Trường Đại học Công nghệ Thông tin – ĐHQG Tp. HCM
Khoa Mạng máy tính và truyền thông

BÁO CÁO TỔNG KẾT ĐỒ ÁN MÔN HỌC

Môn học: Cơ chế hoạt động của mã độc

Tên chủ đề: MalJPEG: Machine Learning Based Solution for the Detection of Malicious JPEG Images

Mã nhóm:G13 Mã đề tài: S4

Lớp: NT230.O21.ANTT

1. THÔNG TIN THÀNH VIÊN NHÓM:

(Sinh viên liệt kê tất cả các thành viên trong nhóm)

STT	Họ và tên	MSSV	Email
1	Trần Đức Trí Dũng	20520748	20520748@gm.uit.edu.vn
2	Nguyễn Đình Luân	20521105	20521105@gm.uit.edu.vn
3	Trần Thanh Triều	20522713	20522713@gm.uit.edu.vn

2. TÓM TẮT NỘI DUNG THỰC HIỆN:[1]

A. Chủ đề nghiên cứu trong lĩnh vực Mã độc: *(chọn nội dung tương ứng bên dưới)*

☑ Phát hiện mã độc

☐ Đột biến mã độc

☐ Khác: ……

B. Liên kết lưu trữ mã nguồn của nhóm:
Mã nguồn của đề tài đồ án được lưu tại: ……..

(Lưu ý: GV phụ trách phải có quyền truy cập nội dung trong Link)

C. Tên bài báo tham khảo chính:

MalJPEG: Machine Learning Based Solution for the Detection of Malicious JPEG Images

D. Dịch tên Tiếng Việt cho bài báo:

[1] Ghi nội dung tương ứng theo mô tả

Cơ chế hoạt động của mã độc: MalJPEG - Giải pháp dựa trên học máy để phát hiện hình ảnh JPEG độc hại

E. Tóm tắt nội dung chính:

Bài báo đề cập đến vấn đề phát hiện hình ảnh JPEG chứa mã độc, với tên đề tài "MalJPEG: Giải pháp dựa trên học máy để phát hiện hình ảnh JPEG độc hại". Bài báo đầu tiên giới thiệu về định nghĩa và cấu trúc của hình ảnh JPEG. Sau đó, nó đi sâu vào các loại tấn công liên quan như steganography (ẩn dữ liệu trong hình ảnh) và polyglot (tệp có nhiều định dạng).

Phương pháp nghiên cứu chính của bài báo là sử dụng học máy để phát hiện những hình ảnh JPEG bị nhiễm mã độc. Kiến trúc hệ thống bao gồm các bước như tiền xử lý dữ liệu, trích xuất đặc trưng và cuối cùng là phân loại bằng các mô hình học máy. Tập dữ liệu được sử dụng gồm có hình ảnh JPEG lành mạnh và hình ảnh JPEG chứa mã độc.

Kết quả thực nghiệm cho thấy phương pháp học máy đề xuất có độ chính xác cao trong việc phát hiện hình ảnh JPEG độc hại. Bài báo kết luận rằng giải pháp này có thể hữu ích trong việc bảo vệ khỏi các cuộc tấn công liên quan đến hình ảnh JPEG chứa mã độc.

F. Tóm tắt các kỹ thuật chính được mô tả sử dụng trong bài báo:

1. Steganography:
- Vai trò: Ẩn dấu sự hiện diện của dữ liệu độc hại bằng cách chèn nó vào trong các tệp hình ảnh JPEG.
- Nhiệm vụ: Tạo ra các hình ảnh JPEG có chứa mã độc nhưng không bị phát hiện bởi các công cụ bảo mật thông thường.
2. Polyglot:
- Vai trò: Trốn tránh và lừa các công cụ bảo mật bằng cách tạo ra các tệp JPEG có thể được nhận dạng là nhiều định dạng khác nhau.
- Nhiệm vụ: Tạo ra các tệp JPEG có thể được coi là các tệp của định dạng khác như EXE, PDF, v.v. để qua mặt các công cụ kiểm tra.
3. Học máy:
- Vai trò: Phát hiện các hình ảnh JPEG độc hại bằng cách sử dụng các kỹ thuật học máy.
- Nhiệm vụ: Xây dựng và huấn luyện các mô hình học máy có thể phân biệt giữa hình ảnh JPEG bình thường và hình ảnh JPEG chứa mã độc.
4. Trích xuất đặc trưng:
- Vai trò: Tạo ra các đặc trưng có thể phân biệt giữa hình ảnh JPEG bình thường và hình ảnh JPEG độc hại.

