RỦI RO VI PHẠM RIÊNG TƯ DỮ LIỆU TRONG HỌC SÂU

Trần Trương Tuấn Phát[1,2], Đặng Trần Khánh[1*]

[1]*Trường Đại học Công nghiệp Thực phẩm TP.HCM*

[2]*Trường Đại học Bách Khoa - ĐHQG TP.HCM*

*Email: *khanh@hufi.edu.vn*

Ngày nhận bài: 10/06/2022; Ngày chấp nhận đăng: 13/7/2022

TÓM TẮT

Nhờ vào sự vượt trội về khả năng dự đoán của các phương pháp học sâu, ứng dụng trí tuệ nhân tạo nói chung và học sâu nói riêng đã giải quyết được nhiều vấn đề thực tế và ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực, ngành nghề. Tuy nhiên, mặc dù các mô hình học máy dựa trên học sâu mạnh trong nhiều tác vụ và bài toán nhưng vẫn chưa hoàn thiện. Điển hình là các mô hình này rất dễ bị tấn công và vi phạm các tiêu chí về an toàn thông tin. Trong đó, rủi ro vi phạm về riêng tư dữ liệu là một vấn đề nhức nhối vì nó không chỉ ảnh hưởng đến hệ thống, người cung cấp dịch vụ, người dùng mà còn cả đến sự an toàn, lòng tin của con người vào việc sử dụng công nghệ và các vấn đề xã hội, pháp lý. Trong bài báo này, chúng tôi tổng hợp và phân tích các công trình liên quan đến vấn đề vi phạm riêng tư dữ liệu trong học sâu trong những năm gần đây, từ đó đề xuất mô hình và đưa ra những cảnh báo khi xây dựng các mô hình học sâu.

Từ khóa: Riêng tư dữ liệu, học sâu, dữ liệu lớn, bảo mật dữ liệu, điều khiển truy xuất.

1. MỞ ĐẦU

Nhờ sự phát triển của các công nghệ phần cứng và dữ liệu lớn, các mô hình học sâu dựa vào mạng nơron lần lượt vượt qua các phương pháp học máy trước đó trong hàng loạt các lĩnh vực, đặc biệt là trong thị giác máy tính [1-3] và xử lý ngôn ngữ tự nhiên [4-6]. Tuy vậy, gần đây các mô hình xây dựng dựa trên phương pháp học sâu bị khai thác và chứng minh có khả năng không an toàn trước nhiều rủi ro và các cuộc tấn công: tấn công trốn tránh (*adversarial*/*evasion attack)* [7-9], tấn công cửa sau (*backdoor attack*) [10, 11] làm vi phạm tính toàn vẹn (*integrity*) của an toàn thông tin; bên cạnh đó, tấn công đầu độc dữ liệu (*data poisoning attack*) [12, 13] làm vi phạm tính toàn vẹn (*integrity*) và sẵn sàng (*availability*), tấn công trích xuất mô hình (*model extraction attack*) [14, 15], tấn công đảo ngược mô hình (*model inversion attack*) [16] làm vi phạm tính bảo mật (*confidentiality*); tấn công riêng tư dữ liệu (*privacy attack*) [17-22] làm vi phạm tính bảo mật (*confidentiality*) và tính riêng tư dữ liệu (*data privacy*),v.v. Việc liên tiếp bị khai thác và tìm ra những điểm yếu mới khiến cho tính an toàn khi áp dụng rộng rãi công nghệ ứng dụng học sâu là một câu hỏi lớn.

Riêng tư dữ liệu hay riêng tư người dùng có rất nhiều khía cạnh định nghĩa, tuy nhiên, ở đây là có thể hiểu là các công ty, tổ chức cung cấp dịch vụ trên không gian mạng đó phải có nghĩa vụ bảo vệ quyền riêng tư đó cho người dùng. Đó có thể là những thoả thuận giữa khách hàng, người dùng với công ty, tổ chức đó thông qua cách điều khoản, chính sách riêng tư (*privacy policies, privacy regulations*). Ví dụ, như là bảo vệ về việc chia sẻ dữ liệu cho một công ty, tổ chức khác nữa hoặc về mục đích sử dụng của dữ liệu,…

Xét nguyên nhân, dữ liệu riêng tư có thể bị lộ theo hai cách trực tiếp hoặc gián tiếp. Nguyên nhân trực tiếp đến từ chính những công nghệ, dịch vụ, kênh trao đổi thông tin, nơi lưu

trữ không bảo vệ được sự riêng tư cho người dùng. Cụ thể hơn, nó có thể đến từ các công ty/tổ chức công nghệ cung cấp dịch vụ không hoàn thiện về tính bảo vệ riêng tư, hay đến từ chính bản thân của người dùng sử dụng sai cách vô tình công khai dữ liệu riêng tư của mình. Một trong những những ví dụ tiêu biểu hiện nay là mạng xã hội Facebook khi liên tục bị cáo buộc và phạt khi vi phạm các quy định bảo vệ quyền riêng tư vào năm 2018-2019. Ngoài việc vi phạm riêng tư thường đến từ sự không cẩn trọng, không quan tâm của các công ty/tổ chức cung cấp dịch vụ lẫn người dùng, nhưng nó còn có thể đến từ việc cố tình khai thác trái phép, điển hình là các tổ chức bán dữ liệu hay sử dụng tài nguyên đó để thực hiện mục đích trái phép, không đúng cam kết với khách hàng. Ví dụ điển hình là việc lộ thông tin của trang web tìm kiếm nổi tiếng đầu thế kỷ XX - AOL (2004) - gần 100 triệu người bị vi phạm quyền riêng tư trong không gian mạng trong vụ việc này.

Tuy nhiên, kể cả khi được xem xét cẩn thận về các quá trình chia sẻ, thu thập, sử dụng, lưu trữ thì quyền riêng tư của chủ thể dữ liệu vẫn có thể bị vi phạm do những cá nhân/tổ chức có hiểu biết công nghệ cố gắng khai thác thông tin riêng tư. Những nguyên nhân này có thể xem là gián tiếp vì phải qua quá trình nghiên cứu, tìm hiểu, phân tích để khai thác thông tin riêng tư [23, 24]. Các công nghệ trí tuệ nhân tạo, học sâu đang dần len lỏi vào hầu hết các lĩnh vực trong cuộc sống. Quá trình học của các thuật toán, mô hình học sâu đã giúp chúng ta đưa ra những quyết định, những dự đoán cho một dữ liệu đầu vào mới sau quá trình huấn luyện trên nhiều dữ liệu đã biết trước đó. Tuy nhiên, chính nhờ khả năng như vậy trí tuệ nhân tạo, học sâu có thể trở thành công cụ để khai thác quyền riêng tư. Ví dụ, bằng việc cho học sâu học trên những dữ liệu nhạy cảm, ta có thể làm cho nó có khả năng đưa ra tiên đoán khá chính xác về dữ liệu riêng tư của một người khác. Bài viết này tập trung vào vấn đề rủi ro vi phạm tính riêng tư của các mô hình học sâu nên các nguyên nhân gây vi phạm được khảo sát là nguyên nhân gián tiếp.

