THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  https://www.youtube.com/watch)
- Link slides (dạng .pdf đặt trên Github của nhóm):
  (ví dụ: https://github.com/mynameuit/CS2205.xxx/TenDe
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Kiều Hồng Khang
- MSSV: 240201042

- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 7.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 3
- Link Github:
  https://github.com/mynameuit/CS2205.xxx/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÁT HIỆN TẤN CÔNG DDOS ĐỐI KHÁNG BẰNG MÔ HÌNH GAN VỚI BỘ PHÂN BIỆT KÉP (GAN-DD)

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DETECTING ADVERSARIAL DDOS ATTACKS VIA GANS WITH DUAL DISCRIMINATORS

TÓM TẮT *(Tối đa 400 từ)*

Các cuộc tấn công từ chối dịch vụ phân tán (DDoS) đang trở thành mối đe dọa nghiêm trọng với các hệ thống mạng hiện nay do ngày càng tăng cả về số lượng và mức độ tinh vi. Một thách thức mới đó là sự kết hợp của kỹ thuật tấn công đối kháng (adversarial attacks), có khả năng gây nhiễu loạn dữ liệu đầu vào, nhằm đánh lừa và làm giảm đáng kể hiệu quả các mô hình phát hiện truyền thống.

Để khắc phục vấn đề trên, nghiên cứu này đề xuất sử dụng mô hình Generative Adversarial Networks with Dual Discriminators (GAN-DD). Mô hình này phát triển dựa trên mạng đối sinh GAN, nhưng điểm khác biệt cơ bản là tích hợp hai bộ phân biệt song song. Thiết kế hai bộ phân biệt giúp mô hình tăng khả năng học được các đặc trưng sâu sắc hơn từ dữ liệu, đồng thời có khả năng phân biệt rõ ràng giữa lưu lượng hợp lệ và lưu lượng tấn công, ngay cả khi dữ liệu bị tác động bởi kỹ thuật tấn công đối kháng.

Nghiên cứu sẽ thực hiện các bước cụ thể bao gồm: xây dựng và huấn luyện mô hình GAN-DD với tập dữ liệu lưu lượng mạng thực tế, tạo ra các mẫu lưu lượng bất thường dựa trên kỹ thuật đối kháng, và kiểm tra khả năng phát hiện của mô hình trước những dạng tấn công này.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong bối cảnh Internet ngày càng phát triển mạnh mẽ, các cuộc tấn công từ chối dịch vụ phân tán (DDoS) không chỉ gia tăng về số lượng mà còn trở nên phức tạp hơn. Đặc

biệt, sự xuất hiện của các kỹ thuật tấn công đối kháng (adversarial attacks) đã khiến cho việc phát hiện và phòng chống DDoS gặp phải nhiều khó khăn và thách thức.

Mặc dù các phương pháp học máy và học sâu đang được sử dụng rộng rãi trong lĩnh vực an ninh mạng, chúng vẫn tồn tại hạn chế khi đối mặt với các mẫu lưu lượng tấn công được ngụy trang tinh vi bởi kỹ thuật đối kháng.

Đề tài này giới thiệu mô hình GAN-DD – một biến thể của mạng đối sinh (GAN) với cấu trúc hai bộ phân biệt (dual discriminators), nhằm cải thiện khả năng phân loại lưu lượng mạng hợp lệ và lưu lượng tấn công, bao gồm cả các trường hợp dữ liệu bị nhiễu do tấn công đối kháng.

Hướng tiếp cận này được kế thừa và phát triển dựa trên nghiên cứu của Shieh et al. (2022), trong đó mô hình GAN-DD với hai bộ phân biệt riêng biệt đã được đề xuất: một bộ chuyên phân biệt dữ liệu thật và giả, bộ còn lại tập trung nhận dạng dữ liệu bị nhiễu đối kháng. Mô hình này đã cho thấy hiệu quả vượt trội trong việc phát hiện các cuộc tấn công DDoS đối kháng, nhất là trong những tình huống phức tạp mà các mô hình học sâu truyền thống khó xử lý tốt.

Trên cơ sở đó, đề tài hiện tại sẽ triển khai lại kiến trúc GAN-DD trên các bộ dữ liệu mới, đánh giá chi tiết hiệu quả, đồng thời thực hiện các điều chỉnh và tối ưu hóa nhằm nâng cao hơn nữa khả năng ứng dụng thực tế của mô hình trong các hệ thống mạng hiện đại.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Xây dựng kiến trúc mô hình GAN-DD có khả năng phát hiện lưu lượng DDoS nhiễu đối kháng với độ chính xác và độ nhạy vượt trội so với các mô hình truyền thống.
2. Thực nghiệm trên tập dữ liệu thực tế và dữ liệu đối kháng để đánh giá độ hiệu quả mô hình trong môi trường tấn công thực tế.
3. Đề xuất chiến lược triển khai hệ thống phát hiện GAN-DD trong các hệ thống

mạng thật, bao gồm phân tích thách thức và hướng khắc phục.

NỘI DUNG VÀ PHƯƠNG PHÁP

- NỘI DUNG NGHIÊN CỨU

1. Khảo sát các phương pháp phát hiện tấn công DDoS truyền thống

Phân tích các kỹ thuật phát hiện DDoS phổ biến như detection theo ngưỡng, theo hành vi (anomaly-based), SVM, Decision Tree, Random Forest và các mô hình học sâu như CNN, LSTM. Nhận diện điểm mạnh, hạn chế và lý do các phương pháp này dễ bị tấn công đối kháng vượt qua.

2. Tìm hiểu mô hình GAN-DD

Trình bày kiến trúc của mô hình GAN-DD gồm:

- Generator (G) sinh lưu lượng mạng giả mạo.
- Discriminator 1 (D1) phân biệt dữ liệu thật/giả.
- Discriminator 2 (D2) phát hiện dữ liệu bị nhiễu đối kháng.

