhỏ hơn một ngưỡng nào đó. Trong trường hợp này, ta chấp nhận có một số điểm bị phân lớp sai để tránh overfitting. Class cho leaf node này có thể được xác định dựa trên class chiếm đa số trong node.

• Nếu khoảng cách từ node đó đến root node đạt tới một giá trị nào đó. Việc hạn chế chiều sâu của tree này làm giảm độ phức tạp của tree và phần nào giúp tránh overfitting.

• Nếu tổng số leaf node vượt quá một ngưỡng nào đó.

• Nếu việc phân chia node đó không làm giảm entropy quá nhiều (information gain nhỏ hơn một ngưỡng nào đó).

Một số thuật toán khác

Ngoài ID3, C4.5, ta còn một số thuật toán khác như:

• Thuật toán CHAID: tạo cây quyết định bằng cách sử dụng thống kê chi-square để xác định các phân tách tối ưu. Các biến mục tiêu đầu vào có thể là số (liên tục) hoặc phân loại.

• Thuật toán C&R: sử dụng phân vùng đệ quy để chia cây. Tham biến mục tiêu có thể dạng số hoặc phân loại.

• MARS

• Conditional Inference Trees

**Ưu/nhược điểm của thuật toán cây quyết định**

**Ưu điểm**

Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó:

• Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.

• Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả.

• Có thể làm việc với cả dữ liệu số và dữ liệu phân loại.

• Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê.

• Có khả năng là việc với dữ liệu lớn.

**Nhược điểm**

Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

• Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.

• Cây quyết định hay gặp vấn đề overfitting.

#### Logistic Regression

Phương pháp hồi quy logistic là một mô hình hồi quy nhằm dự đoán giá trị đầu ra *rời rạc* (*discrete target variable*) *y* ứng với một véc-tơ đầu vào **x**. Việc này tương đương với chuyện phân loại các đầu vào **x** vào các nhóm *y* tương ứng.

Ví dụ, xem một bức ảnh có chứa một con mèo hay không. Thì ở đây ta coi đầu ra *y*=1 nếu bước ảnh có một con mèo và *y*=0 nếu bức ảnh không có con mèo nào. Đầu vào **x** ở đây sẽ là các pixel một bức ảnh đầu vào.

Chart, scatter chart

Description automatically generated

Mô hình Logistic Regression:

Sử dụng phương pháp thống kê ta có thể coi rằng khả năng một đầu vào x nằm vào một nhóm *y*0​ là xác suất nhóm *y*0​ khi biết x: *p*(*y*0​∣x). Dựa vào công thức xác xuất hậu nghiệm ta có:

Ảnh có chứa văn bản

Mô tả được tạo tự động

Đặt :

Ảnh có chứa văn bản

Mô tả được tạo tự động

Ta có:

Chart, box and whisker chart

Description automatically generated

Hàm *σ*(*a*) ở đây được gọi là **hàm sigmoid** (logistic sigmoid function). Hình dạng chữ S bị chặn 2 đầu của nó rất đặt biệt ở chỗ dạng phân phối đều ra và rất mượt.

*Ảnh có chứa văn bản, thiết bị điện tử

Mô tả được tạo tự động*

Công thức tính xác suất lúc này:

Diagram

Description automatically generated

Trong đó,  x là thuộc tính đầu vào còn w là trọng số tương ứng.

#### Random Forest

Random Forest là một thuật toán học có giám sát. Random Forest (Rừng ngẫu nhiên) sử dụng các cây (tree) để làm nền tảng.Nó là một tập hợp của các Decision Tree, mà mỗi cây được chọn theo một thuật toán dựa vào ngẫu nhiên.

Điểm mạnh của Random Forest :

• Random Forest có thể sử dụng cho cả bài toán Classification và Regression

• Random Forest làm việc được với dữ liệu thiếu giá trị

• Khi Forest có nhiều cây hơn, chúng ta có thể tránh được việc Overfitting với tập dữ liệu

• Có thể tạo mô hình cho các giá trị phân loại

Cách hoạt động của Random Forest :

Random Forest hoạt động bằng cách đánh giá nhiều Cây quyết định ngẫu nhiên, và lấy ra kết quả được đánh giá tốt nhất trong số kết quả trả về.

Để biểu diễn dự đoán sử dụng Random Forest đã huấn luyện, ta sử dụng các bước bên dưới :

1. Lấy các test features và sử dụng các Cây quyết định đã tạo ra để dự đoán kết quả, lưu nó vào một danh sách.

2. Tính toán số lượng vote trên toàn bộ Forest cho từng kết quả

3. Lấy kết quả có số lượng vote lớn nhất làm kết quả cuối cho mô hình

#### SupportVectorMachine (SVM)

**SVM** là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, vẽ đồ thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

Chart, scatter chart

Description automatically generated

Support Vectors hiểu một cách đơn giản là các đối tượng trên đồ thị tọa độ quan sát, Support Vector Machine là một biên giới để chia hai lớp tốt nhất.