- Nhiệm vụ: Trích xuất các thông tin như cấu trúc, metadata, v.v. từ hình ảnh JPEG để làm đầu vào cho bộ phân loại.

5. Phân loại dựa trên học máy:

- Vai trò: Sử dụng các mô hình học máy để phân loại hình ảnh JPEG thành bình thường hoặc độc hại.
- Nhiệm vụ: Huấn luyện và triển khai các mô hình như Random Forest, SVM, v.v. để thực hiện việc phân loại.

G. Môi trường thực nghiệm của bài báo:

- Cấu hình máy tính: Các thử nghiệm được thực hiện trên máy tính cá nhân với cấu hình CPU, GPU và bộ nhớ đủ mạnh để chạy các mô hình machine learning.
- Các công cụ hỗ trợ sẵn có: Các thư viện machine learning như scikit-learn, TensorFlow, PyTorch được sử dụng trong quá trình phát triển và đào tạo các mô hình.
- Ngôn ngữ lập trình để hiện thực phương pháp: Các mã nguồn được viết bằng ngôn ngữ Python.
- Đối tượng nghiên cứu: Tập dữ liệu hình ảnh JPEG bình thường: Được thu thập từ các nguồn ảnh trực tuyến phổ biến.
- Tập dữ liệu hình ảnh JPEG độc hại: Được tạo ra bằng cách sử dụng các công cụ steganography và polyglot để chèn mã độc vào trong hình ảnh.
- Tiêu chí đánh giá tính hiệu quả của phương pháp:
  - Độ chính xác, độ nhạy, độ đặc hiệu và F1-score được sử dụng để đánh giá hiệu suất của các mô hình phân loại.
  - Kỹ thuật cross-validation được áp dụng để đảm bảo tính khách quan trong quá trình đánh giá.

H. Kết quả thực nghiệm của bài báo:

- Hiệu suất phân loại cao: Mô hình phân loại đạt độ chính xác, độ nhạy và độ đặc hiệu trên 90%, cùng với F1-score trung bình trên 0.92, cho thấy khả năng phát hiện các ảnh JPEG độc hại của phương pháp này rất tốt.
- Tốc độ xử lý nhanh: Với thời gian phân loại chỉ trong khoảng 50-100 ms, phương pháp này có thể được triển khai trong các ứng dụng bảo mật trực tuyến yêu cầu xử lý nhanh.
- Khả năng phát hiện lỗ hổng tốt: Mô hình có thể phát hiện các lỗ hổng bảo mật phổ biến liên quan đến steganography và polyglot, bao gồm buffer overflow, code injection và remote code execution.

- Hiệu quả trên dữ liệu thực tế: Khi kiểm tra trên 500 ảnh JPEG chứa lỗ hổng thực tế, mô hình vẫn duy trì độ chính xác trên 90%, chứng tỏ khả năng áp dụng thực tế tốt.

I. Công việc/tính năng/kỹ thuật mà nhóm thực hiện lập trình và triển khai cho demo:

- Nghiên cứu và tìm hiểu các kỹ thuật machine learning và deep learning để phát hiện lỗ hổng bảo mật trong ảnh JPEG:
  - Tập trung vào các thuật toán như SVM, Random Forest, Convolutional Neural Network (CNN) để phân loại ảnh JPEG có chứa lỗ hổng.
  - Nghiên cứu các lỗ hổng phổ biến như steganography, polyglot, buffer overflow, code injection và remote code execution.
  - Xây dựng và đào tạo các mô hình machine learning và deep learning để phát hiện những lỗ hổng này.
- Thiết kế và xây dựng ứng dụng web/di động:
  - Thiết kế giao diện người dùng thân thiện, dễ sử dụng.
  - Xây dựng các chức năng chính như đăng nhập/đăng ký, tải lên ảnh, phát hiện lỗ hổng và quản lý lịch sử.
  - Tích hợp các mô hình machine learning và deep learning vào ứng dụng để phát hiện lỗ hổng trong ảnh JPEG.
- Triển khai và kiểm thử ứng dụng:
  - Triển khai ứng dụng trên nền tảng web hoặc di động.
  - Thực hiện kiểm thử toàn diện, đảm bảo tính bảo mật, tính sẵn sàng, khả năng mở rộng và hiệu suất.
  - Đánh giá hiệu quả của ứng dụng trong việc phát hiện lỗ hổng bảo mật trong ảnh JPEG.