Trình bày cách phối hợp giữa các thành phần và cơ chế huấn luyện đối kháng với hàm mất mát phù hợp.

3. So sánh GAN-DD với các biến thể khác của GAN

So sánh GAN-DD với các kiến trúc khác như:

- Vanilla GAN
- WGAN, WGAN-GP
- LSGAN

Các tiêu chí được sử dụng để so sánh bao gồm: độ chính xác, thời gian hội tụ, khả năng phát hiện tấn công đối kháng và sự ổn định trong quá trình huấn luyện mô hình.

4. Ứng dụng GAN-DD vào phát hiện tấn công DDoS

Áp dụng mô hình vào bài toán phân loại lưu lượng mạng (hợp lệ/tấn công).

Thực hiện pipeline: tiền xử lý dữ liệu → huấn luyện mô hình → đánh giá → xuất kết quả dự đoán.

5. Đánh giá hiệu quả của GAN-DD

Dùng các chỉ số đánh giá: Accuracy, Precision, Recall, F1-score.

So sánh mô hình với các baseline như SVM, DNN, Random Forest.

Thử nghiệm trong điều kiện có dữ liệu nhiễu đối kháng để đánh giá tính bền vững.

6. Phân tích thách thức và giải pháp

Thảo luận các thách thức khi triển khai thực tế như:

- Tài nguyên tính toán (GPU, RAM).
- Nguy cơ overfitting.
- Khả năng tổng quát hóa kém với dữ liệu mới.

Đề xuất hướng cải tiến như:

- Mô hình GAN nhẹ (Lightweight GAN),
- Học liên miền (Domain Adaptation),
- Kết hợp attention hoặc kỹ thuật regularization.

- PHƯƠNG PHÁP NGHIÊN CỨU

1. Thu thập và tiền xử lý dữ liệu

Sử dụng tập dữ liệu công khai **CICIDS2017** làm nguồn dữ liệu đầu vào. Dữ liệu được tiền xử lý qua các bước:

- Loại bỏ bản ghi lỗi hoặc thiếu giá trị;
- Mã hóa nhãn và chuẩn hóa đặc trưng;
- Chia dữ liệu thành tập huấn luyện và kiểm thử;
- Sinh thêm các mẫu dữ liệu bị nhiễu đối kháng bằng kỹ thuật như **FGSM** để mô phỏng các kịch bản tấn công thực tế.

2. Thiết kế mô hình GAN-DD

Mô hình GAN-DD bao gồm ba thành phần chính:

- **Generator (G)**: tạo dữ liệu lưu lượng mạng giả mạo có đặc trưng giống dữ liệu tấn

công thực.

- **Discriminator 1 (D1)**: phân biệt giữa dữ liệu thật và dữ liệu do Generator sinh ra.
- **Discriminator 2 (D2)**: phát hiện dữ liệu bị nhiễu đối kháng.

Mô hình được xây dựng bằng framework **TensorFlow/Keras**, sử dụng các kiến trúc mạng sâu (Deep Neural Network) với hàm mất mát là sự kết hợp giữa binary cross-entropy và adversarial loss.

3. Huấn luyện mô hình

Áp dụng chiến lược huấn luyện luân phiên (alternating training) giữa G và hai bộ phân biệt D1, D2. Các kỹ thuật như early stopping, dropout, và batch normalization được sử dụng để kiểm soát overfitting và đảm bảo tính ổn định trong quá trình huấn luyện.

4. Đánh giá và so sánh

Hiệu quả của mô hình được đánh giá bằng các chỉ số: Accuracy, Precision, Recall, F1-score.

Kết quả của GAN-DD được so sánh với các mô hình baseline như:

- SVM, Random Forest, DNN;
- Các biến thể GAN khác như Vanilla GAN, WGAN, LSGAN.

5. Phân tích và đề xuất cải tiến

Cuối cùng, kết quả được phân tích để xác định ưu/nhược điểm của mô hình GAN-DD, từ đó đề xuất các hướng cải tiến như:

- Tối ưu kiến trúc mạng để giảm chi phí tính toán;
- Tăng khả năng tổng quát hóa bằng multi-domain learning;
- Ứng dụng các kỹ thuật nâng cao như attention mechanism hoặc lightweight GAN cho môi trường tài nguyên hạn chế.

KẾT QUẢ MONG ĐỢI

- Phát triển mô hình GAN-DD có khả năng phát hiện hiệu quả các cuộc tấn công

DDoS, kể cả khi dữ liệu bị nhiễu đối kháng.

- Đạt độ chính xác trên 90%, F1-score ≥ 85% trên tập dữ liệu thực nghiệm.
- Hiệu quả mô hình vượt trội so với các phương pháp truyền thống và các biến thể GAN cơ bản.
- Đề xuất giải pháp triển khai mô hình trong môi trường thực tế, đảm bảo hiệu suất và khả năng ứng dụng.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 2, 2672–2680.

[2]. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of Wasserstein GANs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; 5769–5779

[3]. Nguyen, T.D.; Le, T.; Vu, H.; Phung, D. Dual Discriminator Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017; pp. 2667–2677.

[4]. Zhang, X.; Zhao, Y.; Zhang, H. Dual-discriminator GAN: A GAN way of profile face recognition. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; 162–166

[5] Chin-Shiuh Shieh, Thanh-Tuan Nguyen, Wan-Wei Lin, Yong-Lin Huang, Mong-Fong Horng, Tsair-Fwu Lee, Denis Miu: Detection of Adversarial DDoS Attacks Using Generative Adversarial Networks with Dual Discriminators. Symmetry 14(1): 66 (2022). DOI: 10.3390/sym14010066