Trong phần còn lại, chúng tôi sẽ lần lượt trình bày: Giới thiệu về học sâu và các khái niệm được sử dụng trong trong bài – Khái niệm học sâu (Phần 2), các tấn công vào các mô hình học sâu làm vi phạm riêng tư, đặc biệt chúng tôi tập trung phân tích tấn công suy luận thành viên - rủi ro vi phạm riêng tư của các mô hình học sâu (Phần 3), cuối cùng là phần tổng kết và đưa ra những nhận xét, các hướng phát triển của lĩnh vực nghiên cứu (Phần 4).

2. KHÁI NIỆM HỌC SÂU

Mặc dù lĩnh vực trí tuệ nhân tạo (*Artificial Intelligence - AI*) đã có lịch sử khá lâu nhưng AI thực sự bùng nổ và hồi sinh từ 1987–1993 là nhờ vào các mô hình học dựa trên học sâu [25]. Học sâu là một tập con của các phương pháp học máy dựa vào các mạng nơ-ron nhân tạo, được lấy cảm hứng từ cách tổ chức thần kinh của con người.

Về mặt cấu trúc, một mô hình học sâu gồm nhiều lớp (*layers*) phức tạp, biến đổi phi tuyến, còn được gọi là hàm kích hoạt (*activation functions*), tiêu biểu là sigmoid và rectified linear units (ReLUs) và học được cách biểu diễn (representations) và đưa ra dự đoán. Bên cạnh đó, để huấn luận cấu trúc này có thể học được ta cần định nghĩa một hàm mất mát (*loss function*) để tối thiểu hoá đầu ra của cấu trúc với dữ liệu thực tế. Giả sử dữ liệu cần học $\{x_1, x_2, \ldots, x_n\}$, ta cần tìm tập tham số của mô hình θ so cho $\measuredangle(\theta)=\frac{1}{n}\measuredangle(\theta, x_i)$ đạt giá trị nhỏ nhất có thể. Thuật toán xuống đồi (*gradient descent*) được dùng để tìm điểm cực tiểu có khả năng đạt giá trị nhỏ nhất này. Vì các mô hình học sâu thường được huấn luyện trên một tập dữ liệu rất lớn nên ta thường không thể bỏ tất cả các dữ liệu đầu vào để học trong một lần mà phải huấn luyện theo từng lô (*batch*) và dùng mini-batch hay "stochastic gradient descent" để tìm các điểm cực tiểu địa phương. Qua quá trình phát triển các mạng nơ-ron ngày càng sâu với nhiều lớp và cấu trúc khác nhau như tích chập (*convolution*), hồi quy (*recurrent*), chuẩn hoá bó (*batch norm*),… với số lượng tham số mô hình ngày càng khổng lồ.

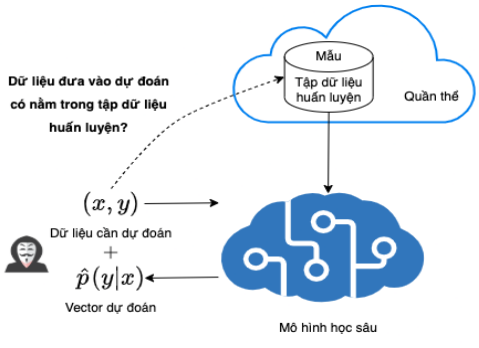
3. RỦI RO VI PHẠM RIÊNG TƯ DỮ LIỆU TRONG HỌC SÂU

Thế nào là một mô hình học sâu vi phạm tính riêng tư? Theo Dalenius [26] thì từ đầu ra dự đoán (thường là một vector dự đoán) thì kẻ tấn công có thể suy luận thêm những thông tin khác về tập dữ liệu huấn luyện và thông số mô hình huấn luyện thì mô hình học sâu đó có khả năng làm lộ tính riêng tư. Cụ thể hơn đối mô hình học trên những dữ liệu nhạy cảm thì từ đầu ra của mô hình, kẻ tấn công có thể khai thác trực tiếp những thông tin sau: biết được một điểm dữ liệu/một cá nhân thuộc tập dữ liệu huấn luyện, xây dựng ngược lại được tập dữ liệu huấn luyện hay tìm được các đặc điểm, tính chất nhạy cảm của tập dữ liệu huấn luyện hoặc một hay một số lớp đại diện trong tập dữ liệu huấn luyện. Tuy nhiên để ngăn chặn hoàn toàn điều này là rất khó đạt được vì nhiều nguyên nhân, đặc biệt nếu kẻ tấn công có kiến thức nền về tập dữ liệu huấn luyện hoặc quần thể nơi tập dữ liệu đó được lấy mẫu. Do đó, khi xét về tính riêng tư nghĩa ta xét về tính riêng tư của một điểm dữ liệu/một cá nhân trong một tập dữ liệu được sử dụng [27, 28]. Tính riêng tư cá nhân hay điểm dữ liệu này có thể hiểu là với một cá nhân/điểm dữ liệu bất kỳ trong tập dữ liệu huấn luyện hay rộng hơn là trong quần thể lấy mẫu thì từ kết quả đầu ra của mô hình ta không thể suy luận thêm thông tin gì từ điểm dữ liệu/cá nhân này.

Trong những năm gần đây, các công trình nghiên cứu đã chỉ ra các công nghệ xây dựng dựa trên mạng nơ-ron học sâu có rủi ro về vi phạm dữ liệu riêng tư của người cung cấp dữ liệu. Rủi ro này thể hiện qua các cuộc tấn công như: trích xuất mô hình [14, 15], suy luận thuộc tính [16], suy luận tính chất [20, 30] và suy luận thành viên [17-19]. Trong đó tấn công suy luận thành viên được xem là dấu hiệu của việc lộ thông tin cá nhân.

3.1. Tấn công suy luận thành viên

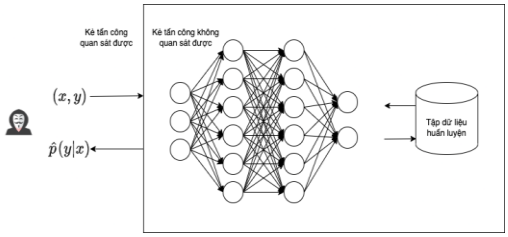
Tấn công suy luận thành viên (*Hình 1*) tìm cách suy luận một cá nhân, một điểm dữ liệu nào đó có thuộc tập dữ liệu được sử dụng để xây dựng mô hình hay không. Trong nhiều công trình, độ chính xác trong tấn công suy luận thành viên được dùng làm thước đo của việc một mô hình học sâu có rủi ro lộ tính riêng tư [27, 28] bởi vì, đối với mô hình được huấn luyện trên tập dữ liệu nhạy cảm, việc kẻ tấn công suy luận được một cá nhân/điểm dữ liệu thuộc tập dữ liệu đó thì hiển nhiên riêng tư của chủ dữ liệu bị vi phạm. Ví dụ: mô hình huấn luyện trên tập dữ liệu là những bệnh nhân lao thì hiển nhiên nếu suy luận được một người nào đó thuộc tập dữ liệu huấn luyện thì người đó bị mắc bệnh lao. Từ việc "suy luận được thành viên" thì kẻ tấn công hoàn toàn có thể tiến hành thêm những suy luận vi phạm dữ liệu cá nhân khác như: ghi nhận lại dữ liệu hồ sơ bằng cách kết hợp các nguồn khác nhau, rồi tổng hợp suy luận các tính chất, đặc tính nhạy cảm khác từ các nguồn. Tấn công suy luận thành viên cũng đã được nghiên cứu từ lâu trước đó cho việc bảo vệ tính riêng tư dữ liệu cho địa điểm [31, 32], gen [33, 34], các mô hình học máy truyền thống [35],…