J. Các khó khăn, thách thức hiện tại khi thực hiện:

- Tính phức tạp của các lỗ hổng bảo mật trong ảnh JPEG:
  - Các lỗ hổng như steganography, polyglot, buffer overflow, code injection và remote code execution có cấu trúc và cơ chế hoạt động phức tạp.
  - Việc phát hiện và phân loại chính xác các loại lỗ hổng này trong ảnh JPEG là một thách thức lớn.
- Thiếu dữ liệu mẫu:
  - Việc thu thập đủ lượng dữ liệu mẫu ảnh JPEG chứa các loại lỗ hổng bảo mật là khó khăn.
  - Việc chuẩn bị tập dữ liệu đủ lớn và đa dạng để huấn luyện các mô hình machine learning và deep learning là một thách thức.
- Tối ưu hóa hiệu suất của mô hình:

- Các mô hình machine learning và deep learning đòi hỏi nhiều tài nguyên tính toán và thời gian huấn luyện.
- Việc tìm ra cấu trúc mô hình, siêu tham số tối ưu để đạt hiệu suất tốt trên các tập dữ liệu khác nhau là một thách thức.

- Tích hợp và triển khai ứng dụng:
  - Việc tích hợp các mô hình machine learning và deep learning vào ứng dụng web/di động có thể gặp một số khó khăn về kỹ thuật.
  - Triển khai ứng dụng và đảm bảo tính sẵn sàng, bảo mật và hiệu suất là một thách thức.

3. TỰ ĐÁNH GIÁ MỨC ĐỘ HOÀN THÀNH SO VỚI KẾ HOẠCH THỰC HIỆN:

70%

4. NHẬT KÝ PHÂN CÔNG NHIỆM VỤ:

STT	Công việc	Phân công nhiệm vụ
1	Nghiên cứu bài báo	Cả nhóm
2	Tìm hiểu về mã độc	Nguyễn Đình Luân, Trần Thanh Triều
3	Thực hiện demo mã độc	Nguyễn Đình Luân
4	Huấn luyện mô hình	Trần Đức Trí Dũng
5	Kiểm tra, đánh giá, chỉnh sửa mô hình	Trần Đức Trí Dũng
6	Viết báo cáo	Trần Đức Trí Dũng

BÁO CÁO TỔNG KẾT CHI TIẾT

Phần bên dưới của báo cáo này là tài liệu báo cáo tổng kết - chi tiết của nhóm thực hiện cho đề tài này.

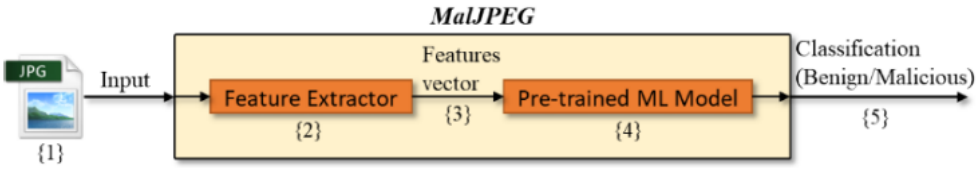
Qui định: Mô tả các bước thực hiện/ Phương pháp thực hiện/Nội dung tìm hiểu (Ảnh chụp màn hình, số liệu thống kê trong bảng biểu, có giải thích)

A. Phương pháp thực hiện

Chương 3: PHƯƠNG PHÁP NGHIÊN CỨU

1. Tổng quan phương pháp

Về phương pháp, tác giả đã thực hiện trích xuất 10 đặc trưng từ cấu trúc tệp nhị phân của các hình ảnh JPEG. Các đặc trưng trên đều là các đặc trưng đơn giản và có thể trích xuất dễ dàng từ các hình ảnh JPEG. Tuy nhiên, chúng đều là những đặc trưng có tính phân biệt cao có thể giúp phân biệt các hình ảnh JPEG lành tính và các hình ảnh bị chèn mã độc cho các mục đích xấu. Sau đó, tác giả sử dụng các đặc trưng đó cho việc huấn luyện các mô hình học máy. Cuối cùng, các mô hình đã được huấn luyện sẽ dùng vào việc dự đoán các hình ảnh JPEG là lành tính hay độc hại.



