NHỮNG RỦI RO VÀ CÁC PHÒNG CHỐNG VI PHẠM TÍNH RIÊNG TƯ TRONG MÔ HÌNH HỌC CỘNG TÁC

Hà Lê Hoài Trung[1], Đặng Trần Khánh[2*]

[1]*Trường Đại học Công nghệ Thông tin - ĐHQG TP.HCM*

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

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

Ngày nhận bài: 10/6/2022; Ngày chấp nhận đăng: 13/7/2022

TÓM TẮT

Với việc phát triển khoa học như ngày nay, con người tận hưởng cuộc sống hiện đại và nhiều tiện nghi hơn, đồng thời cũng tạo ra nhiều dữ liệu. Các dữ liệu này được lưu trữ trong các thiết bị và miền ứng dụng khác nhau, đồng thời xã hội cũng ngày càng nhận thức rõ hơn về các vấn đề bảo mật, tính riêng tư của dữ liệu, việc huấn luyện các mô hình học máy tập trung hay các mô hình trí tuệ nhân tạo (AI) truyền thống đang phải đối mặt với những thách thức về tính hiệu quả và quyền riêng tư. Trong những năm gần đây, học liên kết (Federated Learning) đã nổi lên như một giải pháp thay thế và tiếp tục phát triển mạnh trong lĩnh vực trí tuệ nhân tạo phục vụ cuộc sống con người. Các mô hình học liên kết hiện tại có một số lỗ hổng dễ bị tấn công bởi người tấn công nằm bên trong hoặc bên ngoài hệ thống, ảnh hưởng đến quyền riêng tư của dữ liệu và tính bảo mật của hệ thống. Bên cạnh việc huấn luyện các mô hình toàn cục yêu cầu tính bảo mật là điều quan trọng trong các thiết kế các hệ thống học liên kết có đảm bảo quyền riêng tư và có khả năng chống lại các loại tấn công khác nhau. Nghiên cứu này trình bày toàn diện về quyền riêng tư và tính bảo mật trong học liên kết, bao gồm: 1) các mối đe dọa; 2) các cuộc tấn công và phòng thủ về quyền riêng tư. Các kỹ thuật chính cũng như các giả định cơ bản được áp dụng bởi các cuộc tấn công và phòng thủ khác nhau trong học liên kết cũng được trình bày giúp hiểu rõ hơn về bản chất và điều kiện thực hiện tấn công. Cuối cùng, các hướng nghiên cứu trong tương lai nhằm bảo vệ tính riêng tư trong mô hình học liên kết cũng sẽ được thảo luận chi tiết.

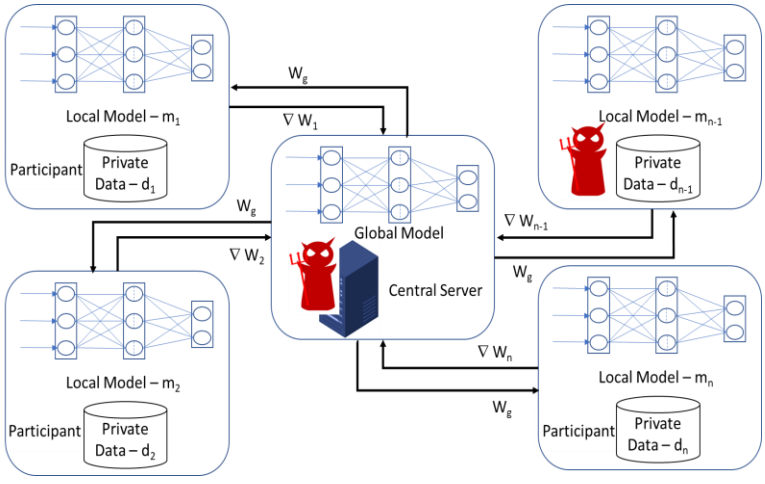
Từ khóa: Học cộng tác, tính riêng tư, mạng sinh đối kháng, tấn công, phòng chống.