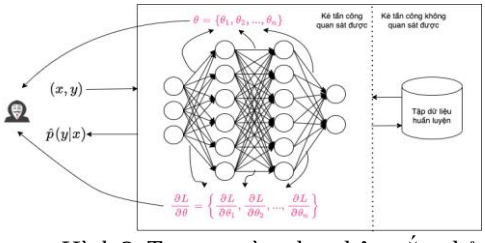
Hình 1. Tấn công suy luận thành viên

Xét về khả năng quan sát của kẻ tấn công thì kẻ tấn công không thể quan sát, không biết gì về mô hình chỉ có thể gửi dữ liệu và nhận về kết quả dự đoán, thậm chí số lượng gửi dữ liệu để đoán có thể có giới hạn [17]. Với ngữ cảnh này có thể phân loại là tấn công hộp đen (*black-box*) (*Hình 2*). Ví dụ về ngữ cảnh này là các dịch vụ học máy (*machine learning as a service*) của Google (Google Prediction API), Amazon Machine Learning (Amazon ML), Microsoft Azure Machine Learning (Azure ML), BigML,… Với các dịch vụ này chúng ta có thể tải lên tập dữ liệu huấn luyện, chọn mô hình và các dịch vụ này sẽ huấn luyện thành mô hình cho chúng ta. Chúng ta chỉ cần dùng nó như một dịch vụ và không biết các tham số bên trong, thậm chí là kiến trúc mô hình. Ngược lại, nếu kẻ tấn công có thể quan sát được tham số mô hình khi gửi dữ liệu dự đoán thì ngữ cảnh này kẻ tấn công có nhiều thông tin hơn, có thể xem đây là ngữ cảnh hộp trắng (*white-box*) [19].

3.1.1. Tấn công hộp đen

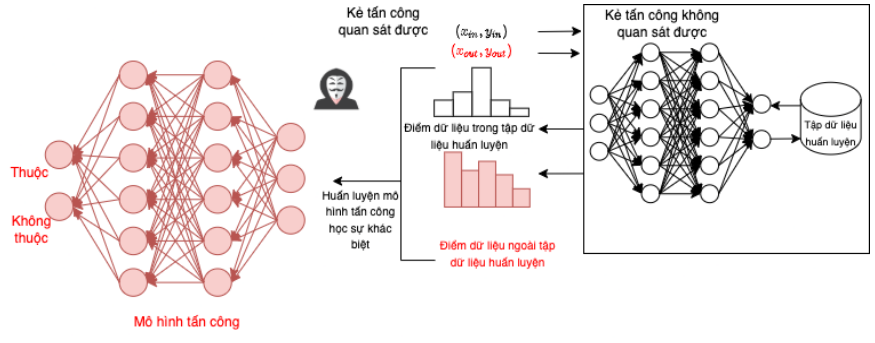


Hình 2. Trong trường hợp hộp đen kẻ tấn công sử dụng dữ liệu đầu vào và kết quả dự đoán trả ra từ một dịch vụ AI hộp đen để thực hiện suy luận thành viên



Hình 3. Trong trường hợp hộp trắng kẻ tấn công có khả năng quan sát được các thông số học và gradient trong quá trình dự đoán. Do đó kẻ này có thể sử dụng dữ liệu đầu vào, kết quả dự đoán, thông số học và gradient từng lớp để suy luận thành viên

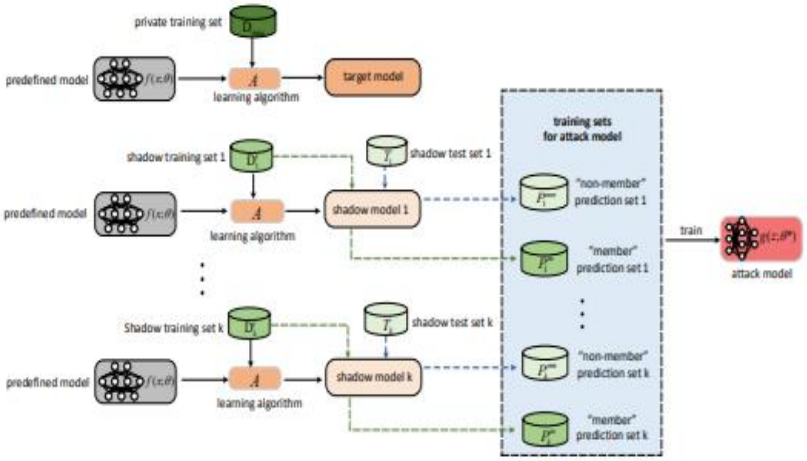
Trong trường hợp tấn công hộp đen, kẻ tấn công bị giới hạn kiến thức, quyền tiếp cận và mô hình nên phải cố gắng để có thể suy luận thông tin thành viên của một cá nhân/điểm dữ liệu đầu vào có thuộc tập dữ liệu huấn luyện của mô hình nạn nhân hay mô hình đối tượng (*victim* or *target model*) hay không. Ở đây, kẻ tấn công không có kiến thức về mô hình nhưng vẫn có thể có kiến thức về tập dữ liệu, cụ thể về phân bố của tập dữ liệu. Trong [17], các tác giả đã đề xuất một phương pháp có thể dùng để thực hiện tấn công suy luận tổng quát. Phương pháp này dựa trên quan sát "hành vi" mô hình có xu hướng trả về kết quả dự đoán của một dữ liệu đầu vào không thuộc tập dữ liệu thuộc tập huấn luyện khác với một điểm dữ liệu thuộc tập dữ liệu huấn luyện. Điều này là có khả năng xảy ra khá cao vì với những những dữ liệu thuộc vào tập dữ liệu huấn luyện thì kết quả trả ra của mô hình chắc chắn hơn (ví dụ, điểm tin cậy của một lớp là cao hơn hẳn so với các lớp còn lại) với những dữ liệu đầu vào lạ không thuộc tập dữ liệu huấn luyện.



Hình 4. Xây dựng mô hình tấn công học sự khác biệt giữa vector dự đoán của điểm dữ liệu trong tập dữ liệu huấn luyện và điểm dữ liệu trong tập dữ liệu

Vấn đề là làm sao để xây dựng tập dữ liệu bao gồm dữ liệu trong và dữ liệu ngoài tập dữ liệu huấn luyện? Có ba cách cơ bản: Thứ nhất là dựa vào chính mô hình. Kẻ tấn công gửi dữ liệu dự đoán và dò xem dữ liệu nào nào có điểm tin cậy thuộc một lớp cao hơn hẳn các lớp khác với một ngưỡng cố định. Nếu tìm được một không gian các điểm dữ liệu của một lớp như vậy, từ đó có thể xây dựng một tập dữ liệu để huấn luyện mô hình tấn công từ đây; Cách thứ hai, nếu kẻ tấn công có kiến thức về phân bố dữ liệu của tập dữ liệu huấn luyện thì có thể lấy mẫu tuân theo phân bố này; Cuối cùng là nếu có thể tiếp cận một số dữ liệu mẫu thì kẻ này có thể xây dựng những dữ liệu nhiễu (*noisy data*) từ những dữ liệu đã biết này. Tập dữ liệu mà kẻ tấn công dùng mọi cách trong khả năng có được để huấn luyện mô hình tấn công gọi là tập dữ liệu mờ (*shadow dataset*).