Hình.1 Kiến trúc của MalJPEG

2. Kiến trúc hệ thống

2.1. Tiền xử lý dữ liệu

Sau khi kiểm tra và nghiên cứu thủ công cấu trúc tệp, tác giả đã nhận ra sự khác biệt về cấu trúc tệp của hình ảnh JPEG lành tính và độc hại. Ví dụ, một số tệp JPEG độc hại chứa dữ liệu (thường là code) sau điểm đánh dấu cuối tệp (EOI). Ngoài ra, tác giả đã phân tích và thống kê sự phân bố tần số và kích thước của điểm đánh dấu JPEG trong cả hình ảnh JPEG độc hại và lành tính, đồng thời xác định các đặc điểm chủ yếu phân biệt giữa hình ảnh JPEG lành tính và độc hại. Các đặc trưng được trích xuất rất đơn giản và hầu hết chúng đều dựa trên sự hiện diện và kích thước của các điểm đánh dấu cụ thể trong cấu trúc tệp hình ảnh JPEG. Ngoài ra, các đặc trưng có thể dễ dàng được trích xuất tĩnh (không thực sự hiển thị hình ảnh) khi phân tích tệp hình ảnh JPEG.

#	Feature Name	Description	Info Gain Rank
1	Marker_EOI_content _after_num	Number of bytes after the EOI (end of file) marker.	0.058
2	Marker_DHT_size_max	Maximal DHT marker size found in the file.	0.025
3	File_size	Image file size in bytes.	0.023
4	Marker_APP1_size_max	Maximal APP1 marker size found in the file.	0.023
5	Marker_COM_size_max	Maximal COM marker size found in the file.	0.017
6	Marker_DHT_num	Number of DHT markers found in the file.	0.016
7	File_markers_num	Total number of markers found in the file.	0.014
8	Marker_DQT_num	Number of DQT markers found in the file.	0.012
9	Marker_DQT_size_max	Maximal DQT marker size found in the file.	0.012
10	Marker_APP12_size_max	Maximal APP12 marker size found in the file.	0.011

Bảng 1: Các đặc trưng của MalJPEG dựa trên Information Gain Rank

Bảng 1 chứa tập hợp các đặc trưng MalJPEG, các đặc trưng được sắp xếp theo thứ hạng sao cho đặc trưng đầu tiên nổi bật nhất. Information Gain xếp hạng một đặc trưng (thuộc tính) bằng cách đo mức giảm entropy của một tập hợp nhất định sau khi chia nó dựa trên một đặc trưng cụ thể. Cụ thể, nó trừ đi entropy có trọng số của từng tập hợp con khỏi entropy ban đầu của toàn bộ tập hợp. Entropy đặc trưng cho sự rối loạn trong một tập hợp các trường hợp tùy ý. Nếu tập hợp hoàn toàn đồng nhất thì entropy bằng 0, nếu tập hợp được chia đều thì nó có entropy bằng 1. Information Gain xếp hạng cao hơn cho các đặc trưng góp phần đáng kể vào việc phân biệt giữa các lớp độc hại và lành tính.

Ở bước tiền xử lý dữ liệu này, module trích xuất sẽ đọc tệp nhị phân của hình ảnh JPEG, sau đó trích xuất 10 đặc trưng được liệt kê ở bảng 3.1. Kết quả trích xuất sẽ được sử dụng cho quá trình huấn luyện và dự đoán kết quả.

