ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

ĐỀ CƯƠNG ĐỀ TÀI KHÓA LUẬN TỐT NGHIỆP

1. Tên đề tài hoặc hướng NC:

Tên tiếng Việt: **Phát sinh dữ liệu tấn công chống lại IDS bằng mô hình đối kháng tạo sinh**

Tên tiếng Anh: **Generative Adversarial Networks for Attack against Intrusion Detection System.**

2. Ngành và mã ngành đào tạo: An toàn thông tin **Mã ngành:** 7480202

3. Họ tên học viên thực hiện đề tài, khóa-đợt học:

Sinh viên: LÊ KHẮC TIẾN **MSSV:** 16521221 **Khóa:** 11

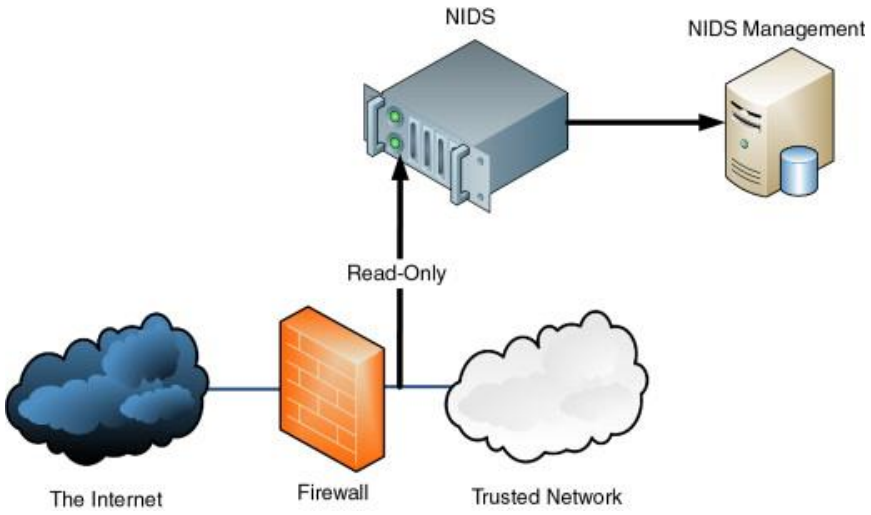
Người hướng dẫn: TS. Phạm Văn Hậu

Địa chỉ email, điện thoại liên lạc của người hướng dẫn:

Email: haupv@uit.edu.vn

4. Tổng quan tình hình NC:

Với sự lan rộng của các mối đe dọa bảo mật trên internet, hệ thống phát hiện xâm nhập (IDS) trở thành công cụ thiết yếu để phát hiện và phòng tránh tấn công mạng được thể hiện dưới dạng lưu lượng độc hại. IDS giám sát lưu lượng mạng và đưa ra cảnh báo nếu lưu lượng không an toàn được xác định, phát hiện bởi bộ phân tích. Mục đích chính của IDS 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.