Sau khi đã có tập dữ liệu mờ, kẻ tấn công chia ra thành n tập nhỏ, không giao nhau (giả sử như vậy nếu có đủ dữ liệu) để huấn luyện ra mô hình mờ (s*hadow model*). Về cơ bản, cần huấn luyện các mô hình mờ này sao cho nó càng giống với mô hình đối tượng nhất có thể. Nếu kẻ tấn công có kiến thức về mô hình (mô hình học, kiến trúc) thì việc huấn luyện này đơn giản, tuy nhiên trong ngữ cảnh hộp đen thì kẻ tấn công không có thông tin mô hình. Trong một số trường hợp, ví dụ dịch vụ học máy thì kẻ tấn công có thể tải lên tập dữ liệu và yêu cầu huấn luyện ra mô hình mờ trên tập dữ liệu đó giống như mô hình đối tượng mà không cần quan tâm là mô hình gì, kiến trúc và cách huấn luyện ra sao. Về cơ bản huấn luyện được càng nhiều mô hình thì độ chính xác của cuộc tấn công càng cao. Do huấn luyện trên tập dữ liệu mờ nên chúng ta biết được dữ liệu nào thuộc và không thuộc dữ liệu huấn luyện của từng mô hình mờ. Từ đây có thể xây dựng tập dữ liệu bao gồm kết đoán của các mô hình mờ với nhãn "in"/"out" để huấn luyện mô hình tấn công.



Hình 5. Tấn công suy luận thành viên bằng cách xây dựng một mạng nơ-ron dự đoán xem điểm dữ liệu có thuộc vào tập dữ liệu của mô hình đối tượng không [43]

Phương pháp xây tấn công suy luận thành viên bằng cách xây dựng mô hình mờ được sử dụng rộng rãi nhất và được khảo sát trong nhiều nhất công trình về rủi ro vi phạm tính riêng tư của học sâu [36]. Bên cạnh đó, thay vì xây dựng mô hình tấn công thì một số công trình cũng đề xuất cách suy luận mới dựa vào: kết quả xuất ra của hàm mất mát [37] - nếu kết quả của hàm mất mát nhỏ hơn một ngưỡng cố định thì khả năng điểm dữ liệu đó thuộc vào tập dữ liệu huấn luyện; điểm tin cậy [22] - nếu điểm tin cậy lớn hơn một ngưỡng cố định thì khả năng điểm dữ liệu thuộc vào tập dữ liệu huấn luyện; entropy của vector dự đoán [22, 38]- entropy của vector dự đoán càng nhỏ thì khả năng điểm dữ liệu đó thuộc vào tập dữ liệu huấn luyện. Các phương pháp này đều khá thành công khi suy luận ra thành viên với độ chính xác cao.

Công trình [17] lần đầu tiên đề xuất được một phương pháp tấn công suy luận thành viên tổng quát vào các mô hình học sâu. Các tác giả trong [22] đã cải thiện phương pháp này bằng cách chỉ dùng một mô hình mờ duy nhất và dữ liệu xây dựng không nhất thiết cùng phân phối với mô hình nạn nhân. Các phương pháp tấn công thường quan sát các vector dự đoán là các điểm tin cậy hay các điểm tin cậy được che đậy (*masking*) - làm tròn hoặc tìm ra top-k; các công trình [18, 39], chứng minh rằng với vector dự đoán là các nhãn duy nhất thì tấn công suy luận vẫn có thể thành công. Trong [18], các tác giả dùng phương pháp tăng cường dữ liệu (*data augmentation*) để xây dựng thêm dữ liệu và nhận thấy rằng những dữ liệu này có nhãn bền vững hơn nếu thuộc trong tập dữ liệu huấn luyện. Tương tự, [39] đã xây dựng các ví dụ nghịch cảnh (*adversarial examples*) và nhận thấy rằng nếu dữ liệu không thuộc tập dữ liệu huấn luyện thì dễ dàng xây dựng hơn dựa vào độ lớn của xáo trộn nhiễu (*perturbation*). Qua đó, chúng ta thấy được rằng rủi ro vi phạm tính riêng tư, cụ thể là rủi ro bị suy luận thành viên của các mô hình học sâu hiện hữu.

3.1.2. Tấn công hộp trắng

Tấn công hộp trắng (*Hình 3*) được xem là mở rộng của tấn công hộp đen vì kẻ tấn công hoàn toàn có thể dùng phương pháp tấn công hộp đen như trên để suy luận thành viên. Tuy nhiên, ở đây kẻ tấn công có nhiều thông tin hơn tấn công hộp đen, cụ thể là hàm mất mát, thông số mô hình và gradient của hàm loss tương ứng với và có thể dùng các phương pháp phức tạp hơn để tăng độ chính xác của suy luận thành viên. Tuy nhiên, không phải tất cả những thông tin biết thêm điều thực sự có giá trị nên công trình [19] đã đề xuất một phương pháp dựa vào khai thác phương pháp học - stochastic gradient descent (SGD) bởi vì nếu một điểm dữ liệu thuộc vào tập dữ liệu huấn luyện thì nó có xu hướng làm cho gradient hướng về *0*.

Phương pháp tấn công dựa vào SGD cũng thành công trong việc tấn công suy luận thành viên của phương pháp học liên kết (*federated learning*). Học liên kết [40, 41] là phương pháp học phân tán (*distributed learning*) mà nhiều máy/thiết bị ngoại vi (*edge devices*) cộng tác nhau để cùng xây dựng mô hình học máy toàn cục. Các máy/thiết bị ngoại vi trước hết tải mô hình cơ sở từ máy chủ/máy tổng hợp rồi xây dựng một hình cục bộ (l*ocal model*) của riêng mình, liên tục gửi thông số lên trên để tổng hợp thành mô hình toàn cục theo nhiều cách rồi gửi lại mô hình cập nhật cho các máy/thiết bị ngoại vi để tiếp tục huấn luyện. Phương pháp học liên kết có thể là một phương pháp đầy triển vọng về học bảo vệ tính riêng tư (*privacy-preserving learning*) vì cách huấn luyện theo hướng phân tán do đó dữ liệu mang tính riêng tư, nhạy cảm có thể huấn luyện nội bộ trong từng máy/thiết bị và chỉ chuyển những gì học được, ví dụ các tham số, điểm tin cậy hay nhãn lên máy chủ để tổng hợp. Về mặt rủi ro vi phạm tính riêng tư, trong cài đặt học liên kết, có thể coi là một tấn công hộp trắng khi kẻ tấn công có thể là một máy/thiết bị hay thậm chí là máy tổng hợp tham gia vào quá trình huấn luyện. Vì là một thiết bị trong hệ thống nên thay vì thụ động (*passive attack*) khai thác thông tin bằng cách quan sát, kẻ tấn công có thể chủ động (*active attack*) cập nhật lên mô hình toàn cục theo hướng có lợi cho việc khai thác thông tin cho mình, còn gọi là tấn công người trong cuộc (*insider attack*), suy luận một điểm dữ liệu được huấn luyện nội bộ ở thiết bị nào [42]. Thực tế, huấn luyện mô hình bằng phương pháp học liên kết và bảo vệ được tính riêng tư là một bài toán khó vì dữ liệu được phân tán cộng với các kỹ thuật bảo vệ tính riêng tư càng làm ảnh hưởng độ chính xác của dự đoán hơn nữa [43, 44].