1. GIỚI THIỆU

Các thiết bị điện toán ngày càng trở nên phổ biến như ti vi thông minh, điện thoại thông minh, đồng hồ thông minh, nhà thông minh, … Nhưng thiết bị này giúp cho cuộc sống hiện đại trở nên tiện nghi hơn hiện đại hơn, đồng thời tạo ra một lượng lớn dữ liệu. Việc thu thập dữ liệu từ các thiết bị vào các hệ thống cơ sở lưu trữ tập trung rất tốn kém và mất thời gian. Các phương pháp tiếp cận học máy tập trung theo phương pháp truyền thống gặp một số hạn chế về cơ sở hạ tầng như băng thông hạn chế, kết nối mạng không liên tục và các hạn chế về độ trễ [35]. Một mối quan tâm quan trọng khác là quyền riêng tư của dữ liệu và tính bảo mật của người dùng vì dữ liệu sử dụng thường chứa thông tin nhạy cảm [1]. Dữ liệu nhạy cảm như hình ảnh khuôn mặt, dịch vụ dựa trên vị trí hoặc thông tin sức khỏe có thể được sử dụng để quảng cáo và khuyến nghị trên mạng xã hội gây ra những rủi ro về quyền riêng tư tức thời hoặc tiềm ẩn. Do đó, dữ liệu riêng tư không nên được chia sẻ trực tiếp mà không có bất kỳ biện pháp bảo vệ quyền riêng tư nào.

Sau sự vụ vi phạm dữ liệu Facebook (facebook dataleak 2018), mọi người ngày càng lo lắng về tính bảo mật và quyền riêng tư dữ liệu của họ cũng như nơi dữ liệu được sử dụng. Vì việc sử dụng bất hợp pháp và khai thác dữ liệu cá nhân là mối đe dọa đối với cả cá nhân và doanh nghiệp, nên việc rò rỉ dữ liệu có thể gây ra một số hậu quả nghiêm trọng. Do đó, một số quốc gia trên toàn thế giới đang đưa ra luật mới để bảo đảm an ninh dữ liệu và quyền riêng tư như quy định chung về bảo vệ dữ liệu là một trong những ví dụ điển hình về luật mà Liên minh Châu Âu đã ban hành vào năm 2018. GDPR trao cho người dùng quyền cao hơn qua dữ liệu cá nhân của họ. Mục đích và mục tiêu chính của quy định chung về bảo vệ dữ liệu là bảo vệ quyền riêng tư của người dùng và cung cấp bảo mật dữ liệu. Quyền được quên cho phép người dùng xóa dữ liệu của họ theo yêu cầu, do đó giữ cho dữ liệu ở chế độ riêng tư, nhưng đồng thời khiến cho các phép toán tổng hợp dữ liệu trở nên khó khăn hơn hơn.

Việc ban hành các quy định này chắc chắn sẽ đặt ra những thách thức mới trong các thủ tục giao dịch dữ liệu. Những yếu tố này đã dẫn đến sự thay đổi mô hình trong cách xử lý dữ liệu và lưu trữ dữ liệu. Trí tuệ nhân tạo thường được sử dụng để giải quyết các vấn đề phức tạp, nhưng sự ra đời của dữ liệu lớn đã dẫn đến sự xuất hiện của các công nghệ điện toán mới như điện toán đám mây, học liên kết, v.v. thể hiện trong Hình 1. Cơ bản mô hình này là cho phép chủ sở hữu dữ liệu từ các lĩnh vực khác nhau hợp tác để xây dựng mô hình học máy trong khi đảm bảo, bảo vệ quyền riêng tư và bí mật dữ liệu của người dùng. Mối quan tâm về quyền riêng tư của người dùng trong quá trình trao đổi dữ liệu đã được trình bày trong các công trình nghiên cứu của [13, 26]. Vì vậy, trong một vài năm gần đây thế giới thấy một sự thay đổi từ mô hình học máy truyền thống sang học liên kết. Hầu hết tất cả các vấn đề có thể được giải quyết bằng mô hình học máy truyền thống như xử lý hình ảnh / giọng nói / văn bản, xử lý và phân tích dữ liệu được thu thập từ cảm biến, phát triển hệ thống khuyến nghị hiệu quả, v.v. có thể được giải quyết bằng học liên kết.



Hình 1. Quá trình huấn luyện học liên kết. Người tấn công có thể là máy chủ hoặc là người tham gia huấn luyện mô hình

2. PHÂN LOẠI MÔ HÌNH HỌC LIÊN KẾT

2.1. Dựa trên phân bố dữ liệu

Dựa trên sự phân phối các đặc trưng dữ liệu và mẫu dữ liệu của những người tham gia, mô hình học liên kết có thể được phân loại thành 3 dạng học liên kết theo chiều ngang, học liên kết theo chiều dọc và học liên kết chuyển tiếp.