2.2 Phương pháp học máy

Mô hình huấn luyện dựa trên 4 loại thuật toán học máy sau:

- *Decision Tree: là một mô hình học máy sử dụng cây quyết định để thực hiện dự đoán. Nó chia dữ liệu thành các nhánh dựa trên các điều kiện liên quan đến các thuộc tính của dữ liệu. Các nút trong cây quyết định đại diện cho các thuộc tính, các nhánh đại diện cho các điều kiện, và các lá đại diện cho các kết quả dự đoán.*
- *Random Forest: là một mô hình học máy dựa trên tập hợp nhiều cây quyết định (decision tree). Mỗi cây quyết định trong Random Forest được xây dựng bằng cách sử dụng một tập con ngẫu nhiên của dữ liệu đào tạo và chỉ sử dụng một tập con ngẫu nhiên của các thuộc tính. Kết quả dự đoán của Random Forest là sự kết hợp của các dự đoán từ các cây quyết định riêng lẻ.*
- *LightGBM (Light Gradient Boosting Machine): là một mô hình học máy dựa trên gradient boosting. Nó sử dụng một thuật toán chia cây hiệu quả và hỗ trợ các đặc trưng như xử lý dữ liệu tương đối lớn, đặc trưng ít và rời rạc. LightGBM thường nhanh hơn và sử dụng ít bộ nhớ hơn so với các mô hình boosting truyền thống khác.*
- *XGBoost (Extreme Gradient Boosting): là một mô hình học máy dựa trên gradient boosting. Nó sử dụng một quy trình tăng cường để xây dựng một mô hình dự đoán bằng cách kết hợp nhiều mô hình yếu (ví dụ như các cây quyết định) thành một mô hình mạnh hơn. XGBoost có hiệu quả tốt, linh hoạt và có thể sử dụng được trên các bộ dữ liệu lớn.*

Các mô hình học máy trên được lựa chọn do chúng cho hiệu suất tốt trên các tập dữ liệu mất cân đối.

B. Chi tiết cài đặt, hiện thực

1. Tài nguyên

- *CPU: AMD Ryzen 5 5600H*
- *RAM: 16GB*
- *Hệ điều hành: Window 10*

2. Môi trường phát triển

- *Trình soạn thảo: Visual Studio Code*
- *Ngôn ngữ: Python*
- *Thư viện sử dụng: numpy, pandas, sklearn, …*

3. Tập dữ liệu

Dữ liệu được thu thập từ hai nguồn:

- *Hình ảnh lành tính: gồm 32868 mẫu được thu thập từ trang Kaggle*
- *Hình ảnh độc hại: gồm 675 mẫu được thu thập từ Virusshare và VirusTotal*

Có thể thấy trong tập dữ liệu trên, hình ảnh lành tính và độc hại có sự chênh lệch lớn khiến dữ liệu bị mất cân đối. Tuy nhiên theo tác giả, việc dữ liệu mất cân đối là để mô hình có thể được học sát với thực tế hơn, từ đó đưa ra dự đoán chính xác hơn. Bởi vì trong thực tế, các hình ảnh có mã độc tuy rất nhiều nhưng chúng vẫn chiếm tỉ lệ rất nhỏ so với các hình ảnh lành tính.

Dữ liệu được chia thành tập train chiếm 80% tập dữ liệu và tập test chiếm 20% tập dữ liệu.

C. Kết quả thực nghiệm

1. *Kết quả thực nghiệm*

	Random forest	Decision tree	LightGBM	XGBoost
TPR	**0.9645**	0.9467	0.9231	0.9467
FPR	**0.0001**	0.0006	**0.0001**	0.0002
IDR	**0.9644**	0.9462	0.9230	0.9465
AUC	**0.9822**	0.9731	0.9615	0.9733

Bảng 2: Kết quả thực nghiệm

Dựa vào bảng 2, có thể thấy các mô hình đều cho kết quả dự đoán khá tốt (đều ở ngưỡng trên 92%). Một điểm đáng chú ý nữa đó chính là FPR của các mô hình đều rất thấp. Đây là một điều quan trọng khi huấn luyện các mô hình cho việc phân loại.

Chỉ số TPR cap và FPR thấp cho thấy rằng các mô hình trên đều có khả năng phân loại rất tốt. Trong số đó thì Random Forest cho kết quả tốt nhất trong tất cả mô hình ở cả 4 chỉ số. LightGBM cho kết quả kém nhất nhưng nhìn chung thì vẫn đạt được kết quả rất cao.