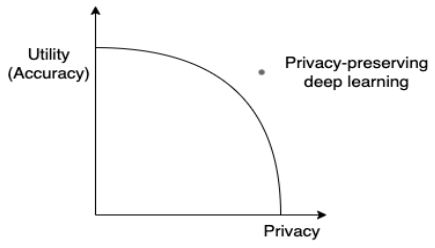
Ngoài ra, hầu hết các công trình các công trình mà chúng tôi đề cập ở trên đều thực hiện trên các mô hình phân loại ảnh, thuộc lĩnh vực thị giác máy tính. Tuy nhiên nhiều công trình với những cách tấn công suy luận thành viên tương tự cũng thành công trong việc tấn công suy luận thành viên từ mô hình sinh (GANs, VAEs) [45, 46] và các mô hình ngôn ngữ [47, 48], các loại dữ liệu khác [49, 50]. Các công trình giúp ta thấy được rằng rủi ro vi phạm riêng

tư của các mạng học sâu là có thật mặc cho kiến trúc, mô hình và cách cài đặt. Nếu chúng ta không kiểm tra, khảo sát kỹ càng thì rủi ro cho vi phạm riêng tư người dùng có thể xảy ra.

3.1.3. Thảo luận

Vậy lý do các mô hình học sâu có rủi ro vi phạm tính riêng tư là do đâu? Các công trình [17, 19, 37] đã chỉ ra rằng quá khớp (*overfitting*) là một nguyên nhân gây ra rủi ro vi phạm tính riêng tư của các mô hình học sâu. Quá khớp là hiện tượng mô hình học tổng quát hoá tốt trên tập dữ liệu huấn luyện nhưng lại dự đoán kém trên tập kiểm tra. Các công trình này đã nghiên cứu mối quan hệ giữa quá khớp, đặc trưng bởi khoảng cách tổng quát hóa (*generalization gap*) bằng hiệu giữa độ chính xác dự đoán trên tập dữ huấn luyện và độ chính xác trên tập kiểm tra ($g = a_{training} - a_{test}$), với độ chính xác của tấn công suy luận thành viên, cụ thể hơn là xác suất thành công suy luận thành viên $> \frac{1}{2}$ với mô hình quá khớp. Khi một mô hình càng quá khớp thì đối với điểm dữ liệu thuộc vào tập dữ liệu huấn luyện, mô hình sẽ dự đoán với điểm tin cậy rất cao, ngược lại với điểm dữ liệu không thuộc tập dữ liệu huấn luyện, mô hình sẽ xuất ra kết quả không chắc chắn và dễ phân biệt. Từ đây ta thấy được rằng để giúp mô hình học sâu giảm bớt rủi ro vi phạm tính riêng tư thì ta phải huấn luyện được một mô hình học sâu đủ tốt và không bị quá khớp. Điều này cũng chính là mục tiêu của quá trình học như vậy điều này càng chứng minh thêm là trong bài toán học của mô hình học sâu thì độ chính xác dự đoán và rủi ro vi phạm tính riêng tư không trực tiếp mâu thuẫn nhau. Chúng ta có thể đồng thời đạt được hai mục tiêu này bằng cách huấn luyện ra một hình dự đoán tốt và chống quá khớp *(Hình 6)*.

Bên cạnh đó, một nguyên nhân nữa khiến các mô hình học sâu dễ vi phạm tính riêng tư là: các mô hình học sâu ngày càng sâu hơn, chứa lượng tham số khổng lồ để có thể học được một lượng dữ liệu lớn. Tuy nhiên việc này cũng vô tình làm các mô hình học sâu ghi nhớ (*memorize*) một số điểm dữ liệu thay vì học để tổng quát hoá tốt hơn *(Hình 7)*. Tổng quát hơn thì rõ ràng kiến trúc và phương pháp học dữ liệu có ảnh hưởng đến tính riêng tư của một điểm dữ liệu/một cá nhân trong tập dữ liệu huấn luyện. Trong [51], các tác giả cũng đã chỉ ra rằng phương pháp học cây quyết định (*decision tree*) dễ dàng bị tấn công suy luận thành viên và vi phạm tính riêng tư do với một điểm dữ liệu mang một đặc trưng mới nó hoàn toàn có thể làm cho mô hình rẽ thêm một nhánh quyết định. Trong khi đó naive Bayes thì bảo vệ tính riêng tư tốt nhất trong so với các mô hình còn lại trong khảo sát gồm mạng nơ-ron học sâu, hồi quy logistic, *k*-nearest neighbor và cây quyết định.



Hình 6. Độ chính xác của mô hình học sâu không trực tiếp ảnh hưởng đến tính riêng tư của dữ liệu mà nó dùng để huấn luyện.

Mô hình	Số lượng tham số	Rủi ro vi phạm riêng tư
AlexNet	2.47 triệu	75.1%
ResNet	1.7 triệu	64.3%
DenseNet	25.62 triệu	74.3%

Hình 7. Các mô hình học sâu ngày càng "sâu" chứa nhiều tham số hơn giúp cho việc học dữ liệu và tổng quát học tốt hơn. Tuy nhiên điều này vô tình làm các mô hình học sâu này ghi nhớ (thay vì học) một số điểm dữ liệu trong chính nó. Do đó rủi ro suy luận thành viên cũng cao hơn.

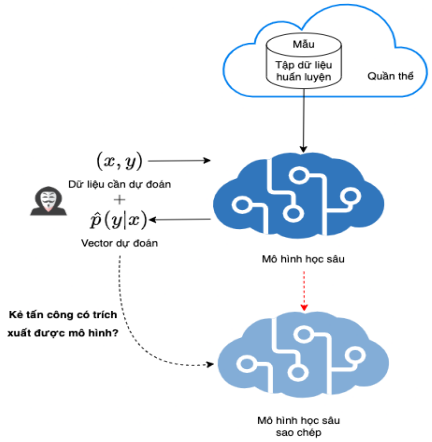
3.2. Các cuộc tấn công khác vi phạm tính riêng tư

Như đã trình bày ở trên, tấn công suy luận thành viên được xem như là thước đo cho việc lộ thông tin mà thường được các công trình sử dụng để kiểm tra về rủi ro vi phạm riêng tư dữ liệu. Tuy nhiên vi phạm riêng tư không chỉ dừng lại ở rủi ro suy luận thành viên, trong phần này chúng tôi sẽ tóm tắt lại một số cuộc tấn công vi phạm riêng tư dữ liệu khác đã xuất hiện trong các công trình gần đây.

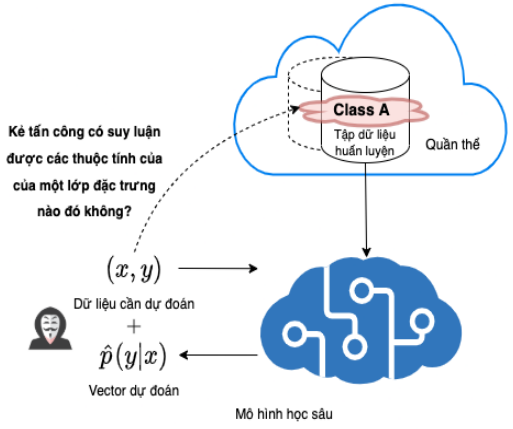
3.2.1. Tấn công trích xuất mô hình

Tấn công trích xuất mô hình (*Hình 8*) [14, 15, 52, 53] là cuộc tấn công mà dựa vào những công khai của mô hình, kẻ tấn công cố gắng suy luận ngược lại mô hình hoặc một mô hình gần đúng để "trộm" mô hình hoặc các khả năng dự đoán của mô hình. F. Tramèr và các cộng sự [20] đề xuất minh hoạ đầu tiên cho việc các mô hình có thể bị trích xuất và trộm, cụ thể trên một mô hình học tên là hồi quy logistic nhân (*kernel logistic regression*) với đặc điểm là mô hình này sẵn bên trong nó đã chứa (nhớ) một số điểm dữ liệu trong tập huấn luyện. Do đó tác giả đề xuất khai thác và trích xuất từ đây.

