


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/lHFBZgpaWsE>
- Link slides (dạng .pdf đặt trên Github):
https://github.com/minh240899/CS2205.MAR2024/blob/main/PE_MALWARE_DETECTION_USING_OPTIMIZED_FEATURES.pdf
- *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*

<ul style="list-style-type: none">● Họ và Tên: Lê Văn Minh● MSSV: 230202011 	<ul style="list-style-type: none">● Lớp: CS2205.APR2023● Tự đánh giá (điểm tổng kết môn): 8.0/10● Số buổi vắng: 2● Số câu hỏi QT cá nhân: 2● Link Github: https://github.com/minh240899/CS2205.MAR2024
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA) PHÁT HIỆN MÃ ĐỘC PE SỬ DỤNG ĐẶC TRƯNG ĐƯỢC TỐI ƯU HÓA
TÊN ĐỀ TÀI TIẾNG ANH (IN HOA) PE MALWARE DETECTION USING OPTIMIZED FEATURES
TÓM TẮT <i>(Tối đa 400 từ)</i> <p>Sự tiến bộ vượt bậc của công nghệ thông tin đã biến máy tính thành một phần không thể thiếu trong cuộc sống hàng ngày, nhưng điều này cũng tạo điều kiện cho các cuộc tấn công mạng nguy hiểm, từ phá hoại máy tính đến lợi dụng kinh tế, đe dọa sự an toàn ngày càng nghiêm trọng và phức tạp. Trong các phương thức tấn công, malware (phần mềm độc hại) nổi bật về độ phổ biến và nguy hiểm, cho phép kẻ tấn công xâm nhập hệ thống, từ ăn cắp dữ liệu bí mật đến kiểm soát toàn bộ hệ thống và yêu cầu tiền chuộc. Nghiên cứu cách phát hiện malware sử dụng tệp PE (Portable Executable) trên hệ điều hành Windows (cụ thể là malware Windows PE) trở nên quan trọng và cấp bách. Để chống lại mối đe dọa này, việc áp dụng các kỹ thuật học máy và học sâu như LSTM và XGBoost là không thể thiếu. Học sâu, một mô hình phức tạp của học máy truyền thống và là một tiến bộ của mạng nơ-ron, đã mang lại những kết quả ấn tượng trong nhiều lĩnh vực, bao gồm an ninh mạng. Các nghiên cứu gần đây cho thấy học sâu đã được sử dụng hiệu quả để phát hiện mã độc. Các phương pháp này bao gồm kết hợp phép chiếu ngẫu nhiên và mạng nơ-ron để phân loại mã độc quy mô lớn, cũng như sử dụng các mô hình ngôn ngữ tiên tiến để phát hiện mã độc dựa trên chỉ thị nguyên thủy. Hơn nữa, các phương pháp học máy và học sâu như LSTM (Long Short-Term Memory) và XGBoost (eXtreme Gradient Boosting) cũng đã được áp dụng và đánh giá hiệu suất trong việc phát hiện malware. Nghiên cứu này thử nghiệm và so sánh hiệu suất của các mô hình LSTM và XGBoost trong việc phát hiện mã độc sử dụng các tệp PE, với hy vọng đóng góp quan trọng vào lĩnh vực này và cung cấp các phương pháp hiệu quả để bảo vệ hệ thống khỏi các cuộc tấn công mạng.</p>

GIỚI THIỆU *(Tối đa 1 trang A4)*

Với sự phát triển nhanh chóng của công nghệ thông tin, máy tính đã trở thành một phần không thể thiếu trong cuộc sống, kéo theo các nguy cơ tấn công mạng ngày càng cao, đặc biệt là từ malware (phần mềm độc hại). Malware cho phép kẻ tấn công xâm nhập hệ thống, ăn cắp dữ liệu, kiểm soát hệ thống và đòi tiền chuộc. Vì vậy, nghiên cứu cách phát hiện malware sử dụng tệp PE (Portable Executable) trên Windows là rất quan trọng.