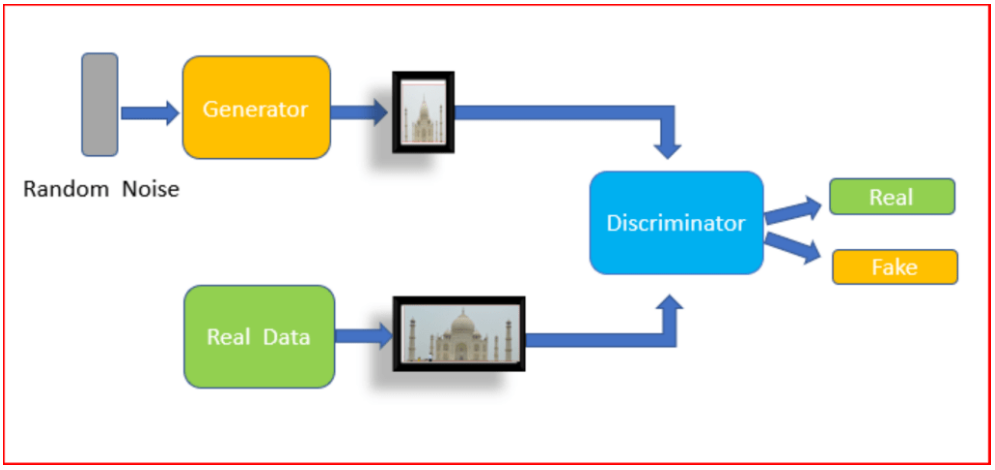


Hình 1: IDS trong hệ thống mạng

Đối với các vấn đề phân loại, thuật toán học máy đã được áp dụng rộng rãi trong IDS và đạt được kết quả khả quan Các thuật toán phát hiện đã được sử dụng để giám sát và phân tích lưu lượng độc hại, bao gồm K-Nearest Neighbor, Support Vector Machine (SVM), Cây quyết định (Decision Tree), v.v [1]. 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.

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 [4]. Kẻ tấn công cố gắng đánh lừa để các mô hình phân loại sai mong muốn bằng cách sử dụng các mẫu lưu lượng độc đối kháng. Và các Mạng Đối kháng Tạo sinh GAN) là phương pháp được lựa chọn tiềm năng cho các cuộc tấn công đối kháng như vậy.



Hình 2 Nguyên tắc hoạt động chính của GAN

Goodfellow và cộng sự đã giới thiệu GAN, một khung để đào tạo các mô hình sinh dữ liệu (generative models) [5], với ý tưởng chính là hai mạng nơ-ron nhiều lớp Generator và Discriminator, chơi một trò chơi minimax để đưa đến một giải pháp tối ưu [6]. Bộ sinh (Generator) tìm hiểu cách thức dữ liệu được hình thành, sau đó đặt câu hỏi: Dựa vào giả định cách thức tạo ra dữ liệu, dữ liệu sẽ được phân vào nhóm nào. Bộ phân biệt (Discriminator) không quan tâm đến cách thức dữ liệu được tạo ra, chỉ phân loại dựa trên đầu vào dữ liệu. Với đặc điểm trên, mục đích của bộ sinh là đánh lừa bộ phân biệt, trong khi bộ phân biệt có vai trò phân loại dữ liệu đầu vào (đúng hoặc sai) để đưa ra phản hồi cho bộ sinh. Trò chơi kết thúc khi bộ phân biệt không thể xác định, phân biệt chính xác các dữ liệu đầu vào được tạo ra bởi bộ sinh.

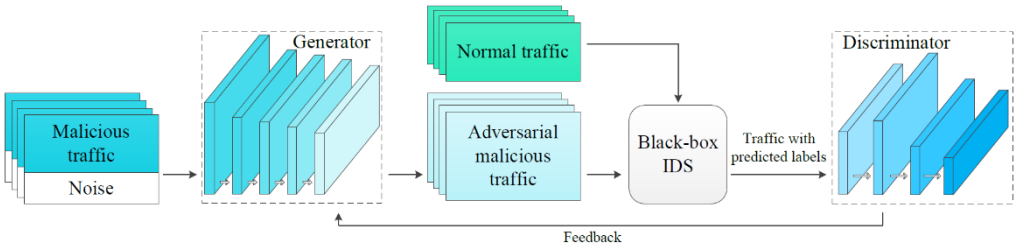
GAN đã cho thấy sự tiên tiến của mình trong việc tạo ra hình ảnh, âm thanh và văn bản [7, 8, 9]. Chia sẻ đặc điểm tương tự của văn bản hoặc câu, bảo mật thông tin cũng là lĩnh vực tiềm năng ho GAN gần đây. Các nghiên cứu hiện tại đã sử dụng GAN để cải thiện phát hiện phần mềm độc hại hoặc tạo các ví dụ về phần mềm độc hại đối với các cuộc tấn công đe dọa hơn [10, 11]. Tuy vậy, ở thời điểm hiện tại, giới nghiên cứu bảo mật đang thiếu các nghiên cứu về GAN được sử dụng trong IDS. Chính vì vậy, đề tài này là hướng nghiên cứu tiềm năng giúp các IDS phát hiện và phòng ngừa các loại tấn công mới chưa được xác định.