Theo học liên kết theo chiều ngang, các tập dữ liệu do mỗi người tham gia sở hữu chia sẻ các đặc trưng tương tự nhưng liên quan đến những người dùng khác nhau. Ví dụ: một số bệnh viện có thể lưu trữ các loại dữ liệu tương tự nhau (ví dụ: nhân khẩu, thông tin bệnh lâm sàng và gen) về các bệnh nhân khác nhau.

Bảng 1. Phân loại mô hình học liên kết ngang

Học liên kết theo chiều ngang	Số lượng tham gia	Tần số tham gia huấn luyện	Yêu cầu đáp ứng tài nguyên
H2B	Nhỏ (vài chục hoặc vài trăm)	Thường xuyên	Cao
H2C	Lớn (vài nghìn cho đến vài triệu)	Không thường xuyên	Thấp

Nếu họ quyết định cùng nhau xây dựng một mô hình học máy sử dụng mô hình liên kết. Trong bài báo này, mô hình học liên kết ngang được phân loại thành học liên kết ngang cho doanh nghiệp (H2B) và học liên kết ngang cho người dùng (H2C). So sánh giữa H2B và H2C được liệt kê trong Bảng 1. Sự khác biệt chính nằm ở số lượng người tham gia, mức độ tham gia huấn luyện học liên kết và có thể ảnh hưởng đến cách các đối tượng tấn công cố gắng xâm nhập hệ thống học liên kết. Theo H2B, số lượng người tham gia thường ít. Các đối tượng này có thể được chọn thường xuyên trong quá trình huấn luyện mô hình học liên kết. Những người tham gia có xu hướng sở hữu sức mạnh tính toán lớn và các công nghệ học máy phức tạp. Trong khi theo H2C, số lượng người tham gia tiềm năng có thể lên đến hàng nghìn hoặc thậm chí hàng triệu người. Trong mỗi vòng lặp huấn luyện, chỉ một số lượng nhỏ trong số hàng trăm, hàng triệu người dùng được chọn.

2.2. Dựa trên kiến trúc

2.2.1 Kiến trúc đồng nhất

Học liên kết với kiến trúc đồng nhất: Chia sẻ gradient thường chỉ giới hạn ở các kiến trúc học liên kết đồng nhất, tức là cùng một mô hình học máy được chia sẻ với tất cả những người tham gia. Những người tham gia nhằm mục đích cộng tác, xây dựng một mô hình chính xác hơn. Cụ thể, các tham số học máy w của mô hình thường thu được thông qua việc giải bài toán tối ưu hóa: $min_w \sum_{i=1}^{n} F(w, D_i)$, trong đó F (w; D_i) là hàm mục tiêu cho tập dữ liệu huấn luyện cục bộ trên người tham gia thứ i và mô tả tham số mô hình w trên tập dữ liệu huấn luyện cục bộ D_i. Các bộ phân loại khác nhau (ví dụ: hồi quy logistic, mạng nơ ron học sâu) sử dụng các hàm mục tiêu khác nhau. Trong học liên kết, mỗi người tham gia duy trì một mô hình học máy cục bộ cho bộ dữ liệu huấn luyện cục bộ của mình. Máy chủ duy trì mô hình học máy toàn cục thông qua tổng hợp các tham số mô hình cục bộ từ n người tham gia. Cụ thể, học liên kết với kiến trúc đồng nhất thực hiện các bước trong Hình 1. Học liên kết với kiến trúc đồng nhất thường có hai dạng [20]: (1) FedSGD, trong đó mỗi người tham gia gửi cập nhật Stochastics Gradient Descent (SGD) đến máy chủ; (2) FedAvg, trong mô hình này người tham gia thực hiện hàng loạt các lần lặp lại tính toán cục bộ giá trị SGD trước khi gửi các bản cập nhật đến máy chủ, phương pháp này có chi phí giao tiếp hiệu quả hơn. Tất cả các phương pháp này đều dựa trên quy tắc tông hợp trung bình tức là lấy giá trị trung bình của các tham số mô hình cục bộ làm mô hình toàn cục. Tuy nhiên, giá trị trung bình của mô hình toàn cục có thể bị thao túng tùy ý ngay cả khi một người tham gia bị tấn công hay bị chiếm quyền [3].