Là một kĩ thuật phân lớp khá phổ biến, SVM thể hiện được nhiều ưu điểm trong số đó có việc tính toán hiệu quả trên các tập dữ liệu lớn.

**Ưu điểm** :

* Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.
* Tiết kiệm bộ nhớ: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.
* Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.

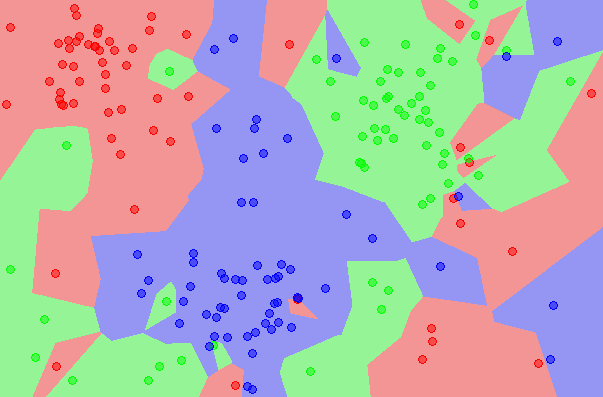
**Nhược điểm**:

* Bài toán số chiều cao: Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả khá tồi.
* Chưa thể hiện rõ tính xác suất: Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm margin từ điểm dữ liệu mới đến siêu phẳng phân lớp mà chúng ta đã bàn luận ở trên.

#### K-nearest Neighbor

KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách chỉ dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận), không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu.

Với KNN, trong bài toán Classification, label của một điểm dữ liệu mới (hay kết quả của câu hỏi trong bài thi) được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set. Label của một test data có thể được quyết định bằng major voting (bầu chọn theo số phiếu) giữa các điểm gần nhất, hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label.



**Ưu điểm:**

* Độ phức tạp tính toán của quá trình training là bằng 0.
* Việc dự đoán kết quả của dữ liệu mới rất đơn giản.
* Không cần giả sử gì về phân phối của các class.

**Nhược điểm :**

* KNN rất nhạy cảm với nhiễu khi K nhỏ.
* KNN là một thuật toán mà mọi tính toán đều nằm ở khâu test. Trong đó việc tính khoảng cách tới từng điểm dữ liệu trong training set sẽ tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.

### Học sâu

Mô phỏng cách bộ não hoạt động, các thuật toán học sâu được dựa trên các neuron nhân tạo được nối với nhau thành các lớp.

Một neuron nhân tạo (còn được gọi là percepton) là một hàm biến đổi toán học nhận một hoặc nhiều đầu vào đã được nhân với các giá trị gọi là "weights", cộng các giá trị đó lại với nhau thành một giá trị duy nhất. Sau đó giá trị này được đưa vào một hàm phi tuyến (được gọi là activation function) và kết quả của hàm này chính là đầu ra của neuron.

A picture containing watch

Description automatically generated

Ta có:

* x biểu diễn các giá trị đầu vào của neuron hiện tại
* w là các giá trị weights có tác dụng biến đổi các giá trị đầu vào thành một định dạng có cấu trúc mà máy có thể hiểu được
* b là giá trị bias có tác dụng tăng sự linh hoạt trong việc học dữ liệu
* z là giá trị trước khi đi vào hàm phi tuyến, với:
* a là giá trị cuối cùng và là đầu ra của neuron hiện tại, sẽ được truyền đến tất cả neuron được nối với neuron hiện tại của hidden layer kế tiếp hoặc được xử lý như là kết quả cuối cùng của mô hình.

Giá trị a được tính với hàm a=f(x), với các hàm f như sau:

* Linear Activation Function:
* Sigmoid Activation Function:
* Hyperbolic Tangent Activation Function:
* Rectified Linear Unit Activation Function: :
* Leaky ReLU Activation Function: :
* Softmax Activation Function: :

#### Multi-layer Perceptron

Mô hình mạng nơron được sử dụng rộng rãi nhất là mô hình mạng nhiều tầng truyền thẳng (MLP: Multi Layer Perceptron). Một mạng MLP tổng quát là mạng có n (n≥2) tầng (thông thường tầng đầu vào không được tính đến): trong đó gồm một tầng đầu ra (tầng thứ n) và (n-1) tầng ẩn.

Diagram

Description automatically generated with medium confidence

**Kiến trúc của một mạng MLP tổng quát có thể mô tả như sau:**

* Đầu vào là các vector (x1, x2, …, xp) trong không gian p chiều, đầu ra là các vector (y1, y2, …, yq) trong không gian q chiều. Đối với các bài toán phân loại, p chính là kích thước của mẫu đầu vào, q chính là số lớp cần phân loại.
* Mỗi neural thuộc tầng sau liên kết với tất cả các nơron thuộc tầng liền trước nó.
* Đầu ra của neural tầng trước là đầu vào của nơron thuộc tầng liền sau nó.

