ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Phân loại lưu lượng mạng mã hóa sử dụng Class Incremental Learning

LÊ XUÂN NAM
nam.lx162814@sis.hust.edu.vn

Ngành: Kỹ thuật máy tính

Giảng viên hướng dẫn: TS. Tống Văn Vạn _____

Chữ kí GVHD

Khoa: Kỹ thuật máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 08/2024

LỜI CẢM ƠN

Trong suốt những năm tháng học tập tại Đại học Bách khoa Hà Nội, em luôn nhận được sự giúp đỡ và động viên từ gia đình, thầy cô, bạn bè. Là sinh viên của Trường, em đã học được rất nhiều điều từ đạo đức, lối sống, phong cách làm việc đến kiến thức chuyên môn từ những người Thầy, người Cô luôn hết mình giúp đỡ sinh viên, em được tiếp xúc và học hỏi từ rất nhiều sinh viên tài năng của trường. Em xin gửi lời cảm ơn tới bố mẹ, vợ, thầy cô bạn bè đã luôn giúp đỡ em trong quá trình học tập và hoàn thiện đồ án tốt nghiệp. Đặc biệt em xin bày tỏ lòng biết ơn sâu sắc đến thầy Tống Văn Vạn, người đã trực tiếp giúp đỡ, hướng dẫn và luôn sát sao đến em trong quá trình em làm đồ án tốt nghiệp. Em xin chân thành cảm ơn.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Internet ngày càng được sử dụng nhiều trên toàn cầu khiến nó trở thành một phần không thể thiếu trong hoạt động hàng ngày của con người cả trong hoạt động cá nhân lẫn doanh nghiệp. Khi Internet ngày càng phổ biến đồng nghĩa với việc xuất hiện càng nhiều các ứng dụng. Đối với các nhà cung cấp dịch vụ mạng (Vietel, VNPT, v.v), phân loại lưu lượng mạng đóng một vai trò rất quan trọng trong việc quản lí mạng, vì nhà mạng có thể triển khai nhiều giải pháp hướng ứng dụng (truyền video, định tuyến, v.v) hay các giải pháp phát hiện bất thường trong hệ thống mạng. Với sự xuất hiện của các thuật toán học sâu (Deep Learning), rất nhiều nghiên cứu đã tập trung vào các giải pháp phân loại lưu lượng mạng sử dụng học sâu để phát hiện các kiểu ứng dụng đã biết.

Tuy nhiên, mỗi khi ứng dụng mới xuất hiện, nếu chỉ đào tạo lại mô hình với bộ dữ liệu mới thì mô hình không thể hoạt động tốt trên bộ dữ liệu cũ đã được đào tạo, còn việc đào tạo lại mô hình phân loại lưu lượng mạng với toàn bộ dữ liệu cũ và dữ liệu mới là không khả thi vì tốn kém về mặt thời gian và tiền bạc. Trong đề tài này, em đề xuất giải pháp phân loại lưu lượng mạng mã hóa sử dụng Class Incremantel Learning nhằm giải quyết các hạn chế trên. Đề tài này gồm hai đóng góp chính. Thứ nhất, đề tài mô hình phân loại lưu lượng mạng mã hóa sử dụng thuật toán học sâu. Thứ hai, đề tài triển khai giải thuật Class Incremental Learning, iCaRL và đánh giá hiệu năng của các giải thuật này đối với bài toán phân loại lưu lượng mạng.

Sinh viên thực hiện

(Ký và ghi rõ họ tên)

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
CIL	Học tăng dần theo lớp (class incremental learning)
CNN	Mạng nơ ron tích chập
iCaRL	Phân loại tăng dần và học biểu diễn (Incremental Classifier and Representation Learning)

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Internet ngày càng được sử dụng nhiều trên toàn cầu khiến nó trở thành một phần không thể thiếu trong hoạt động hàng ngày của con người cả trong hoạt động cá nhân lẫn doanh nghiệp. Theo báo cáo của Liên minh Viễn thông Quốc tế (ITU), người dùng internet trên toàn cầu đạt 4,1 tỷ vào năm 2019, tăng hơn 53% so với năm 2005. Internet phát triển đồng nghĩa với việc giám sát quản lý và bảo mật hệ thống đang gặp phải nhiều thách thức lớn. Chính vì vậy phân loại lưu lượng mạng rất cần thiết để các nhà cung cấp dịch vụ internet có thể lọc nội dung và xác định các hành vi hoặc ứng dụng có hại đối với người dùng hoặc tổ chức.

Khi internet ngày càng tăng đồng nghĩa với việc xuất hiện càng nhiều các ứng dụng. Mỗi khi ứng dụng mới xuất hiện, nếu chỉ đào tạo lại mô hình với bộ dữ liệu mới thì mô hình không thể hoạt động tốt trên bộ dữ liệu cũ đã được đào tạo, còn việc đào tạo lại mô hình phân loại lưu lượng mạng với toàn bộ dữ liệu cũ và dữ liệu mới là không khả thi vì tốn kém về mặt thời gian và tiền bạc. Chính vì vậy Class Incrementel Learning được kỳ vọng giải quyết vấn đề gia tăng dữ liệu nhanh chóng.

