

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/Gw6rLfohxqA>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/Vuoncog/ResearchMethod/blob/master/Slides.pdf>
- Họ và Tên: Nguyễn Văn
Vượng
- MSHV: 250201040
- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Link Github:
<https://github.com/Vuoncog/ResearchMethod>



ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

PHÂN TÍCH MÃ ĐỘC ANDROID BẰNG MÔ HÌNH GENERATIVE
ADVERSARIAL NETWORKS

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ANDROID MALWARE ANALYSIS USING GENERATIVE ADVERSARIAL
NETWORKS

TÓM TẮT (*Tối đa 400 từ*)

Sự gia tăng nhanh chóng và phức tạp của các biến thể mã độc trên hệ điều hành Android đang đặt ra những thách thức nghiêm trọng đối với an ninh thông tin di động. Mặc dù các kỹ thuật học máy đã được ứng dụng rộng rãi trong việc phát hiện xâm nhập, hiệu năng của chúng thường bị giới hạn đáng kể bởi hiện tượng mất cân bằng dữ liệu, khi số lượng mẫu mã độc thực tế thu thập được thấp hơn rất nhiều so với các ứng dụng an toàn. Để khắc phục hạn chế này, nghiên cứu đề xuất một phương pháp tiếp cận mới dựa trên mô hình Generative Adversarial Networks (GAN). Mục tiêu trọng tâm của đề xuất là tái cân bằng phân phối dữ liệu thông qua việc tự động sinh ra các mẫu mã độc nhân tạo mang các đặc trưng sát với thực tế, thay vì chỉ sử dụng các biện pháp lấy mẫu truyền thống. Các mẫu dữ liệu sinh ra sau đó được kết hợp với tập dữ liệu gốc để huấn luyện bộ phân lớp, giúp mô hình học được các đặc trưng tiềm ẩn sâu hơn của mã độc. Kết quả thực nghiệm trên các tập dữ liệu tiêu chuẩn chứng minh rằng phương pháp đề xuất không chỉ cải thiện độ chính xác và chỉ số F1-score so với các thuật toán học máy cơ bản, mà còn nâng cao khả năng tổng quát hóa trong việc nhận diện các biến thể mã độc mới chưa từng xuất hiện trong quá trình huấn luyện.

GIỚI THIỆU (*Tối đa 1 trang A4*)

Sự phổ biến rộng rãi của hệ điều hành Android trên quy mô toàn cầu đã biến nền tảng này trở thành mục tiêu trọng điểm của tội phạm mạng, dẫn đến sự gia tăng theo cấp số nhân cả về số lượng lẫn mức độ tinh vi của các dòng mã độc di động. Các phương pháp phòng thủ truyền thống dựa trên chữ ký đang dần bộc lộ những hạn chế đáng kể, đặc biệt là sự bất lực trước các biến thể mã độc mới hoặc các kỹ thuật che giấu phức tạp như mã độc đa hình. Để đối phó với thách thức này, các kỹ thuật Học máy đã được áp dụng rộng rãi và chứng minh tiềm năng to lớn trong việc tự động hóa quy trình phân tích và nhận diện các hành vi bất thường của ứng dụng.

Hiệu năng của các mô hình học máy trong lĩnh vực an ninh mạng thường bị suy giảm nghiêm trọng bởi vấn đề mất cân bằng dữ liệu. Trong thực tế thu thập, số lượng các ứng dụng an toàn luôn chiếm tỷ lệ áp đảo so với các mẫu mã độc. Sự chênh lệch này khiến các bộ phân lớp khi huấn luyện có xu hướng thiên vị về phía lớp đa số, dẫn đến tỷ lệ bỏ sót các mối đe dọa thực tế cao [1]. Các kỹ thuật cân bằng dữ liệu truyền thống như lấy mẫu thiểu hay lấy mẫu dư đã được sử dụng, chúng vẫn tồn tại những nhược điểm cơ bản: lấy mẫu thiểu làm mất mát thông tin quan trọng, trong khi các thuật toán nội suy tuyến tính thường không phản ánh đúng sự phân bố phức tạp và phi tuyến tính của các đặc trưng mã độc hiện đại.

Nhằm khắc phục những hạn chế nêu trên, nghiên cứu này đề xuất một hướng tiếp cận mới sử dụng Generative Adversarial Networks (GAN) để giải quyết bài toán mất cân bằng dữ liệu [2]. Khác với các kỹ thuật truyền thống, phương pháp được đề xuất tận dụng khả năng của GAN trong việc học phân phối xác suất của dữ liệu gốc để sinh ra

các mẫu mã độc nhân tạo. Các mẫu dữ liệu sinh ra này không chỉ mang các đặc trưng sát với thực tế mà còn có sự đa dạng cao, giúp làm giàu không gian đặc trưng của lớp thiểu số trong tập huấn luyện, từ đó tái cân bằng tỷ lệ phân lớp và nâng cao khả năng tổng quát hóa của mô hình.