	Random forest	Decision tree	LightGBM	XGBoost
TPR	**0.9704**	0.9527	**0.9704**	0.9586
FPR	0.0004	0.0006	0.0004	**0.0001**
IDR	**0.9701**	0.9521	**0.9701**	0.9585
AUC	0.9850	0.9760	0.9850	0.9792

Bảng 3: Kết quả của bài báo

Tuy rằng quá trình thực nghiệm thu được kết quả khá tốt nhưng nó vẫn chưa đặt được hiệu suất cao như kết quả mà tác giả đã đạt được

2. Ưu điểm và nhược điểm

Ưu điểm:

- *Hiệu suất cao*
- *Khả năng tổng quát tốt*

Nhược điểm:

- *Cần liên tục cập nhật vì trong thực tế, tin tặc sẽ luôn tìm cách để chèn payload theo nhiều cách khác nhau để vượt qua trình phát hiện được tích hợp MalJPEG*
- *Phụ thuộc vào nguồn dữ liệu huấn luyện mô hình*

D. Hướng phát triển

Mở rộng tập dữ liệu và cập nhật liên tục: việc cập nhật các mẫu hình ảnh JPEG độc hại mới, sẽ giúp cải thiện hiệu suất của các mô hình phân loại. Việc cập nhật liên tục tập dữ liệu sẽ đảm bảo các mô hình luôn được huấn luyện trên dữ liệu mới nhất, giúp phát hiện các mẫu tấn công mới.

Tích hợp với các hệ thống bảo mật khác: tích hợp phương pháp MalJPEG vào các hệ thống bảo mật khác như tường lửa, hệ thống phát hiện xâm nhập, phòng chống mã độc, ... Điều này sẽ tạo thành một lớp bảo vệ toàn diện chống lại các mối đe dọa từ hình ảnh độc hại.

Triển khai trong môi trường thực tế: triển khai và đánh giá hiệu quả của phương pháp này trong môi trường thực tế, như các cổng web, hệ thống email, ... để đảm bảo tính hiệu quả và khả năng mở rộng trong thực tế.

Các hướng phát triển đề xuất bên trên sẽ giúp nâng cao hiệu suất, mở rộng tính ứng dụng và đảm bảo tính bền vững của giải pháp phát hiện hình ảnh JPEG độc hại dựa trên phương pháp MalJPEG.

Sinh viên báo cáo các nội dung mà nhóm đã thực hiện, có thể là 1 phần hoặc toàn bộ nội dung của bài báo. Nếu nội dung thực hiện có khác biệt với bài báo (như cấu hình, tập dữ liệu, kết quả,...), sinh viên cần chỉ rõ thêm khác biệt đó và nguyên nhân.

Sinh viên đọc kỹ yêu cầu trình bày bên dưới trang này

YÊU CẦU CHUNG

- Sinh viên tìm hiểu và thực hiện bài tập theo yêu cầu, hướng dẫn.
- Nộp báo cáo kết quả chi tiết những việc (**Report**) bạn đã thực hiện, quan sát thấy và kèm ảnh chụp màn hình kết quả (nếu có); giải thích cho quan sát (nếu có).
- Sinh viên báo cáo kết quả thực hiện và nộp bài.

Báo cáo:

- File .PDF. Tập trung vào nội dung, không mô tả lý thuyết.
- Đặt tên theo định dạng: [Mã lớp]-Project_Final_NhomX_Madetai. (trong đó X và Madetai là mã số thứ tự nhóm và Mã đề tài trong danh sách đăng ký nhóm đồ án).

  Ví dụ: [NT521.N11.ANTT]-Project_Final_Nhom03_CK01.
- Nếu báo cáo có nhiều file, nén tất cả file vào file .ZIP với cùng tên file báo cáo.
- Nộp file báo cáo trên theo thời gian đã thống nhất tại courses.uit.edu.vn.

Đánh giá:

- Hoàn thành tốt yêu cầu được giao.
- Có nội dung mở rộng, ứng dụng.

Bài sao chép, trễ, … sẽ được xử lý tùy mức độ vi phạm.

HẾT