Hoạt động của mạng MLP như sau: tại tầng đầu vào các neural nhận tín hiệu vào xử lý (tính tổng trọng số, gửi tới hàm truyền) rồi cho ra kết quả (là kết quả của hàm truyền); kết quả này sẽ được truyền tới các neural thuộc tầng ẩn thứ nhất; các nơron tại đây tiếp nhận như là tín hiệu đầu vào, xử lý và gửi kết quả đến tầng ẩn thứ 2. Quá trình tiếp tục cho đến khi các neural thuộc tầng ra cho kết quả.

**Một số kết quả đã được chứng minh:**

* Bất kì một hàm Boolean nào cũng có thể biểu diễn được bởi một mạng MLP 2 tầng trong đó các neural sử dụng hàm truyền sigmoid.
* Tất cả các hàm liên tục đều có thể xấp xỉ bởi một mạng MLP 2 tầng sử dụng hàm truyền sigmoid cho các neural tầng ẩn và hàm truyền tuyến tính cho các nơron tầng ra với sai số nhỏ tùy ý.
* Mọi hàm bất kỳ đều có thể xấp xỉ bởi một mạng MLP 3 tầng sử dụng hàm truyền sigmoid cho các neural tầng ẩn và hàm truyền tuyến tính cho các neural tầng ra.

#### Convolutional Neural Network:

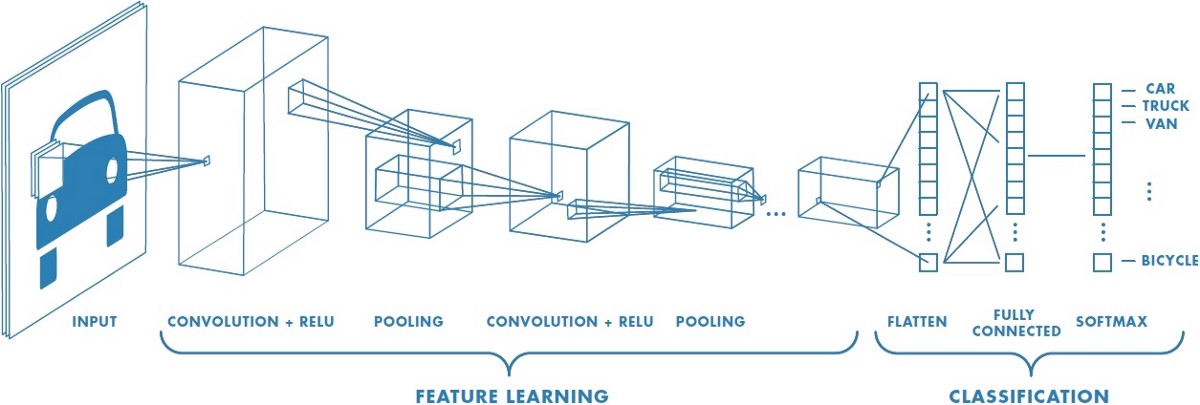
Gồm 4 layer:

Convolutional layer: Đây là lớp quan trọng nhất của CNN, lớp này có nhiệm vụ thực hiện mọi tính toán. Những yếu tố quan trọng của một convolutional layer là: stride, padding, filter map, feature map.

Relu layer: Relu layer là hàm kích hoạt trong neural network và hàm này còn được gọi là activation function. Hàm kích hoạt có tác dụng mô phỏng các neuron có tỷ lệ truyền xung qua axon

Pooling layer: Khi đầu vào quá lớn, những lớp pooling layer sẽ được xếp vào giữa giữa những lớp Convolutional layer để làm giảm parameter. Hiện nay, pooling layer có 2 loại chủ yếu là: max pooling và average.

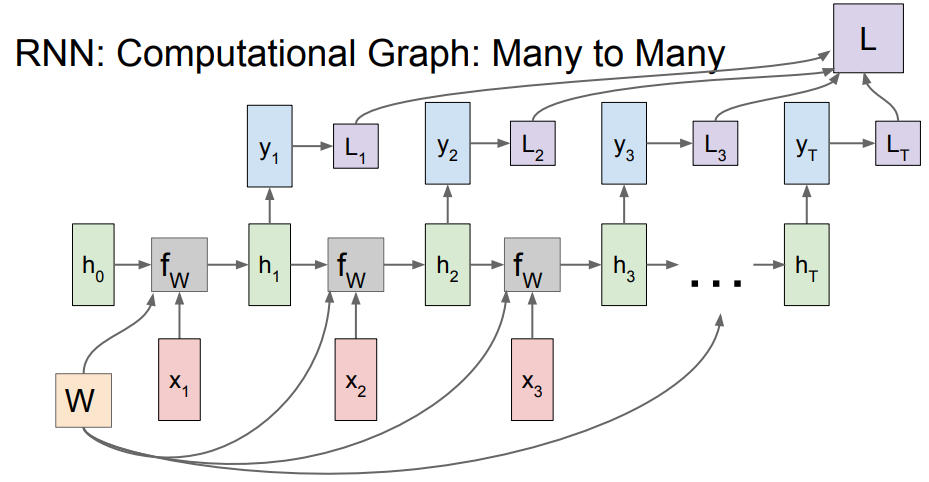
Fully connected layer: Lớp này có nhiệm vụ đưa ra kết quả sau khi lớp convolutional layer và pooling layer đã nhận được data truyền. Lúc này, ta thu được kết quả là model đã đọc được thông tin của data và để liên kết chúng cũng như cho ra nhiều output hơn thì ta sử dụng fully connected layer.