Tình hình nghiên cứu, phát triển trong và ngoài nước

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.

Grosse đề xuất áp dụng thuật toán dựa trên đạo hàm chuyển tiếp của các mạng thần kinh bị tấn công để tạo ra các mẫu mã độc Android đối kháng với các chức năng của mã độc được giữ lại[13].

Thuật toán học tăng cường với một tập hợp các hoạt động bảo tồn chức năng đã được sử dụng để tạo các mẫu phần mềm độc hại đối kháng [14]. Rosenberg đã tạo ra các mẫu đối kháng kết hợp các chuỗi lệnh gọi API và các tính năng tĩnh với khung tạo tấn công từ đầu đến cuối [15]. Al-Dujaili đã trình bày 4 phương pháp để tạo các mẫu mã độc được mã hóa nhị phân với chức năng độc hại được bảo tồn và sử dụng SLEIPNIR để huấn luyện các trình phát hiện mạnh mẽ [16]. Bên cạnh đó, vấn đề nghiên cứu thư rác đối kháng cũng nhận được nhiều sự quan tâm gần đây, lí do là Zhou đã tạo ra thư rác đối kháng với mô hình SVM đối kháng và nghiên cứu cách xây dựng bộ lọc thư rác mạnh mẽ hơn [17].



Hình 3: Phần huấn luyện của IDSGAN.

Bộ dữ liệu huấn luyện được chia ra thành lưu lượng thông thường và lưu lượng độc. Sau khi thêm nhiễu (noise), lưu lượng độc được gửi tới bộ sinh. Lưu lượng độc đối kháng và lưu lượng thông thường được dự đoán bởi IDS hộp đen. Nhãn dự đoán và nhãn gốc được sử dụng trong bộ phân biệt để mô tả IDS hộp đen. Mất mát của bộ sinh được tính toán dựa vào kết quả của bộ phân biệt và các nhãn dự đoán của IDS hộp đen.

GAN đã được áp dụng trong việc tạo ra các mẫu đối kháng trong lĩnh vực bảo mật thông tin. Hu đã đề xuất một khung GAN để tạo các mẫu mã độc đối kháng cho các cuộc tấn công hộp đen [11]. Hu cũng tận dụng một mô hình mới để tạo ra một số chuỗi API đối kháng sẽ được chèn vào chuỗi API ban đầu của mã độc để hình thành các cuộc tấn công, nhằm vượt qua, đánh lừa các hệ thống phát hiện Recurrent Neural Networks (RNN) [18].

5. Tính khoa học và tính mới của đề tài:

Như đã trình bày ở phần 4, các hệ thống phát hiện xâm nhập dần dần lộ ra lỗ hổng của nó trước hình thức tấn công đối kháng. Mô hình đối kháng mà tiêu biểu là GAN đã được áp dụng rộng rãi trong phát hiện phần mềm độc hại, nhưng có rất ít nghiên cứu về các mẫu lưu lượng độc đối kháng với IDS.

Nghiên cứu hướng tới đề xuất áp dụng mạng GAN vào trong quá trình phát sinh dữ liệu tấn công chống lại hệ thống phát hiện xâm nhập (IDSGAN). Mục tiêu của mô hình là triển khai IDSGAN để tạo các mẫu lưu lượng độc có thể đánh lừa và tránh sự phát hiện của các hệ thống phòng thủ điều này sẽ giúp đánh giá khả năng phát hiện các loại tấn công mới, kiểm tra sự an toàn của IDS trước khi được triển khai. Các thuật toán học máy được gán là IDS hộp đen, mô phỏng hệ thống phát hiện xâm nhập trong thực tế không xác định với cấu trúc bên trong của nó. Nghiên cứu thiết kế bộ sinh (Generator) và bộ phân biệt (Discriminator) trên cơ sở Wasserstein GAN vì các đặc tính vượt trội của nó [12]. Bộ sinh tạo ra các tấn công đối kháng: lưu lượng truy cập độc hại. Bộ phân biệt cung cấp thông tin phản hồi cho việc đào tạo bộ sinh và bắt chước IDS hộp đen.