2.2.2. Kiến trúc không đồng nhất

Học liên kết với kiến trúc không đồng nhất: Những nỗ lực gần đây đã mở rộng vấn đề học liên kết cho phép hợp tác huấn luyện các mô hình có kiến trúc không đồng nhất [17]. Việc

huấn luyện mô hình liên kết thông thường chỉ có thể thực hiện trực tiếp trên các tham số mô hình học máy nếu tất cả các mô hình cục bộ có cùng cấu trúc. Chia sẻ dự đoán mô hình thay vì tham số mô hình hoặc cập nhật các tham số mô hình và loại bỏ nguy cơ tấn công suy luận hộp trắng trong học liên kết. Không giống như các thuật toán học tập liên kết hiện có, chắt lọc mô hình liên kết (Federated Model Distillation – FedMD) không buộc mô hình toàn cục vào các mô hình cục bộ. Mỗi mô hình học cục bộ được cập nhật riêng biệt, những người tham gia chia sẻ kiến thức về các mô hình cục bộ của họ thông qua dự đoán của họ trên một tập hợp dữ liệu công khai và không được gắn nhãn [14]. Một lợi ích của việc chia sẻ mô hình học theo phương pháp FedMD là giảm chi phí giao tiếp mà không ảnh hưởng đáng kể đến độ hiệu dụng của mô hình.

3. NHỮNG ĐE DỌA VÀ PHƯƠNG PHÁP BẢO VỆ TÍNH RIÊNG TƯ TRONG MÔ HÌNH HỌC LIÊN KẾT

3.1. Những rủi ro

Học liên kết cung cấp mô hình huấn luyện có nhận thức về quyền riêng tư, như không yêu cầu chia sẻ dữ liệu và cho phép người tham gia và rời khỏi quá trình huấn luyện một cách tự do. Tuy nhiên, các nghiên cứu gần đây đã chứng minh rằng học liên kết có thể không phải lúc nào cũng đảm bảo về quyền riêng tư cho người tham gia. Các giao thức học liên kết hiện tại dễ bị tấn công bởi: (1) một máy chủ độc hại nhằm mục đích suy ra thông tin nhạy cảm từ các bản cập nhật tham số mô hình riêng lẻ theo thời gian, can thiệp vào quá trình huấn luyện hoặc kiểm soát chế độ xem của những người tham gia về các tham số toàn cục; (2) nếu một trong số những người tham gia là kẻ tấn công thì họ có thể suy ra thông tin nhạy cảm của những người tham gia khác, thông qua việc giả mạo tổng hợp tham số toàn cục hoặc tấn công đầu độc vào mô hình toàn cục.

Bảng 2. Tổng hợp các loại tấn công trong mô hình học liên kết

Loại tấn công	Mục tiêu tấn công		Vai trò người tấn công		Kịch bản huấn luyện		Mức độ phức tạp tấn công		
	Mô hình	Dữ liệu huấn luyện	Người tham gia	Máy chủ	H2B	H2C	Số lần lặp tấn công		Yêu cầu thêm thông tin
							1 lần	Nhiều lần	
Suy diễn lớp đại diện	Không	Có	Có	Có	Có	Không	Không	Có	Có
Suy diễn thành viên trong tập huấn luyện	Không	Có	Có	Có	Có	Không	Không	Có	Có
Suy diễn thuộc tính	Không	Có	Có	Có	Có	Không	Không	Có	Có
Suy diễn dữ liệu huấn luyện và dữ liệu gán nhãn	Không	Có	Không	Có	Có	Không	Có	Có	Không

Về vấn đề rò rỉ quyền riêng tư, các gradient trao đổi trong suốt quá trình huấn luyện có thể tiết lộ thông tin nhạy cảm [5, 10], thậm chí gây ra rò rỉ các tham số mô hình học sâu [14],

cho bên thứ ba hoặc máy chủ trung tâm [2]. Ví dụ, như đã đề cập trong [5], ngay cả một phần nhỏ của gradient cũng có thể tiết lộ một lượng lớn thông tin nhạy cảm về dữ liệu cục bộ. Các nghiên cứu gần đây cho thấy rằng, chỉ cần quan sát các gradient, người tấn công có thể đánh cắp thành công dữ liệu huấn luyện [30] thông qua phương pháp tấn công xây dựng lại dữ liệu huấn luyện và mạng sinh đối kháng (Generative adversarial network – GAN).