Hiện nay để huấn luyện các mô hình học sâu giải quyết các bài toán thực tế và xử lý được nhiều tác vụ cùng lúc ngày càng tốn kém [5, 6]. Vì vậy không sai khi nói rằng mô hình học sâu là một tài sản quý giá mà một tổ chức, doanh nghiệp phải bỏ rất nhiều tiền và tài nguyên để huấn luyện và sử dụng cho các vấn đề của mình gặp phải. Do đó nếu mô hình bị rò rỉ, bị trộm thì đây là thiệt hại không hề nhỏ. Tuy nhiên các mô hình tốn kém ấy lại thường được huấn luyện lại từ mô hình được huấn luyện trước đó (*pre-trained model*), mà mô hình huấn luyện từ mô hình được huấn luyện trước đó dễ bị trích xuất lại hơn [52].



Hình 8. Tấn công trích xuất mô hình



Hình 9. Tấn công suy luận tính chất

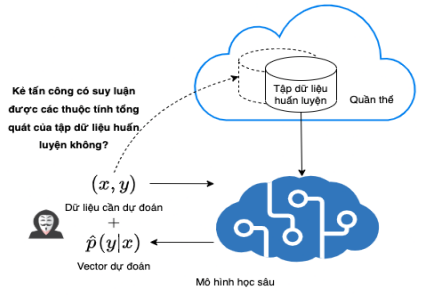
Việc làm sao để phòng chống và ngăn ngừa kẻ tấn công "trộm" mô hình là điều gần như không thể vì nếu xây dựng được tập dữ liệu thì kẻ tấn công hoàn toàn có thể xây dựng mô hình cho mình. Các cách khả dĩ nhất là giới hạn số lượng truy vấn và tìm cách chứng minh ràng buộc mà kẻ tấn công có thể tấn công trích xuất được. Ngoài ra làm sao cho việc xây dựng tập dữ liệu có phân phối gần giống với tập dữ liệu được sử dụng huấn luyện mô hình là khó khăn. Cuối cùng, vấn đề tấn công trích xuất mô hình có thể dẫn đến một bài toán, vấn đề mới là làm sao để đính bản quyền (*watermarking*), nhất là khi từ chi phí và giá trị của mô hình học sâu có thể coi như một tài sản sở hữu trí tuệ (*intellectual property*) [54].

3.2.2. Tấn công suy luận tính chất

Tấn công suy luận tính chất [29, 30] (*Hình 9*) là tấn công từ đầu ra của mô hình kẻ tấn công suy luận ngược lại các tính chất nhạy cảm đại diện của một lớp cụ thể. Ví dụ đối với hệ thống nhận diện khuôn mặt thì kẻ tấn công suy luận được một số người nhận diện được bị cận (không phải là mục đích của hệ thống). Việc khai thác được tính chất nhạy cảm trong cuộc tấn công này là khá ngẫu nhiên về cả xác suất thành công lẫn về xác định thuộc tính nào là nhạy cảm (phụ thuộc vào cảm tính chủ dữ liệu lẫn kẻ tấn công). Do đó về lý thuyết nếu mô hình chống chịu được tấn công suy luận thành viên thì cũng xem như bảo vệ được trước tấn công suy luận tính chất.

3.2.3. Tấn công suy luận thuộc tính và đảo ngược mô hình

Tấn công đảo ngược mô hình (*Hình 10*) [16, 17, 29] là cuộc tấn công mà dựa vào đầu ra của mô hình, kẻ tấn công cố gắng suy luận ngược lại những những đặc trưng chắc chắn của tập dữ liệu huấn luyện hoặc một vài thuộc tính của tập dữ liệu (tấn công suy luận thuộc tính). Cuộc tấn công này khác với tấn công suy luận thành viên là nó không cố gắng lại một hay một số điểm dữ liệu nằm trong tập dữ liệu huấn luyện mà chỉ cố gắng suy luận ra những thuộc tính chung nhất của tập dữ liệu. Do đó, tấn công suy luận thuộc tính và đảo ngược mô hình cũng không nhất thiết làm vi phạm riêng tư nếu những suy luận được về những thuộc tính của tập dữ liệu không phải là dữ liệu nhạy cảm. Ngược lại, nếu những thuộc tính chung nhất, tất yếu của tập dữ liệu để xây dựng mô hình là nhạy cảm thì việc tìm những kỹ thuật để che dấu, bảo vệ nó là điều không thể.



Hình 10. Tấn công suy luận thuộc tính và đảo ngược mô hình

Hình 11. Hình bên phải là dữ liệu gốc để xây dựng mô hình nhận diện khuôn mặt. Hình bên trái là kết quả của tấn công đảo ngược mô hình [17]

M. Fredrikson và các cộng sự [29] đề xuất minh họa đầu tiên cho rủi ro tấn công đảo ngược mô hình trong lĩnh vực gen dược lý học (*pharmacogenetics*). Tiếp đó, các tác giả này cũng minh hoạ cho cuộc tấn công này đối với dữ liệu là ảnh (*Hình 11*) [17]. Có thể thấy rằng đối với bài toán nhận diện khuôn mặt do các dữ liệu tương ứng với mỗi nhãn là cùng một người nên ta có thể dễ dàng biết được dữ liệu suy luận ra được là ai.

4. THẢO LUẬN VÀ HƯỚNG PHÁT TRIỂN

Thông qua những công trình nghiên cứu gần đây, rủi ro vi phạm tính riêng tư của các công nghệ, dịch vụ dựa trên mạng nơ-ron học sâu là có thật. Hiện nay, trên thế giới, ở một số quốc gia thì lộ dữ liệu riêng tư đã xảy ra và gây ra những hậu quả tiêu cực đối với các công ty, tổ chức cung cấp dịch vụ và thậm chí là cả xã hội. Trong đó nguyên nhân, cách thức vi phạm dữ liệu nhạy cảm, riêng tư của các công nghệ phần mềm lại gây khó khăn cho việc tìm hiểu và phòng ngừa cho những người không có kiến thức nền về công nghệ. Ví dụ ở Hoa Kỳ đã tổ chức những phiên điều trần đối với nhà người quản trị của các công ty công nghệ lớn (*big*

tech), tuy nhiên kết quả của một vài phiên điều trần cũng không quá khả quan cho việc hiểu và áp dụng luật pháp và chế tài phù hợp. Những nguyên nhân, cách thức vi phạm dữ liệu lại càng khó khăn hơn đối với các công nghệ đang trong giai đoạn áp dụng rộng rãi và đòi hỏi lượng kiến thức mới để hiểu như học sâu và trí tuệ nhân tạo. Ở Việt Nam, những cuộc tấn công khai thác dữ liệu riêng tư và những dữ liệu nhạy cảm được công khai mặc dù vẫn chưa để lại những hậu quả nặng nề như các nước khác nhưng chúng ta vẫn không thể chủ quan. Với bài báo chuyên khảo những công trình về chủ đề vấn đề riêng tư dữ liệu trong học sâu trong những năm gần đây, chúng tôi hy vọng phần nào đóng góp vào quá trình tìm hiểu và đưa ra các cách bảo vệ trong quá trình huấn luyện và sử dụng các công nghệ được hỗ trợ bởi học sâu.