#### Recurrent Neural Network

Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau. Tức là chúng không liên kết thành chuỗi với nhau. Nhưng các mô hình này không phù hợp trong rất nhiều bài toán như hoàn thành một câu, từ những từ ở phía trước.. vì những dự đoán tiếp theo như từ tiếp theo phụ thuộc vào vị trí của nó trong câu và những từ đằng trước nó.

Và như vậy RNN ra đời với ý tưởng chính là sử dụng một bộ nhớ để lưu lại thông tin từ từ những bước tính toán xử lý trước để dựa vào nó có thể đưa ra dự đoán chính xác nhất cho bước dự đoán hiện tại. Hiện nay RNN được ứng dụng nhiều trong NLP, Dịch máy...



### Mạng sinh đối kháng (GAN)

Mạng sinh đối kháng (GAN) là một mô hình machine learning được Goodfellow và cộng sự giới thiệu trong bài báo khoa học có tên “Generative Adversarial Networks.”. Mô hình này bao gồm hai mạng nơ-ron :

Mô hình sinh mẫu G (Generative model): dùng để nắm bắt phân phối dữ liệu,được xem tương tự như nhóm người cố gắng sản xuất tiền giả và sử dụng nó mà không bị phát hiện.

Mô hình phân biệt D (Discriminative model): dùng để ước tính xác suất của mẫu từ dữ liệu huấn luyện chứ không phải từ G, được xem tương tự như là cảnh sát cố gắng phát hiện tiền giả.

Ảnh sau minh họa luồng xử lý của hai mạng :

Diagram

Description automatically generated

Ta có hàm loss của GAN như sau:



Trong đó:

• 𝑥 : là mẫu thật

• 𝐺(𝑧) : là mẫu được sinh ra

• 𝑧 : là nhiễu ngẫu nhiên

• 𝐸𝑥[𝑓] : kỳ vọng của hàm 𝑓 theo 𝑥.

Như đã đề cập, mục tiêu của bộ phân biệt D là tối đa hoá dự đoán chính xác mẫu thật, tức cố gắng sao cho tỉ lệ của mẫu sinh ra và mẫu thật khác nhau càng lớn càng tốt → **max**. Do đó, D cần được huấn luyện với tăng gradient. Quy tắc cập nhật của D sẽ như sau:

Diagram

Description automatically generated with medium confidence

Tương tự với bộ phân biệt, mục tiêu của bộ sinh mẫu G là đánh lừa bộ phân biệt D và khiến D tin mẫu được tạo ra là thật. Do đó, G cần được huấn luyện để giảm thiểu 1 − 𝐷(𝐺(𝑧)), tức cố gắng sao cho mẫu sinh ra và mẫu thật khác nhau càng ít càng tốt → min. Vì vậy, G cần được huấn luyện với giảm gradient. Quy tắc cập nhật của G sẽ như sau:

A picture containing text, watch

Description automatically generated

Áp dụng cả 3 phương trình, ta có quá trình huấn luyện GAN như sau:

Diagram

Description automatically generated

Mặc khác,WGAN do Martin Arjovsky và cộng sự nghiên cứu không sử dụng hàm loss minmax mà sử dụng cách tiếp cận mới: Thay bộ phân biệt (discrimator) bằng một bộ chấm điểm, với mỗi input vào, bộ chấm điểm sẽ ra một con số, được tính bằng hàm loss, dựa trên khoảng cách Wasserstein :

Ta có công thức tính hàm loss:



Bởi vì đầu ra là một con số có thể bé hơn 0 hoặc lớn hơn 1, ta không thể, và không cần, lấy 0.5 làm mốc xác định mẫu thật hay giả. Ưu điểm của WGAN so với GAN là chúng tránh được vấn đề “vanishing gradient”.

## Nội dung 03: Tìm hiểu về các dataset dùng để training

Dataset là một phần thiết yếu trong quá trình huấn luyện GAN. Thông thường, các dataset được tạo ra bằng cách ghi lại toàn bộ lưu lượng mạng trong một môi trường sử dụng. Các lưu lượng mạng này bao gồm kết nối hợp lệ và các kết nối tấn công, với tỉ lệ tấn công và kĩ thuật khác nhau. Sau đó những bản ghi này thường được “xử lý” lại chỉ còn những đặc tính (feature) cần thiết để giảm dung lượng và tiện lợi cho quá trình tiền xử lý.