Các cuộc tấn công về quyền riêng tư gây ra các mối đe dọa đáng kể đối với học liên kết. Trong quá trình huấn luyện tập trung, máy chủ chịu trách nhiệm về sự riêng tư của tất cả những người tham gia và mức độ bảo mật của mô hình. Tuy nhiên, trong học liên kết, bất kỳ người tham gia nào cũng có thể tấn công máy chủ và do thám những người tham gia khác được tóm tắt trong Bảng 2.

3.1.1. Rủi ro từ mô hình

Trước khi xem xét các cuộc tấn công vào mô hình học liên kết, trước tiên bảng tóm tắt các mối đe dọa mô hình được giới thiệu. Nói chung, các mối đe dọa trong mô hình học liên kết có thể được phân loại thành hai loại: (1) các đối tượng bên trong mô hình và các đối tượng bên ngoài mô hình; (2) giai đoạn huấn luyện và giai đoạn suy luận. Các mối đe dọa mô hình này áp dụng cho cả tính riêng tư và bảo mật.

A. Các đối tượng bên trong mô hình và các đối tượng bên ngoài mô hình

Các cuộc tấn công có thể được thực hiện bởi người tham gia trong quá trình huấn luyện mô hình và người bên ngoài mô hình. Các cuộc tấn công nội gián bao gồm các cuộc tấn công do máy chủ và những người tham gia trong hệ thống học liên kết thực hiện. Các cuộc tấn công từ bên ngoài bao gồm các cuộc tấn công do những người nghe trộm thực hiện dựa trên quá trình quan sát trao đổi thông tin giữa những người tham gia và máy chủ học liên kết và bởi những người sử dụng mô hình học liên kết khi nó được triển khai như một dịch vụ.

Các cuộc tấn công từ bên trong thường nguy hiểm hơn các cuộc tấn công từ bên ngoài, vì nó tăng cường khả năng tấn công thành công của đối thủ. Vì vậy, cuộc khảo sát trong tài liệu nghiên cứu này hướng về các cuộc tấn công chống lại mô hình học liên kết sẽ tập trung chủ yếu vào các cuộc tấn công nội gián.

B. Giai đoạn huấn luyện và giai đoạn suy luận

Giai đoạn huấn luyện: Các cuộc tấn công được thực hiện trong giai đoạn huấn luyện nhằm tìm hiểu, gây ảnh hưởng hoặc làm hỏng chính mô hình học liên kết [5]. Trong giai đoạn huấn luyện, người tấn công thực hiện các cuộc tấn công đầu độc dữ liệu để xâm phạm tính toàn vẹn của tập dữ liệu huấn luyện [23] hoặc các cuộc tấn công đầu độc mô hình để làm mất tính toàn vẹn của quá trình huấn luyện [12]. Người tấn công cũng có thể thực hiện một loạt các cuộc tấn công suy luận vào bản cập nhật tham số mô hình của từng người tham gia hoặc vào bản cập nhật tham số tổng hợp từ tất cả những người tham gia trong quá trình huấn luyện [18].

Giai đoạn suy luận: Các cuộc tấn công được tiến hành trong giai đoạn suy luận được gọi là cuộc tấn công né tránh hoặc thăm dò. Các đối tượng tấn công thường không thay đổi các tham số mô hình mục tiêu, thay vào đó, họ đánh lừa các mô hình này để đưa ra các dự đoán sai hoặc thu thập các đặc điểm của mô hình, sau đó thực hiện các vi phạm về quyền riêng tư và độ ổn định của mô hình. Hiệu quả của các cuộc tấn công như vậy phần lớn được xác định bởi thông tin có sẵn của đối thủ về mô hình. Các cuộc tấn công giai đoạn suy luận có thể được phân loại thành các cuộc tấn công hộp trắng (nghĩa là có toàn quyền truy cập vào mô hình học liên kết) và các cuộc tấn công hộp đen (tức là chỉ có thể truy vấn mô hình học liên kết) [14].