Các đóng góp chính của bài báo bao gồm việc xây dựng một khung phát hiện mã độc tích hợp mô hình sinh dữ liệu sâu, đồng thời thực hiện các đánh giá thực nghiệm toàn diện trên các tập dữ liệu tiêu chuẩn. Kết quả nghiên cứu chứng minh rằng việc sử dụng dữ liệu sinh từ phương pháp đề xuất giúp cải thiện đáng kể độ chính xác và chỉ số F1-score so với các phương pháp cân bằng dữ liệu cơ bản. Quan trọng hơn, mô hình còn thể hiện ưu thế vượt trội trong việc phát hiện các biến thể mã độc chưa từng biết đến, khẳng định tính hiệu quả của việc ứng dụng mô hình sinh trong lĩnh vực bảo mật thiết bị di động.

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

Mục tiêu tổng quát của nghiên cứu là xây dựng và tối ưu hóa một khung phát hiện mã độc trên nền tảng Android có khả năng vận hành hiệu quả trong bối cảnh dữ liệu thực tế thường xuyên gặp tình trạng mất cân bằng. Nghiên cứu tập trung khai thác sức mạnh của các kỹ thuật học sâu, cụ thể là mô hình Generator, nhằm giải quyết các hạn chế của những phương pháp phòng thủ hiện hành, từ đó nâng cao năng lực bảo mật cho hệ sinh thái di động trước các mối đe dọa ngày càng gia tăng.

Nghiên cứu đi sâu vào việc thiết kế và ứng dụng mô hình Generative Adversarial Networks (GAN) để học phân phối xác suất phức tạp của các đặc trưng mã độc.

Trọng tâm của quá trình này là tạo ra các mẫu mã độc nhân tạo mang tính đa dạng cao nhưng vẫn giữ được các đặc trưng hành vi sát với thực tế, qua đó khắc phục những nhược điểm về nội suy tuyến tính thường thấy ở các phương pháp cân bằng dữ liệu truyền thống. Việc bổ sung các mẫu dữ liệu từ mô hình Generator này vào tập huấn luyện đóng vai trò then chốt trong việc tái cân bằng tỷ lệ phân phối giữa lớp ứng dụng an toàn và lớp mã độc, giúp mô hình tránh được hiện tượng thiên vị dữ liệu.

Nghiên cứu hướng đến việc huấn luyện các bộ phân lớp dựa trên tập dữ liệu lai ghép để cải thiện tối đa các chỉ số hiệu năng định lượng như Precision, Recall và F1-score. Một mục tiêu quan trọng khác là kiểm chứng và nâng cao khả năng tổng quát hóa của hệ thống trong việc nhận diện các biến thể mã độc chưa từng biết đến, đảm bảo tính ổn định và hiệu quả thực tiễn của giải pháp đề xuất trước các cuộc tấn công mạng ngày càng tinh vi và khó lường.