### DARPA (Lincoln Laboratory 1998-99)

Tập dữ liệu được xây dựng để phân tích an ninh mạng và đã đặt ra các vấn đề liên quan đến việc đến việc đưa các cuộc tấn công giả tạo và lưu lượng truy cập. Bộ dữ liệu này bao gồm các mối quan hệ hoạt động qua e-mail, duyệt web, FTP, Telnet, IRC và SNMP. Nó chứa các cuộc tấn công như DoS, Đoán mật khẩu, tràn bộ đệm, FTP từ xa, lũ Syn, Nmap và Rootkit. Tập dữ liệu này không đại diện cho lưu lượng mạng trong thực tế và chứa các điểm bất thường như không có kết quả dương tính giả. Ngoài ra, bộ dữ liệu đã lỗi thời để đánh giá hiệu quả của các IDS trên các mạng hiện đại, cả về kiểu tấn công và cơ sở hạ tầng mạng. Hơn nữa, nó thiếu các bản ghi dữ liệu tấn công thực tế.

### KDD’99 (University of California, Irvine 1998-99):

Tập dữ liệu này là phiên bản cập nhật của DARPA98, bằng cách xử lý phần tcpdump. Nó chứa các cuộc tấn công: Neptune -DoS, pod-DoS, Smurf-DoS, and buffer-overflow. Lưu lượng truy cập và tấn công được hợp nhất để đi vào một môi trường giả lập. Tập dữ liệu này có một số lượng lớn các bản ghi dư thừa và bị đính kèm bởi các lỗi dữ liệu dẫn đến kết quả thử nghiệm bị sai lệch. NSL-KDD được tạo bằng KDD (Tavallaee et al., 2009) để giải quyết một số thiếu sót của KDD

### DEFCON (The Shmoo Group, 2000-2002):

Tập dữ liệu DEFCON-8 được tạo vào năm 2000 chứa các cuộc tấn công quét cổng và tràn bộ đệm, trong khi tập dữ liệu DEFCON-10, được tạo vào năm 2002, chứa quét và quét cổng, gói tin xấu, đặc quyền quản trị viên và các cuộc tấn công FTP bằng giao thức Telnet. Trong tập dữ liệu này, lưu lượng được tạo ra trong cuộc thi “Cap ture the Flag (CTF)” khác với lưu lượng mạng trong thế giới thực vì nó chủ yếu bao gồm lưu lượng truy cập xâm nhập trái ngược với lưu lượng truy cập nền bình thường. Bộ dữ liệu này được sử dụng để đánh giá các kỹ thuật điều chỉnh cảnh báo (Nehinbe, 2010) (Group, 2000).

### Dataset NSL-KDD

Được Tavallaee và cộng sự công bố năm 2009, Dataset này bắt nguồn từ KD99, mỗi bản ghi có 41 trường thông tin Dataset gồm: “Intrinsic”: Bản chất của kết nối, “Content”:Nội dung của kết nối có thể liên quan đến cuộc tấn công, “Time-based traffic”: Theo dõi những kết nối trong vòng 2 giây trước có cùng đích đến hoặc dịch vụ mạng, “Host-based traffic”: Theo dõi 100 kết nối gần nhất có cùng đích đến hoặc dịch vụ mạng

Graphical user interface, text, application

Description automatically generated

Bộ huấn luyện/kiểm thử đầy đủ:(KDDTrain+ có hơn 125.000 bản ghi, KDDTest+ hơn 22.000 bản ghi)

Bộ huấn luyện rút gọn KDDTrain+\_20Percent : Trích 20% từ bộ huấn luyện đầy đủ

Bộ kiểm thử KDDTest-21: Trích những traffic khó nhất từ bộ kiểm thử đầy đủ

Ngoài ra, giá trị của các trường dữ liệu có thể được tạm chia làm 3 phần:

* Các giá trị không phải số: Là các trường dữ liệu như protocol ( giao thức kết nối: tcp, udp, icmp), service ( dịch vụ đang nghe trên cổng: pop, http,...), label ( nhãn của kết nối, xác định là kết nối hợp hệ, hay là loại tấn công độc hại: normal, dos, ..)
* Các giá trị là số: Là các trường dữ liệu có giá trị là số nguyên hoặc số thực.
* Các giá trị nhị phân: Là các trường dữ liệu có giá trị là 0 hoặc 1

Các loại tấn công mạng: R2L (Remote to Local), U2R (User to Root), DoS (Denial of Service), Probe.

Hạn chế: Không thể hiện được vết của các cuộc tấn công ở mức độ thấp, tinh vi.