Phân tích các gói tin mã hóa là một nhiệm vụ quan trọng đã được các nhà khoa học nghiên cứu từ lâu. Tuy nhiên các phương pháp truyền thống dựa vào các quy tắc và đặc điểm của gói tin gặp khó khăn trong việc phân loại và phát hiện các đe dọa từ các gói tin do sự hạn chế trong việc phát hiện các đặc điểm lưu lượng mạng được mã hóa mới.

Ngày nay, học sâu là một nhánh của trí tuệ nhân tạo đã cho thấy khả năng vượt trội trong việc phân tích và trích xuất các đặc điểm phức tạp của các bộ dữ liệu mà không cần sự can thiệp thủ công từ con người. Học sâu đã mang lại nhiều tiến bộ trong xử lý ngôn ngữ tự nhiên, thị giác máy tính, nhận dạng giọng nói, ...Việc áp dụng học sâu vào phân tích các gói tin mã hóa đã đạt được những tiến bộ nhất định mặc dù còn nhiều thách thức cần được giải quyết.

1.2 Các giải pháp hiện tại và hạn chế

Các phương pháp phân loại lưu lượng mạng truyền thống, chẳng hạn như dựa trên cổng, kiểm tra gói sâu và dựa trên thống kê đều bị hạn chế trong việc xác định các đặc điểm lưu lượng được mã hóa mới.

Giải pháp sử dụng cổng: Sử dụng cổng (port) để phân loại lưu lượng mạng mã hóa là phương pháp dựa trên giả định rằng các dịch vụ mạng cụ thể thường sử

dụng các cổng tiêu chuẩn như HTTPS sử dụng cổng 443, nó thường được sử dụng để truyền tải dữ liệu nhạy cảm trên web, SSL/TLS: các giao thức bảo mật được sử dụng để mã hóa lưu lượng truy cập mạng thường được sử dụng trên các cổng như 443(HTTPS), 993(IMAPS), ... Tuy nhiên các dịch vụ có thể được cấu hình để sử dụng các cổng không tiêu chuẩn, một cổng có thể được sử dụng cho nhiều loại dịch vụ khác nhau dẫn đến việc phân loại dựa trên cổng gặp khó khăn. Ngoài ra nếu chỉ dựa trên cổng để phân loại lưu lượng mạng sẽ không biết được thông tin chi tiết về nội dung của lưu lượng đó.

Kiểm tra gói sâu: Kiểm tra gói sâu (Deep Packet Inspection) là kỹ thuật phân tích tiêu đề và nội dung bên trong gói tin, nó có khả năng kiểm tra nội dung gói tin trong thời gian thực, giúp xác định chính xác loại ứng dụng hoặc giao thức mà khi dùng cổng không thể phát hiện. DPI thường được sử dụng trong các hệ thống tường lửa, hệ thống phát hiện xâm nhập hoặc trong các giải pháp quản lý băng thông hoặc giám sát an ninh mạng. So với phương pháp sử dụng cổng, DPI có độ chính xác cao hơn ngoài ra nó có thể giúp phát hiện các cuộc tấn công mạng hoặc phần mềm độc hại bằng cách phân tích sâu vào nội dung tin. Tuy nhiên ngoài sử dụng chi phí cao do yêu cầu tài nguyên tính toán mạnh mẽ, DPI không thể giải mã nội dung mã hóa mà không có khóa mã hóa.

Dựa trên thống kê: Phân loại lưu lượng mạng dựa vào thống kê là phương pháp phân tích các đặc trưng thống kê của lưu lượng mạng, chẳng hạn như kích thước gói tin, tần suất, độ trễ và các đặc trưng khác sau đó sử dụng các mô hình học máy để xác định loại ứng dụng hoặc giao thức đang được sử dụng. Mặc dù có nhiều ưu điểm hơn phương pháp sử dụng cổng và phương pháp kiểm tra gói sâu tuy nhiên phương pháp dựa trên thống kê vẫn tồn tại các hạn chế: cần dữ liệu huấn luyện để xây dựng mô hình học máy chính xác, khó khăn khi lưu lượng biến đổi và yêu cầu tài nguyên tính toán lớn.

Các phương pháp phân loại dựa trên học sâu xem xét các tính năng dựa trên gói được khám phá để giải quyết các hạn chế của các phương pháp truyền thống khi phân loại lưu lượng mạng mã hóa.

1.3 Mục tiêu và định hướng giải pháp

1.3.1 Mục tiêu

Với những vấn đề và bối cảnh hiện tại được nêu ra ở các mục trên, mục tiêu đặt ra của đồ án là đề xuất, triển khai và đánh giá được mô hình phân loại lưu lượng mạng mã hóa. Thứ nhất mô hình đề xuất phải phân loại chính xác được các lưu lượng mạng thuộc các lớp đã được đào tạo. Thứ hai, với những lớp lưu lượng mạng mới chưa gặp bao giờ, mô hình đề xuất phải có khả năng phân loại nó mà không