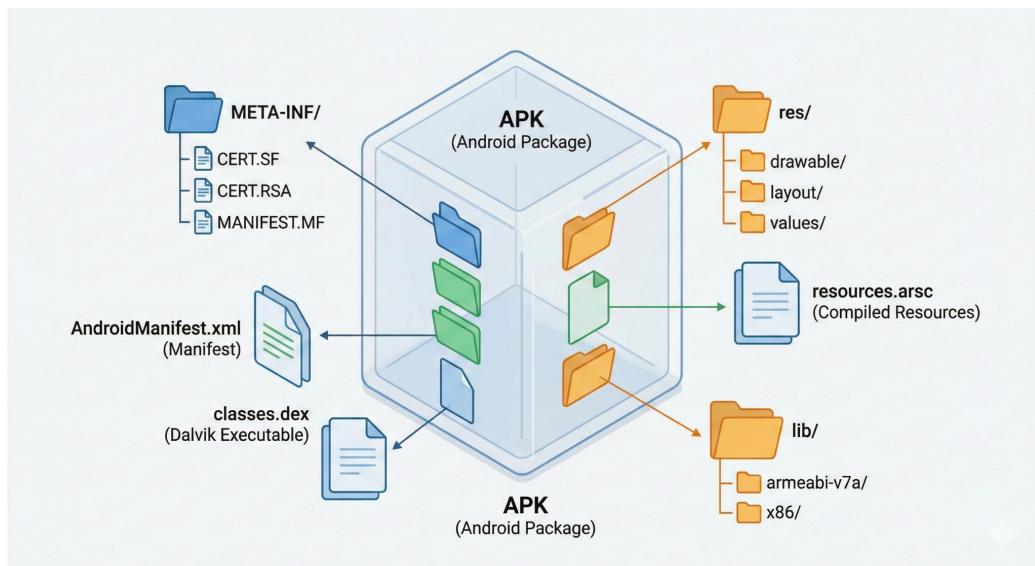
NỘI DUNG VÀ PHƯƠNG PHÁP

Trích xuất đặc trưng bắt đầu bằng việc xử lý tệp tin gói ứng dụng Android (APK). Về mặt kỹ thuật, APK là một định dạng nén lưu trữ chứa toàn bộ mã nguồn và tài nguyên cần thiết để cài đặt ứng dụng. Để xây dựng không gian đặc trưng cho mô hình, thực hiện kỹ thuật phân tích tĩnh thông qua các bước sau:

Công cụ kỹ thuật dịch ngược được sử dụng để giải nén tệp APK, bóc tách hai thành phần cốt lõi là tệp cấu hình *AndroidManifest.xml* và các tệp thực thi *classes.dex*.

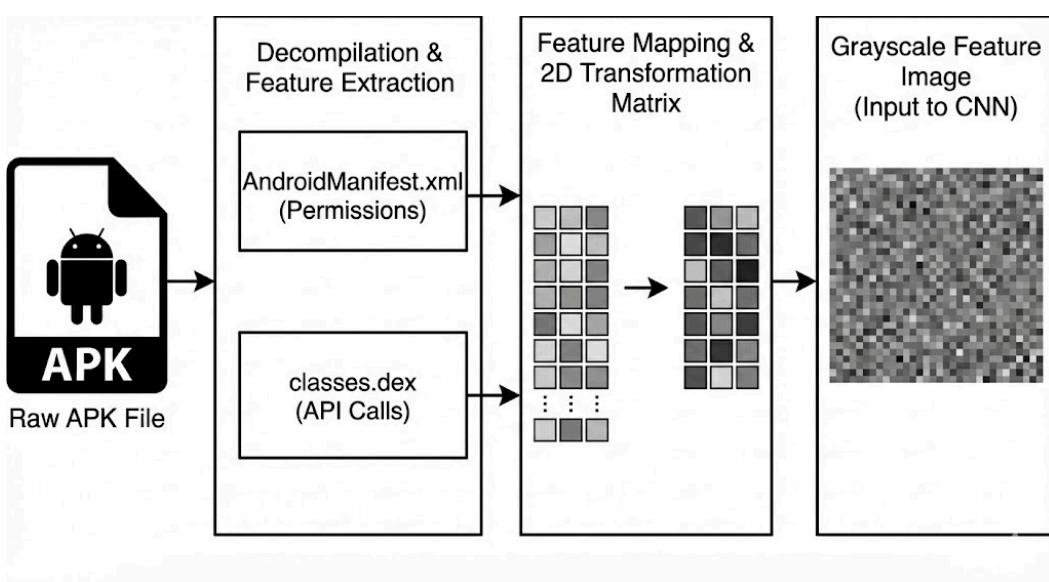
AndroidManifest.xml: Trích xuất danh sách các Quyền hạn mà ứng dụng yêu cầu người dùng cấp phép.

Tệp thực thi *classes.dex*: Nơi chứa mã bytecode Dalvik thực thi logic của ứng dụng.