Bên cạnh đó, chúng tôi nhận thấy một số thách thức, những điểm có thể trở thành trọng tâm nghiên cứu trong các công trình nghiên cứu tiếp theo:

Các công trình học sâu có rủi ro đồng thời cả về an toàn thông tin (*security*) và riêng tư. Các công trình, bài báo nghiên cứu thường chỉ phân tích và phòng chống rủi ro trong một hay một vài cuộc tấn công thuộc một tiêu chí an toàn thông tin cụ thể nào đó, ví dụ các cuộc tấn công nhắm vào tính toàn vẹn, tính sẵn sàng, hay các cuộc tấn công vào tính bí mật. Tuy nhiên trong thực tế một để xây dựng được một hệ thống đáng tin tưởng (*trustworthy*) [55] thì các tiêu chí này phải cùng đi với nhau. Do đó các công trình nghiên cứu cần nghiên cứu về vấn đề các mạng nơ-ron học sâu thoả cả về an toàn lẫn riêng tư thông tin [56].

Giống như các công nghệ khác như cơ sở dữ liệu, ứng dụng web,… Các phương pháp học cũng dần tiến đến việc phân tán việc học (*distributed learning*) và học cộng tác (*collaborative learning*). Các phương pháp học này càng được củng cố mạnh mẽ hơn với các xu hướng công nghệ về mặt phần cứng như điện toán biên (*edge computing*) và điện toán sương mù (*fog computing*) thay thế một phần hay hoàn toàn điện toán đám mây (*cloud computing*) trong tương lai. Việc bảo vệ riêng tư trong ngữ phân tán sẽ có những bài toán và đặc điểm khác với phương pháp huấn luyện trung tâm. Trong các kỹ thuật thuộc loại học này thì học liên kết đang nổi lên như là một phương pháp được sử dụng. Trong thời gian tới, các công trình nghiên về bảo vệ tính riêng tư cho học liên kết chắc chắn sẽ được đẩy mạnh.

Cuối cùng, tấn công suy luận thành viên hay còn có thể được riêng tư thành viên dần trở thành tiêu chí để đánh giá rủi ro vi phạm riêng tư của các công nghệ, không chỉ là trong các mô hình học sâu. Những cuộc tấn công suy luận thành viên này đã và đang được nghiên cứu và đánh giá kỹ càng hơn trong các mô hình học sâu khác nhau. Tuy nhiên việc nghiên cứu những ứng dụng rộng hơn của cuộc tấn công này để hỗ trợ nhiều vấn đề khác những trong học sâu vẫn còn khá hạn chế. Liệu riêng tư thành viên có thể giúp ta giải quyết, mô hình hoá nhiều vấn đề liên quan khác về riêng tư dữ liệu của học sâu là một câu hỏi hay và cần nhiều công trình hơn để khai phá.

TÀI LIỆU THAM KHẢO

1. Krizhevsky A., Sutskever I., and Hinton G.E. - Imagenet classification with deep convolutional neural networks, In Proceedings of the 25th International Conference on Neural Information Processing Systems (2012) (1) 1097–1105.
2. Ren S., He K., Girshick R., and Sun J. - Faster r-cnn: Towards real-time object detection with region proposal networks, In Proceedings of the 28th International Conference on Neural Information Processing Systems (2015) (1) 91-99.
3. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., and Houlsby N. - An image is worth 16x16 words: Transformers for image recognition at scale, In The International Conference on Learning Representations (2021) 1-22.

4. Vaswani A., Shazeer S., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., and Polosukhin I. - Attention is all you need, In Proceedings of the 31st International Conference on Neural Information Processing Systems (2017) 6000–6010.
5. Devlin J., Chang M.-W., Lee K., and Toutanova K. - Bert: Pre-training of deep bidirectional transformers for language understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019) (1) 4171-4186.
6. Brown T. B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D. M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., and Amodei D. - Language models are few-shot learners, In Proceedings of the 34th International Conference on Neural Information Processing Systems (2020) 1877–1901.
7. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., and Fergus R. - Intriguing properties of neural networks, In The International Conference on Learning Representations (2014) 1-10.
8. Goodfellow I. J., Shlens J., and Szegedy C. - Explaining and harnessing adversarial examples, In The International Conference on Learning Representations (2015) 1-11.
9. Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z. B., and Swami A. - Practical black-box attacks against machine learning, Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (2017) 506-519.
10. Chen X., Liu C., Li B., Lu K., and Song D. - Targeted backdoor attacks on deep learning systems using data poisoning, online (2017) 1-18.
11. Gu T., Dolan-Gavitt B., and Garg S. - Badnets: Identifying vulnerabilities in the machine learning model supply chain, IEEE Access **7** (2019) 47230-47244.
12. Shafahi A., Huang W. R., Najibi M., Suciu O., Studer C., Dumitras T., and Goldstein T. - Poison frogs! targeted clean-label poisoning attacks on neural networks, Advances in Neural Information Processing Systems **31** (2018) 6106-6116.
13. Dang T.K., Truong P.T.T., and Tran P.T. - Data poisoning attack on deep neural network and some defense methods, In 2020 International Conference on Advanced Computing and Applications (2020) 15–22.
14. Tramèr F., Zhang F., Juels A., Reiter M.K., and Ristenpart T. - Stealing machine learning models via prediction apis, 25th USENIX Security Symp. (2016) 601-618.
15. Jagielski M., Carlini N., Berthelot D., Kurakin A., and Papernot N. - High accuracy and high fidelity extraction of neural networks, 29th USENIX Security Symposium (2020) 1345-1362.
16. Fredrikson M., Jha S., and Ristenpart T. - Model inversion attacks that exploit confidence information and basic countermeasures, Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (2015) 1322–1333.
17. Shokri R., Stronati M., Song C., and Shmatikov V. - Membership inference attacks against machine learning models, IEEE Symp. on Security and Privacy (2017) 3-18.
18. Choquette-Choo C.A., Tramèr F., Carlini N., and Papernot N. - Label-only membership inference attacks, Proceedings of the 38th International Conference on Machine Learning (2021) 1964-1974.