### Dataset CICIDS2017

Được công bố năm 2017 bởi Viện an ninh mạng Canada (CIC), gồm các loại tấn công mạng: DoS, DdoS, Bruteforce FTP, Bruteforce SSH, Bruteforce, Portscan, Botnet, SQL injection, XSS, Heartbleed, các loại xâm nhập khác..

Hạn chế: Mặc dù đây là một trong những dataset mới nhất, nhưng nó vẫn mắc vài hạn chế như: Dữ liệu quá lớn, nhiều (288602) bản ghi thiếu thông tin, phân bố tấn công không đều…

## Nội dung 04: Nghiên cứu phân tích, đánh giá mô hình IDSGAN

### Cấu trúc IDSGAN:

Trong IDSGAN, trình tạo (generator) sửa đổi một số đặc tính cụ thể để tạo các mẫu lưu lượng tấn công đối nghịch. Trình phân biệt (disciminator) được đào tạo để bắt chước IDS hộp đen và hỗ trợ đào tạo generator. IDS hộp đen được triển khai bằng các thuật toán học máy để phát hiện các cuộc tấn công. Bằng cách làm cho các trọng số của generator khác với IDS trong đào tạo, các mẫu đối nghịch có thể được tạo để trốn tránh việc phát hiện IDS. Mô hình IDSGAN được mô tả như sau:

Diagram

Description automatically generated

Mặc dù mục đích chính của việc tạo các mẫu tấn công đối nghịch là để trốn tránh IDS, tiền đề là mẫu đối nghịch được tạo ra nên giữ lại chức năng tấn công của lưu lượng.

Mỗi loại tấn công khác nhau đều có những đặc tính đại diện (functional feature) cho cuộc tấn công. Những thuộc tính này nên được giữ lại nguyên bản. Ngược lại, với những đặc tính khác ta có thể giữ lại hoặc thay thế. Trong dataset NSL-KDD, những functional feature của từng loại tấn công được trình bày như trong bảng sau:

Table

Description automatically generated

#### Trình tạo (generator):

Là một phần quan trọng của mô hình, trình tạo (generator) đóng vai trò tạo ra mẫu lưu lượng tấn công đối với cuộc tấn công qua mặt IDS.

Để chuyển đổi một mẫu gốc thành một mẫu đối nghịch, mỗi vectơ đầu vào của các mẫu lưu lượng phải bao gồm một vectơ mẫu gốc M m-chiều và vectơ nhiễu N n-chiều. Là phần mẫu gốc, M đã được xử lý trước. Để phù hợp với vectơ được xử lý trước, các phần tử của N phần nhiễu bao gồm các số ngẫu nhiên trong một phân bố đồng đều trong phạm vi [0,1].

Cấu trúc đề xuất của generator có một mạng thần kinh (neural network) với 5 lớp tuyến tính. Lớp ReLU F = max (0; x) được sử dụng để kích hoạt đầu ra của 4 lớp tuyến tính trước. Để làm cho các mẫu đối nghịch đáp ứng công thức của vectơ mẫu gốc M, lớp đầu ra phải có m đơn vị. Việc cập nhật các tham số trong mạng này dựa trên phản hồi từ disciminator.

#### Trình phân biệt (discriminator):

Trình phân biệt là một mạng lưới thần kinh nhiều lớp để phân loại lưu lượng tấn công và lưu lượng bình thường. Dữ liệu huấn luyện của nó bao gồm lưu lượng tấn công đối nghịch và lưu lượng bình thường.

Là một trong những mục đích chính của discriminator, công việc bắt chước khả năng của IDS hộp đen cần kết quả phát hiện của IDS hộp đen đối với các mẫu lưu lượng tấn công bình thường và đối nghịch. Đầu tiên, các mẫu lưu lượng thường và các mẫu lưu lượng tấn công đối nghịch được phân loại theo IDS hộp đen. Sau đó, kết quả phát hiện của IDS hộp đen được sử dụng làm nhãn mục tiêu của dữ liệu huấn luyện discriminator để làm cho phân loại của discriminator, tương tự như IDS hộp đen.

Diagram

Description automatically generated

#### Thuật toán huấn luyện

Hàm mất mát của generator như sau:



trong đó Sattack là mẫu lưu lượng tấn công ban đầu; G đại diện cho generator và D đại diện cho discriminator . Để đào tạo và tối ưu hóa generator để đánh lừa IDS hộp đen, chúng ta cần giảm thiểu LG.

Hàm mất mát của discrimator như sau:



trong đó s là các mẫu lưu lượng cho việc đào tạo discriminator; Bnormal và Battack tương ứng là các mẫu bình thường và các mẫu tấn công đối nghịch được dự đoán bởi IDS hộp đen.

Thuật toán mẫu IDSGAN như sau:

Text

Description automatically generated

## Nội dung 05: Thực hiện hóa mô hình IDSGAN trên máy tính, thử nghiệm tấn công các thuật toán máy học dùng làm IDS

Nhóm lần lượt thực hiện các bước: Tiền xử lý dữ liệu, huấn luyện IDS black-box, huấn luyện GAN, chạy mô hình IDSGAN trên các mô hình đã được huấn luyện.