C. Quyền riêng tư: bán trung thực và độc hại

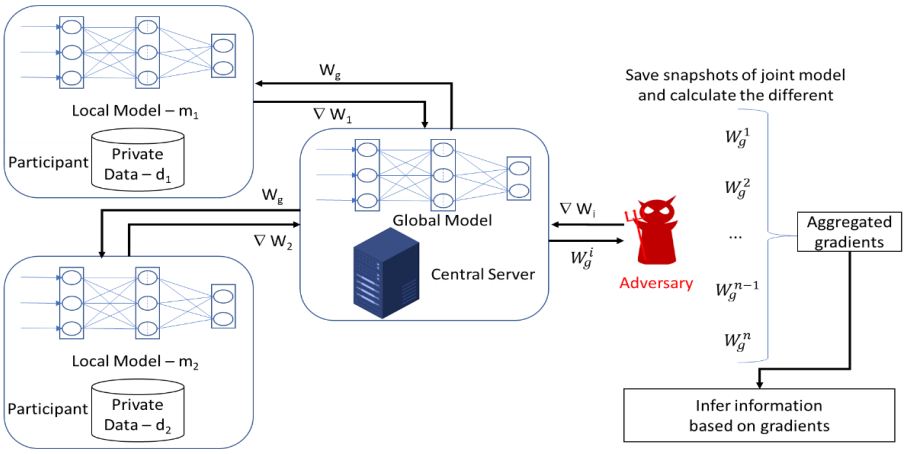
Trong trường hợp bán trung thực: người tấn công thực hiện tấn công một cách thụ động hoặc là những người trung thực nhưng tò mò. Họ cố gắng tìm hiểu trạng thái riêng tư của

những người tham gia khác mà không làm ảnh hưởng đến mô hình học liên kết. Người tấn công chỉ quan sát, và thu thập thông tin nhận được, tức là các thông số của mô hình toàn cục.

Trong trường hợp tấn công độc hại: người tấn công thực hiện tấn công theo cách tương tác (thực hiện các truy vấn đến mô hình) hoặc độc hại cố gắng tìm hiểu trạng thái riêng tư của những người tham gia trung thực và đồng thời làm ảnh hưởng đến mô hình học liên kết bằng cách sửa đổi, làm sai lệch các tham số mô hình học.

3.1.2. Tấn công tính riêng tư

Mặc dù học liên kết ngăn cản người tham gia chia sẻ trực tiếp dữ liệu cá nhân của họ, nhưng một loạt các nghiên cứu đã chứng minh rằng việc trao đổi gradient trong học liên kết cũng có thể làm rò rỉ thông tin nhạy cảm về dữ liệu cá nhân của người tham gia cho những người tấn công thụ động hoặc chủ động [18, 25]. Ví dụ: gradient của hai thời điểm liên tiếp của các thông số mô hình học liên kết có thể làm rò rỉ các đặc trưng không mong muốn của dữ liệu huấn luyện người tham gia cho những người tấn công, vì các mô hình học sâu có xu hướng nhận ra và ghi nhớ nhiều đặc trưng của dữ liệu hơn mức cần thiết cho nhiệm vụ học tập chính [26]. Hình 2 minh họa tập hợp thông tin mà người tấn công có thể suy ra từ các gradient (Δw_1^t), hoặc tương đương sự khác biệt của hai thời điểm liên tiếp của các tham số mô hình (tức là $w^{t+1} - w^t$).



Hình 2. Rò rỉ tính riêng tư từ mô hình học liên kết

Lý do tại sao gradient có thể gây rò rỉ thông tin riêng tư là vì gradient được lấy từ dữ liệu huấn luyện riêng của người tham gia và một mô hình học máy có thể được coi là đại diện đặc trưng của tập dữ liệu đã được huấn luyện. Trong mô hình học sâu, gradient của một lớp được tính toán dựa trên các đặc trưng của lớp đó và sai số với lớp sau. Trong trường hợp các lớp được kết nối đầy đủ theo tuần tự, các gradient của trọng số là tích vô hướng của các đặc trưng của lớp hiện tại và sai số từ lớp sau [18]. Do đó, các quan sát về gradient có thể được sử dụng để suy ra một lượng đáng kể thông tin cá nhân, chẳng hạn như đại diện lớp, thành viên của tập huấn luyện và thuộc tính của một tập con dữ liệu huấn luyện. Thậm chí người tấn công có thể suy ra các nhãn được đánh dấu cho dữ liệu từ các gradient được chia sẻ và khôi phục các mẫu trong tập huấn luyện ban đầu mà không cần biết trước về dữ liệu huấn luyện [30].