Các kỹ thuật học máy và học sâu như LSTM và XGBoost đóng vai trò quan trọng trong việc này. Học sâu, một tiến bộ của học máy truyền thống, đã cho thấy hiệu quả cao trong nhiều lĩnh vực, bao gồm an ninh mạng. Các nghiên cứu cho thấy học sâu có thể phát hiện mã độc hiệu quả, chẳng hạn như sử dụng mạng nơ-ron để phân loại mã độc quy mô lớn và các mô hình ngôn ngữ tiên tiến để phát hiện mã độc dựa trên chỉ thị nguyên thủy.

Nghiên cứu này so sánh hiệu suất của các mô hình LSTM và XGBoost trong việc phát hiện mã độc sử dụng tệp PE, với mục tiêu cung cấp phương pháp hiệu quả bảo vệ hệ thống khỏi các cuộc tấn công mạng. Mục tiêu bao gồm nghiên cứu về mã độc PE, đặc trưng của nó, và các kỹ thuật máy học để rút trích đặc trưng từ mã độc PE. Quá trình nghiên cứu sẽ cải tiến việc rút trích đặc trưng để phân loại bằng XGBoost và sử dụng LSTM để phân loại mã độc PE dựa trên các đặc trưng tối ưu.

Nội dung nghiên cứu gồm tìm hiểu các công trình đã công bố, xây dựng mô hình phát hiện mã độc PE bằng LSTM và XGBoost, và đánh giá mô hình trên bộ dữ liệu BODMAS từ Đại học Illinois Urbana-Champaign. Quá trình xây dựng mô hình bắt đầu từ việc thu thập và chuẩn bị dữ liệu, tiền xử lý và trích xuất các đặc trưng cần thiết. Các mô hình sẽ được huấn luyện và đánh giá trên dữ liệu đã xử lý, với hiệu suất đo lường bằng các chỉ số như độ chính xác, độ nhạy và độ đặc hiệu. Cuối cùng, mô hình sẽ được điều chỉnh để đạt hiệu suất tối ưu trong việc phát hiện mã độc PE

MỤC TIÊU

- Nghiên cứu về mã độc PE, bao gồm cách thức hoạt động và các đặc trưng liên quan đến loại mã độc này.
- Nghiên cứu kỹ thuật sử dụng máy học và học sâu để rút trích các đặc trưng từ mã độc PE, nhằm cải tiến quá trình rút trích và tạo ra tập đặc trưng tối ưu cho phân loại sử dụng XGBoost, đánh giá sự đóng góp của từng đặc trưng.
- Sử dụng mô hình LSTM để phân loại mã độc PE dựa vào các đặc trưng đã được tối ưu từ quá trình rút trích trước đó.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nghiên cứu về mã độc PE và các đặc trưng

Tìm hiểu cách thức hoạt động và các đặc trưng của mã độc PE (Portable Executable).

Xem xét các công trình khoa học đã được công bố liên quan đến chủ đề này để hiểu rõ hơn về các phương pháp hiện có.

Xây dựng và đánh giá mô hình phát hiện mã độc PE

Xây dựng mô hình bằng LSTM và XGBoost: Quá trình này bắt đầu từ việc thu thập và chuẩn bị dữ liệu, bao gồm các mẫu mã độc PE và các mẫu không độc hại để huấn luyện và kiểm tra mô hình. Dữ liệu được tiền xử lý để chuẩn hóa và trích xuất các đặc trưng cần thiết.

Huấn luyện và đánh giá mô hình: Các mô hình LSTM và XGBoost được xây dựng và huấn luyện trên dữ liệu đã tiền xử lý. LSTM học các mẫu và mối quan hệ phức tạp giữa các chuỗi dữ liệu, trong khi XGBoost tối ưu hóa các dự đoán. Mô hình sau đó được đánh giá trên một tập dữ liệu kiểm tra độc lập với các phép đo hiệu suất như độ chính xác, độ nhạy và độ đặc biệt.

Điều chỉnh và cải tiến mô hình: Dựa trên kết quả đánh giá, mô hình được điều chỉnh để đạt hiệu suất tốt nhất, bao gồm điều chỉnh siêu tham số và áp dụng các kỹ thuật bổ sung như tái chọn mẫu và tăng cường dữ liệu.