19. Nasr M., Shokri R., and Houmansadr A. - Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning, IEEE Symposium on Security and Privacy (2019) 739-753.
20. Carlini N., Tramer F., Wallace E., Jagielski M., Herbert-Voss A., Lee K., Roberts A., Brown T., Song D., Erlingsson U., Oprea A., and Raffel C. - Extracting training data from large language models, USENIX Security Symposium (2021) 2633-2650.
21. Long Y., Wang L., Bu D., Bindschaedler V., Wang X., Tang H., Gunter C.A., and Chen K. - A pragmatic approach to membership inferences on machine learning models, in 2020 IEEE European Symposium on Security and Privacy (2020) 521–534.
22. Salem A., Zhang Y., Humbert M., Berrang P., Fritz M., and Backes M., Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, 26th Annual Network and Distributed System Security Symp. (2019) 1-15.
23. Sweeney L. - K-anonymity: A model for protecting privacy, in International Journal on Uncertainty, Fuzziness and Knowledge-based Systems **10** (5) (2002) 557– 570.
24. Machanavajjhala A., Gehrke J., Kifer D., and Venkitasubramaniam M. - L-diversity: Privacy beyond k-anonymity, In Proceedings of the 22nd International Conference on Data Engineering (2006) 1-12.
25. LeCun, Y., Bengio, Y. & Hinton, G - Deep learning. Nature **521** (2015) 436–444.
26. Dalenius T. - Towards a methodology for statistical disclosure control, Statistik Tidskrift (1977) 429–444.
27. Li N., Qardaji W., Su D., Wu Y., and Yang W. - Membership privacy: A unifying framework for privacy definitions, in Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (2013) 889–900.
28. Murakonda S.K. and Shokri R. - Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning, 13th Workshop on Hot Topics in Privacy Enhancing Technologies (2020), 1-3.
29. Fredrikson M., Lantz E., Jha S., Lin S., Page D., and Ristenpart T. - Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing, in Proceedings of the 23rd USENIX Conference on Security Symposium (2014) 17–32.
30. Ganju K., Wang Q., Yang W., Gunter C., and Borisov N. - Property inference attacks on fully connected neural networks using permutation invariant representations, Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (2018) 619–633.
31. Dwork C., Smith A., Steinke T., Ullman J., and Vadhan S. - Robust traceability from trace amounts, in 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (2015) 650–669.
32. Pyrgelis A., Troncoso C., and De Cristofaro E. - Knock knock, who's there? membership inference on aggregate location data, 25th Annual Network and Distributed System Security Symposium (2018) 1-15.
33. Erlich, Y., Narayanan, A - Routes for breaching and protecting genetic privacy. Nat Rev Genet **15** (2014) 409–421.
34. Wright C.E., Barbara J.E., James W.H., Mark A.R. - The law of genetic privacy: applications, implications, and limitations, Journal of Law and the Biosciences **6** (1) (2019) 1–36.

35. Ateniese G., Felici G., Mancini L.V., Spognardi A., Villani A., and Vitali D. - Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers, Intl. Journal of Security and Networks **10** (3) (2015) 37-150.
36. Hu H., Salcic Z., Sun L., Dobbie G., Yu P.S., and Zhang X., Membership inference attacks on machine learning: A survey, ACM Computing Surveys (2022) 1-41.
37. Samuel Y., Irene G., Matt F., and Somesh J. - Privacy risk in machine learning: Analyzing the connection to overfitting, In IEEE 31st Computer Security Foundations Symposium (2018) 268–282.
38. Song L. and Prateek M. - Systematic evaluation of privacy risks of machine learning models. In 30th USENIX Security Symp. (USENIX Security 21) (2021) 2615–2632.
39. Li Z. and Zhang Y. - Membership Leakage in Label-Only Exposures, In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (2021) 880-895.
40. McMahan B., Moore E., Ramage D., Hampson S., and Arcas B. A.y. - Communication-efficient learning of deep networks from decentralized data, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (2017) 1273–1282.
41. Kairouz P., McMahan H.B., Avent B., Bellet A., Bennis M., Bhagoji A.N., Bonawitz K., Charles Z., Cormode G., Cummings R., and et al. - Advances and open problems in federated learning, Foundations and Trends in Machine Learning **14** (1-2) (2021) 1-210.
42. Hu H., Salcic Z., Sun L., Dobbie G., and Zhang X. - Source inference attacks in federated learning, IEEE International Conference on Data Mining (2021) 1102-1107.
43. Kim M., Günlü O.; Schaefer R.F. - Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication, IEEE International Conference on Acoustics, Speech and Signal Processing (2021) 2650-2654.
44. Truex S., Baracaldo N., Anwar A., Steinke T., Ludwig H., Zhang R., and Zhou Y. - A hybrid approach to privacy-preserving federated learning, Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (2018) 1-11.
45. Hayes J., Melis L., Danezis G., and Cristofaro E.D. - Logan: Membership inference attacks against generative models , Proceedings on Privacy Enhancing Technologies 2019 (1) (2019) 133-152.
46. Chen D., Yu N., Zhang Y., and Fritz M. - GAN-leaks: A taxonomy of membership inference attacks against generative models, in Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (2020) 343-362.
47. Mireshghallah F., Goyal K., Uniyal A., Berg-Kirkpatrick T., and Shokri R. - Quantifying privacy risks of masked language models using membership inference attacks, online (2022) 1-16.
48. Hisamoto S., Post M., and Duh K. - Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system?, Transactions of the Association for Computational Linguistics **8** (2020) 49–63.
49. Wang Y., Huang L., Yu P. S., and Sun L., Membership inference attacks on knowledge graphs, online (2021) 1-11.
50. Shah M., Szurley J., Mueller M., Mouchtaris A., and Droppo J. - Evaluating the vulnerability of end-to-end automatic speech recognition models to membership inference attacks (2021) 891–895.

51. Truex S., Liu L., Gursoy M.E., Yu L., and Wei W. - Demystifying Membership Inference Attacks in Machine Learning as a Service. IEEE Transactions on Services Computing **01** (2019) 1–17.
52. Krishna K., Tomar G.S., Parikh A.P., Papernot N., and Iyyer M. - Thieves on sesame street! model extraction of bert-based apis, 8th International Conference on Learning Representations (2020) 1-19.
53. Orekondy T., Schiele B., and Fritz M. - Knockoff nets: Stealing functionality of black-box models, 2019 IEEE conf. on computer vision and pattern recogn. (2019) 4954-4963.
54. Maini P., Yaghini M., and Papernot N. - Dataset inference: Ownership resolution in machine learning, 9th Intl. Conference on Learning Representations (2021) 1-22.
55. Papernot N. - What does it mean for machine learning to be trustworthy?, 1st ACM Workshop on Security and Privacy on Artificial Intelligence (2020) 1-25.
56. Phan N., Thai M.T., Hu H., Jin R., Sun T., and Dou D. - Scalable differential privacy with certified robustness in adversarial learning, Proceedings of the 37th International Conference on Machine Learning (2020) 7683-7694.

ABSTRACT

RISKS OF DATA PRIVACY VIOLATION IN DEEP LEARNING

Tran Truong Tuan Phat[1,2], Dang Tran Khanh[1*]
[1]*Ho Chi Minh City University of Food Industry, Vietnam*
[2]*Ho Chi Minh City University of Technology, VNU-HCM, Vietnam*
[*]Email: *khanh@hufi.edu.vn*

Thanks to the superior predictability of deep learning methods, artificial intelligence (AI)-applied technologies solve a wide range of problems and are increasingly widely used in many fields and industries. However, deep learning-based machine learning models are good at many tasks, problems but not perfect, typically these models are very vulnerable to various attacks which violate information security criteria. In particular, the risk of data privacy breaches is an itchy issue because it not only affects the system, service providers, users but also the safety and trust of people in using these technologies, thereby seriously leading to social and legal issues. In this article, we summarize and analyze the related works of privacy violation issues in deep learning in recent years, thereby modeling and giving warnings when building deep learning models.

Keywords: Data privacy, deep learning, big data, data security, access control.