Có một số thuật toán nhóm đã chạy huấn luyện IDS black-box thành công / hoặc huấn luyện GAN thành công, nhưng không thể chạy mô hình IDSGAN. Tuy nhiên, nhóm vẫn thống kê trên báo cáo cho mục đích tham khảo.

### Thực hiện tiền xử lý dữ liệu

Để phục vụ việc huấn luyện các mô hình, dataset NSL-KDD cần được tiền xử lý một cách phù hợp.

Dữ liệu được tiền xử lý thông qua các bước như sau:



#### Xử lý thô:

Trong dataset, có những trường dữ liệu không phải là số , hoặc chỉ là giá trị nhị phân (0 hoặc 1), hoặc là trường dữ liệu không cần dùng đến. Những trường dữ liệu này sẽ được xử lý.

* Đối với những trường dữ liệu không phải số: gồm “protocol type” (biểu thị giao thức tầng Transport (giao vận) đang được sử dụng), “service” (gồm thông tin về dịch vụ dang kết nối đến), “flag” ( biểu thị trạng thái của kết nối, và/ hoặc trạng thái của cờ TCP). Các giá trị của những trường dữ liệu được mô tả như trong phần 3 “Nội dung 03: Tìm hiểu về các dataset dùng để training".

Nhóm đề xuất phương pháp xử lý các trường dữ liệu không phải là số bằng cách thay thế tên gọi của chúng bằng thứ tự của chúng trong mảng, như minh họa dưới đây:

Text, letter

Description automatically generated

Sau khi xử lý:

Text

Description automatically generated

* Đối với trường dữ liệu là nhị phân: Nhóm đề xuất chuyển thành object.
* Đối với trường dữ liệu không cần thiết (trường “score”): Nhóm đề xuất xóa.

#### Normalize data (Chuẩn hóa dữ liệu)

Với các giá trị trong tập số nguyên và số thực, nhóm đề xuất chuẩn hóa sử dụng hàm min – max như sau:

Với:

* là giá trị trước khi chuẩn hóa
* là giá trị sau khi chuẩn hóa
* là giá trị lớn nhất của trường dữ liệu
* là giá trị nhỏ nhất của trường dữ liệu

Sau khi chuẩn hóa, ta sẽ có các giá trị nằm trong khoảng [0..1]

#### Separate (Phân loại dữ liệu)

Nhóm sẽ tiến hành phân loại dữ liệu:

Đối với bộ train (KDDTrain+): Sử dụng các dữ liệu được phân loại là “normal”, “dos”, “u2r”, “r2l”.

Đối với bộ test (KDDTest+): Sử dụng các dữ liệu độc hại để phục vụ cho GAN: “dos”, “u2r&r2l”.

Mỗi loại dữ liệu được tách thành một file riêng.

#### Split (Tách dữ liệu)

Để đảm bảo tính chính xác, nhóm tách bộ dataset KDD Train+ đã phân loại ở trên làm hai phần bằng nhau, một phần dùng cho training black-box IDS và một phần dùng để training các phần của GAN.

### Thực hiện huấn luyện Black-box IDS

Các thuật toán được nhóm hiện thực hóa:

Machine Learning:

* + Decision Tree,
  + KNN,
  + Logistic Regesion,
  + Random Forest,
  + SVM,

Deep Learning:

* + MLPs,
  + CNNs,
  + RNNs.

Sau đây là bảng giá trị Detection Rate (tỉ lệ phát hiện) và Accuracy ( độ chính xác) của từng thuật toán IDS:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Loại thuật toán | | DOS | | U2R&R2L | |
| Tên thuật toán | DL | ML | Accuracy% | Detection Rate% | Accuracy% | Detection Rate% |
| Decision Tree |  | X | 90.11 | 81.51 | 84.32 | 35.87 |
| KNN |  | X | 90.46 | 79.64 | 77.51 | 3.83 |
| Logistic Regesion |  | X | 89.83 | 79.40 | 76.74 | 0.44 |
| Random Forest |  | X | 90.40 | 79.14 | 79.03 | 10.13 |
| SVM |  | X | 90.52 | 79.25 | 76.67 | 0.20 |
| MLPs | X |  | 87.05 | 80.83 | 74.37 | 0.04 |
| CNNs | X |  | 88.03 | 73.57 | 76.41 | 0.25 |
| RNNs | X |  | 87.48 | 72.09 | 76.87 | 1.15 |

### Thực hiện huấn luyện GAN

Các thuật toán được nhóm hiện thực hóa:

Machine Learning:

* + Decision Tree,
  + KNN,
  + Logistic Regesion,
  + Random Forest,
  + SVM,

Deep Learning:

* + MLPs,

Thuật toán huấn luyện GAN được miêu tả như hình sau:

Text

Description automatically generated with low confidence

Sau khi huấn luyện GAN, nhóm xin trình bày Origin DR ( ODR, tỉ lệ phát hiện ban đầu), Adversiral DR (ADR, tỉ lệ phát hiện với dữ liệu đối kháng), EIR ( evasion increase rate, tỉ lệ né tránh tăng thêm), với:

, và

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Loại thuật toán | | DOS | | | U2R&R2L | | |
| Tên thuật toán | DL | ML | ODR | ADR | EIR | ODR | ADR | EIR |
| Decision Tree |  | X | 81.68 | 18.45 | 77.41 | 36.13 | 16.25 | 55.03 |
| KNN |  | X | 79.63 | 21.60 | 72.88 | 3.78 | 0 | 100 |
| Logistic Regesion |  | X | 79.39 | 16.07 | 79.76 | 0.4 | 0 | 100 |
| Random Forest |  | X | 79.06 | 0.63 | 99.20 | 10.25 | 0.01 | 99.90 |
| SVM |  | X | 79.17 | 0.58 | 99.27 | 0.19 | 0 | 100 |
| MLPs | X |  | 82.23 | 31.00 | 62.31 | 0.50 | 0 | 100 |
| CNNs | X |  | n/a | | | | | |
| RNNs | X |  | n/a | | | | | |

### Thực hiện tạo mẫu đối kháng từ GAN và đối chiếu với dataset

Ở bước này, IDSGAN sẽ lấy dữ liệu từ dataset và so sánh với dữ liệu đối kháng và lưu vào file csv để phục vụ mục đích sử dụng sau này.

# Kết luận và hướng phát triển

Với mục đích tạo ra các cuộc tấn công đối kháng để trốn tránh hệ thống phát hiện xâm nhập, IDSGAN là một khuôn khổ mới của các mạng đối kháng tạo sinh (GAN) dựa trên Wasserstein GAN, bao gồm trình tạo (Generator), trình phân biệt (Discriminator), IDS hộp đen. IDSGAN cho thấy khả năng tốt của mình trong việc tạo ra các mẫu lưu lượng tấn công của các tấn công khác nhau, khiến tỷ lệ phát hiện của các mô hình IDS hộp đen khác nhau giảm xuống xấp xỉ 0. Hơn nữa, để đánh giá mức độ ảnh hưởng của việc thay đổi số lượng các đặc tính chưa được đổi, nó cho thấy rằng khả năng trốn tránh của các mẫu lưu lượng tấn công đối nghịch sẽ duy trì hoặc giảm nhẹ khi thêm các đặc tính không đổi (unmodified features), phản ánh sự mạnh mẽ mạnh mẽ của IDSGAN. Mức độ thay đổi hiệu suất của IDSGAN sau khi bổ sung các đặc tính không đổi dựa trên loại tấn công và loại thuật toán phát hiện của IDS hộp đen. Hiệu năng tuyệt vời được trình bày ở trên bởi mô hình cho thấy tính khả thi và tính linh hoạt rộng rãi của IDSGAN.

# Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| Thành viên | Nội dung | Tự đánh giá |
| Phạm Nguyễn Việt Tân | Tìm hiểu về các thuật toán ML,DL | 10/10 |
| Phạm Ngọc Thành | Tìm hiểu về GAN, IDSGAN, thực hiện hóa mô hình | 10/10 |
| Nguyễn Khải Đăng | Tìm hiểu về các dataset | 10/10 |
| Phạm Bảo Hà | Tìm hiểu về IDS | 10/10 |

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, "A Detailed Analysi`s of the KDD CUP 99 Data Set," in Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009. |
| [2] | I. Sharafaldin, A. Habibi Lashkari and A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, 2018. |
| [3] | K. Scarfone and P. Mell, Guide to Intrusion Detection and Prevention Systems (IDPS), 2007. |
| [4] | S. Panigrahi, "A detailed analysis of CICIDS2017 dataset for designingIntrusion Detection Systems," International Journal of Engineering & Technology, 2018. |
| [5] | Z. Lin, Y. Shi and Z. Xue, "IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection," September 2018. |
| [6] | T.-T.-H. Le, Y. Kim and H. Kim, "Network Intrusion Detection Based on Novel Feature Selection Model and Various Recurrent Neural Networks," Applied Sciences, vol. 9, p. 2019. |
| [7] | S. Hitawala, Dhanabal, L and D. Shantharajah, "A Study on NSL-KDD Dataset for IntrusionDetection System Based on ClassificationAlgorithms," nternational Journal of Advanced Research in Computer and Communication Engineering, vol. 12, p. 1, 2015. |
| [8] | I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," June 2014. |
| [9] | S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT," Procedia Computer Science, vol. 167, p. 1561–1573, 2020. |
| [10] | M. Arjovsky, S. Chintala and L. Bottou, "Wasserstein GAN," January 2017. |