Đánh giá trên cơ sở dữ liệu BODMAS: Bộ dữ liệu BODMAS (Blue Hexagon Open Dataset for Malware Analysis) được sử dụng để đánh giá mô hình. BODMAS chứa các mẫu phần mềm độc hại và không độc hại được thu thập và chăm sóc kỹ lưỡng, cung cấp một cơ sở dữ liệu phong phú để kiểm tra và so sánh hiệu suất mô hình.

Phương pháp thực hiện:

1. Sử dụng XGBoost để chọn lọc đặc trưng:

XGBoost là một thuật toán học máy mạnh mẽ, nổi bật với khả năng chọn lọc đặc trưng thông qua thông số "feature importance". Feature importance đo lường mức độ ảnh hưởng của mỗi đặc trưng đến kết quả dự đoán của mô hình.

Trong XGBoost, feature importance được tính dựa trên mức độ giảm giá trị của hàm mất mát khi một đặc trưng cụ thể được sử dụng để phân tách dữ liệu. Đặc trưng quan trọng hơn sẽ có mức độ giảm lớn hơn, giúp cải thiện hiệu suất và hiệu quả của mô hình.

2. Sử dụng LSTM để phân loại mã độc PE:

LSTM là một loại mạng neural tái phát hiện được thiết kế để xử lý dữ liệu chuỗi, có khả năng nhớ và học các mẫu dài hạn từ dữ liệu đầu vào. LSTM được sử dụng để học các mẫu và mối quan hệ phức tạp giữa các chuỗi dữ liệu, giúp phát hiện các hành vi độc hại trong mã độc.

Các đặc trưng tối ưu từ XGBoost được sử dụng làm đầu vào để huấn luyện mô hình LSTM. Điều này giúp LSTM nhận biết các mẫu dựa trên chuỗi dữ liệu, cải thiện khả năng phát hiện các hành vi độc hại và các biến thể malware.

3. Kết hợp sử dụng LSTM và các đặc trưng tối ưu từ XGBoost cung cấp một phương pháp mạnh mẽ và hiệu quả cho việc phát hiện mã độc trong môi trường an ninh mạng, tận dụng cả khả năng xử lý dữ liệu chuỗi của LSTM và khả năng chọn lọc đặc trưng của XGBoost để xây dựng hệ thống phát hiện malware linh hoạt và mạnh mẽ.

KẾT QUẢ MONG ĐỢI

- Xây dựng mô hình phát hiện mã độc PE bằng LSTM và XGBoost.
- Đánh giá mô hình trên cơ sở dữ liệu BODMAS cho kết quả thực nghiệm cao và thời gian xử lý tốt hơn.
- So sánh mô hình được đề xuất với các phương pháp được công bố trong các bài nghiên cứu gần đây, cho thấy phương pháp được đề xuất có khả năng nhận diện mã độc PE tốt và thực hiện nhanh

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1] T. Rezaei and A. Hamze, "An Efficient Approach For Malware Detection Using PE Header Specifications," 2020 6th International Conference on Web Research (ICWR), Tehran, Iran, 2020, pp. 234-239, doi: 10.1109/ICWR49608.2020.9122312.
- [2] C. Galen and R. Steele, "Evaluating Performance Maintenance and Deterioration Over Time of Machine Learning-based Malware Detection Models on the EMBER PE Dataset," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Paris, France, 2020, pp. 1-7, doi: 10.1109/SNAMS52053.2020.9336538.
- [3] Muhamad Malik Matin and B. Rahardjo, "A Framework for Collecting and Analysis PE Malware Using Modern Honey Network (MHN)," 2020 8th International Conference on Cyber and IT Service Management (CITSM), Pangkal, Indonesia, 2020, pp. 1-5, doi: 10.1109/CITSM50537.2020.9268810.
- [4] M. Kim, "Research on Malware Detection System Using Artificial Intelligence," 2022 IEEE/ACIS 7th International Conference on Big Data, Cloud Computing, and Data Science (BCD), Danang, Vietnam, 2022, pp. 211-213, doi: 10.1109/BCD54882.2022.9900792.
- [5] P. Singh, S. K. Borgohain and J. Kumar, "Performance Enhancement of SVM-based ML Malware Detection Model Using Data Preprocessing," 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 2022, pp. 1-4, doi: 10.1109/ICEFEET51821.2022.9848192.