6. Mục tiêu, đối tượng và phạm vi:

Mục tiêu:

- Nghiên cứu thiết kế IDSGAN, một khung GAN được cải tiến chống lại hệ thống phát hiện xâm nhập, tạo các mẫu lưu lượng độc hại đối kháng để tấn công IDS.

- Để bắt chước các cuộc tấn công và IDS trong thực tế, nghiên cứu này thực hiện tấn công hộp đen và sử dụng các thuật toán máy học làm IDS.

Đối tượng nghiên cứu:

- ✓ Các Machine Learning IDS.
- ✓ Các bộ dataset cho IDS.
- ✓ IDSGAN.

Phạm vi nghiên cứu: Hệ thống IDS có áp dụng có sử dụng phương pháp máy học ở dạng cơ bản.

7. **Nội dung, phương pháp dự định NC**.

Nội dung, phương pháp nghiên cứu chính:

- ✓ Tìm hiểu về IDS
- ✓ Tìm hiểu về máy học và mạng đối kháng tạo sinh (GAN).
- ✓ Xây dựng mô hình IDSGAN tạo các mẫu lưu lượng độc hại đối kháng tấn công IDS.
- ✓ Thực nghiệm sử dụng lưu lượng độc đối kháng của IDSGAN tấn công các thuật toán máy học được dùng làm IDS và đánh giá kết quả.

a) **Nội dung 1**: Tìm hiểu về IDS

- *Mục tiêu*: Nắm được kiến trúc, nguyên tắc hoạt động của IDS.
- *Phương pháp*: Ôn lại kiến thức đã được học trong môn IDS.

b) **Nội dung 2:** Tìm hiểu về máy học và mạng đối kháng tạo sinh (GAN)

- *Mục tiêu*: Nắm được các khái niệm cơ bản trong máy học và cơ chế hoạt động của mạng GAN, Wasserstein GAN là mạng sẽ được sử dụng cho IDSGAN.
- *Phương pháp:* Nghiên cứu tài liệu về máy học, mạng GAN, Wasserstein GAN.

c) **Nội dung 3**: Xây dựng mô hình IDSGAN tạo các mẫu lưu lượng độc đối kháng tấn công IDS

- *Mục tiêu*: Triển khai mô hình IDSGAN tạo ra các mẫu lưu lượng độc đối kháng tấn công IDS.
- *Phương pháp*:
  - Tìm hiểu về các dataset cho IDS.
  - Xây dựng IDSGAN, dựa trên Wasserstein GAN.

d) **Nội dung 4:** Thực nghiệm sử dụng lưu lượng độc đối kháng của IDSGAN tấn công các thuật toán máy học được dùng làm IDS và đánh giá kết quả

- *Mục tiệu*:
  - Thử nghiệm tấn công IDS bằng lưu lượng độc đối kháng tạo ra từ IDSGAN.
  - Thống kê kết quả và đưa ra đánh giá.
  - Đưa ra báo cáo tổng quan về quá trình nghiên cứu thực hiện đề tài.
- *Phương pháp*: Thực nghiệm kết quả tấn công của IDSGAN.

8. **Kế hoạch bố trí thời gian NC**

Kế hoạch sơ lược việc bố trí thời gian nghiên cứu:

Thời gian	Nội dung	Kết quả mong đợi
14/02/2019 – 01/03/2019	Tìm hiểu về IDS	Nắm được kiến trúc, nguyên tắc hoạt động của IDS.
Tháng 4. 2019	Tìm hiểu về máy học và mạng đối kháng tạo sinh (GAN).	Nắm được các khái niệm cơ bản trong máy học và cơ chế hoạt động của mạng GAN, Wasserstein GAN là mạng sẽ được sử dụng cho IDSGAN.
Tháng 5. 2019	Xây dựng mô hình IDSGAN tạo các mẫu lưu lượng độc đối kháng tấn công IDS.	Xây dựng thành công mô hình.
Tháng 6. 2019	Thực nghiệm sử dụng lưu lượng độc đối kháng của IDSGAN tấn công các thuật toán máy học được dùng làm IDS và đánh giá kết quả.	Có được kết quả thực nghiệm và đưa ra được báo cáo tổng quan về quá trình thực hiện đề tài.