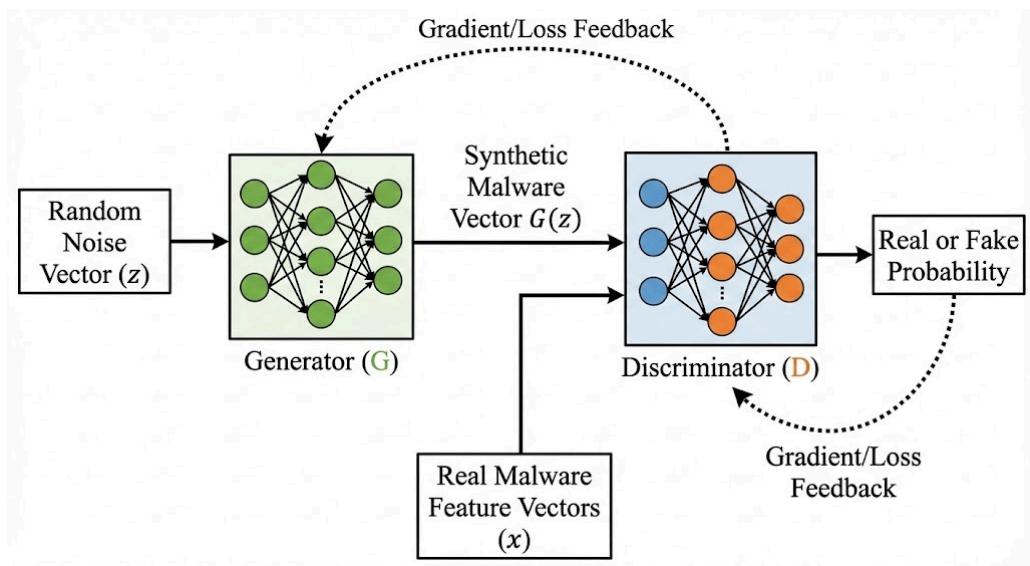
Hình 1. Kiến trúc của tệp tin APK

Tập hợp các đặc trưng thu được từ hai thành phần cốt lõi AndroidManifest.xml và các tệp thực thi classes.dex. Chuẩn hóa các đặc trưng thành một vector nhị phân, trong đó mỗi phần tử đại diện cho trạng thái xuất hiện của một thuộc tính cụ thể trong ứng dụng. Nhằm tận dụng khả năng xử lý không gian của kiến trúc mạng nơ-ron tích chập trong mô hình Generator, vector một chiều này cuối cùng được tái cấu trúc và chuyển đổi thành một ma trận hai chiều dưới dạng hình ảnh xám, đóng vai trò là dữ liệu đầu vào chuẩn hóa cho quá trình huấn luyện [3].



Hình 2. Mô hình trích xuất đặc trưng từ APK sang hình ảnh

Cơ chế hoạt động của mô hình GAN được vận hành dựa trên sự tương tác đối kháng giữa hai mạng nơ-ron thành phần: Mạng Generator (G) và Mạng Discriminator (D). Mạng Generator tiếp nhận một vector nhiễu ngẫu nhiên (z) làm đầu vào và học cách biến đổi nó thành vector đặc trưng mã độc nhân tạo ($G(z)$). Các mẫu giả lập này, cùng với các vector đặc trưng mã độc thực tế (x), được đưa vào Mạng Discriminator để đánh giá và tính toán xác suất thật hoặc giả. Thông qua các luồng phản hồi gradient từ hàm mất mát (Gradient/Loss Feedback), hệ thống thực hiện cập nhật trọng số liên tục cho cả hai phía: Mạng Discriminator tối ưu hóa khả năng nhận diện sự giả mạo, trong khi Mạng Generator học cách tạo ra các cấu trúc dữ liệu ngày càng tinh vi nhằm đánh lừa đối thủ.



Hình 3. Mô hình hoạt động của GAN

KẾT QUẢ MONG ĐỢI

Mục tiêu cốt lõi của nghiên cứu là chứng minh tính hiệu quả vượt trội của việc ứng dụng mô hình GAN trong bài toán phát hiện mã độc Android, đặc biệt là trong bối cảnh dữ liệu thực tế bị mất cân bằng nghiêm trọng. Kết quả dự kiến và quan trọng nhất là khả năng tái thiết lập sự cân bằng phân phối dữ liệu thông qua các mẫu mã độc nhân tạo. Phương pháp đề xuất được kỳ vọng sẽ tạo ra các mẫu mã độc có độ chân thực cao, sở hữu các đặc trưng thống kê tương đồng với mã độc thực tế nhưng vẫn đảm bảo tính đa dạng cần thiết. Việc bổ sung nguồn dữ liệu này vào quá trình huấn luyện sẽ giúp triệt tiêu xu hướng thiên vị của bộ phân lớp đối với các ứng dụng an toàn, từ đó xây dựng được một ranh giới quyết định chính xác và bền vững hơn.

Về mặt định lượng, các thực nghiệm được thiết kế để so sánh hiệu năng của phương pháp đề xuất với các kỹ thuật học máy truyền thống và các phương pháp cân bằng dữ liệu cơ bản như SMOTE. Kết quả mong đợi là sự cải thiện rõ rệt trên các chỉ số đánh giá tiêu chuẩn bao gồm Precision, Recall và F1-Score. Quan trọng hơn cả, nghiên cứu hướng tới việc giảm thiểu tối đa tỷ lệ âm tính giả – tỷ lệ bỏ sót mã độc. Trong lĩnh vực an ninh mạng, việc giảm tỷ lệ này mang ý nghĩa sống còn, và mô hình đề xuất dự kiến sẽ đạt được tỷ lệ phát hiện cao đối với các mẫu mã độc khó nhận biết mà các phương pháp thông thường thường bỏ qua.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1]. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer: SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16: 321-357 (2002)
- [2]. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio: Generative Adversarial Nets. NIPS 2014: 2672-2680
- [3]. Lakshmanan Nataraj, S. Karthikeyan, Gregoire Jacob, B. S. Manjunath: Malware images: visualization and automatic classification. VizSec 2011: 4