✓ Gặp gỡ giảng viên hướng dẫn mỗi hai tuần một lần (vào thứ 3 hoặc thứ 4).

9.Tài liệu tham khảo

[1] Tsai, C.-F.; Hsu, Y.-F.; Lin, C.-Y.; and Lin, W.-Y. 2009. Intrusion detection by machine learning: A review. Expert Systems with Applications 36(10):11994– 12000.

[2] Li, Z.; Qin, Z.; Huang, K.; Yang, X.; and Ye, S. 2017. Intrusion detection using convolutional neural networks for representation learning. In Proceedings of International Conference on Neural Information Processing, 858–866. Guangzhou, China: Springer.

[3] Lin, S. Z.; Shi, Y.; and Xue, Z. 2018. Character-level intrusion detection based on convolutional neural networks. In Proceedings of International Joint Conference of Neural Networks, 3454–3461. Rio de Janeiro, Brazil: IEEE.

[4] Carlini, N., and Wagner, D. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 3–14. Dallas, TX, USA: ACM.

[5] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680.

[6] Lee, H.; Han, S.; and Lee, J. 2017. Generative adversarial trainer: Defense to adversarial perturbations with gan. arXiv preprint arXiv:1705.03387.

[7] Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In CVPR, volume 2, 4.

[8] Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 34–41. New Orleans, Louisiana: AAAI.

[9] Su, H.; Shen, X.; Hu, P.; Li, W.; and Chen, Y. 2018. Dialogue generation with gan. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 8163–8164. New Orleans, Louisiana: AAAI.

[10] Kim, J.-Y.; Bu, S.-J.; and Cho, S.B. 2017. Malware detection using deep transferred generative adversarial networks. In Proceedings of International Conference on Neural Information Processing, 556–564. Guangzhou, China: Springer.

[11] Hu, W., and Tan, Y. 2017b. Generating adversarial malware examples for black-box attacks based on gan. arXiv preprint arXiv:1702.05983.

[12] Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. arXiv preprint arXiv:1701.07875.

[Davis and Clark 2011] Davis, J. J., and Clark, A. J. 2011. Data preprocessing for anomaly based network intrusion detection: A review. computers & security 30(6-7):353–375.

[13] Grosse, K.; Papernot, N.; Manoharan, P.; Backes, M.; and McDaniel, P. 2016. Adversarial perturbations against deep neural networks for malware classification. arXiv preprint arXiv:1606.04435.

[14] Anderson, H. S.; Kharkar, A.; Filar, B.; and Roth, P. 2017. Evading machine learning malware detection. Black Hat.

[15] Rosenberg, I.; Shabtai, A.; Rokach, L.; and Elovici, Y. 2018. Generic black-box end-to-end attack against state of the art api call based malware classifiers. arXiv preprint arXiv:1804.08778.

[16] Al-Dujaili, A.; Huang, A.; Hemberg, E.; and OReilly, U.-M. 2018. Adversarial deep learning for robust detection of binary encoded malware. In Proceedings of the 39th IEEE Symposium on Security and Privacy, 76–82. San Francisco, CA, USA: IEEE.

[17] Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; and Xi, B. 2012. Adversarial support vector machine learning. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 1059–1067. Beijing, China: ACM.

[18] Hu, W., and Tan, Y. 2017a. Blackbox attacks against rnn based malware detection algorithms. arXiv preprint arXiv:1705.08131.

TP. HCM, ngày tháng năm 2020

NGƯỜI HƯỚNG DẪN	**SINH VIÊN KÝ TÊN**
(*Họ tên và chữ ký*)	(*Họ tên và chữ ký*)
PHẠM VĂN HẬU	LÊ KHẮC TIẾN