A. Suy diễn lớp đại diện

Hitaj lần đầu tiên đưa ra một cuộc tấn công suy luận tích cực được gọi là cuộc tấn công mạng sinh đối kháng (GAN) chống lại các mô hình học cộng tác. Trong cuộc tấn công này, một người tham gia đóng vai trò người tấn công, và cố ý làm tổn hại bất kỳ người tham gia nào khác. Cuộc tấn công GAN khai thác bản chất thời gian thực của quá trình học liên kết cho

phép người tấn công huấn luyện GAN để tạo ra các mẫu dữ liệu của dữ liệu huấn luyện được nhắm mục tiêu. Các mẫu được tạo dường như đến từ cùng một miền phân phối với dữ liệu huấn luyện. Do đó, tấn công GAN không nhằm mục đích tái tạo lại các đầu vào huấn luyện chính xác mà chỉ nhắm vào các đại diện của lớp huấn luyện. Cần lưu ý rằng tấn công GAN giả định toàn bộ dữ liệu huấn luyện cho một lớp nhất định đến từ một người tham gia duy nhất, có nghĩa là các đại diện do GAN xây dựng chỉ tương tự với dữ liệu huấn luyện khi tất cả các thành viên trong lớp đều giống nhau. Điều này giống như các cuộc tấn công đảo ngược mô hình trong cài đặt học máy tập trung [21]. Lưu ý rằng những giả định này có thể ít thực tế hơn trong học liên kết. Vì tấn công GAN yêu cầu một lượng đáng kể tài nguyên tính toán để huấn luyện mô hình GAN, nên nó ít phù hợp hơn với các tình huống H2C.

B. Suy diễn thành viên trong tập huấn luyện

Với một dữ liệu chính xác, các cuộc tấn công suy luận thành viên (membership inference attacks – MIA) nhằm xác định xem một dữ liệu có được sử dụng để huấn luyện mô hình hay không [22]. Ví dụ: người tấn công có thể suy ra liệu một hồ sơ bệnh nhân cụ thể có được sử dụng để huấn luyện tạo ra một bộ phân loại liên quan đến một căn bệnh nhất định hay không. Học liên kết mở ra những khả năng mới cho những cuộc tấn công như vậy. Trong học liên kết, người tấn công có thể suy luận xem một dữ liệu cụ thể có thuộc dữ liệu huấn luyện của một người tham gia cụ thể (nếu bản cập nhật đối tượng mục tiêu là từ một người tham gia duy nhất) hoặc bất kỳ người tham gia nào (nếu bản cập nhật tham số mô hình đối tượng mục tiêu được tổng hợp). Ví dụ, trong quá trình huấn luyện mô hình học liên kết, các gradient khác 0 của lớp trung gian một mô hình học sâu xử lý ngôn ngữ tự nhiên được huấn luyện trên dữ liệu văn bản có thể tiết lộ những từ nào có trong một lần lặp huấn luyện của những người tham gia trung thực [18].

C. Suy diễn thuộc tính

Người tấn công có thể thực hiện cả tấn công suy luận thuộc tính thụ động và chủ động để suy ra các thuộc tính nhất định của dữ liệu huấn luyện những người tham gia khác [18]. Các cuộc tấn công suy luận thuộc tính giả định rằng người tấn công có dữ liệu huấn luyện bổ trợ được gắn nhãn chính xác với thuộc tính mục tiêu. Người tấn công thụ động chỉ có thể quan sát các gradient và thực hiện suy luận bằng cách huấn luyện một bộ phân loại thuộc tính nhị phân. Một người tấn công tích cực có thể khai thác tính năng học đa tác vụ để đánh lừa mô hình học liên kết để học cách tách biệt giữa dữ liệu có và không có thuộc tính của đối tượng đích để lấy ra nhiều thông tin hơn. Người tham gia tấn công có thể suy diễn ra khi nào một thuộc tính tồn tại hoặc không tồn tại trong dữ liệu huấn luyện (ví dụ: xác định thời điểm một người xuất hiện lần đầu tiên trong các bức ảnh được sử dụng để huấn luyện bộ phân loại giới tính). Việc giả định dữ liệu huấn luyện bổ trợ trong các cuộc tấn công suy luận thuộc tính có thể hạn chế khả năng ứng dụng của nó trong mô hình H2C.

D. Suy ra nhãn dữ liệu của tập huấn luyện

Một nghiên cứu gần đây có tên Deep Leakage from Gradient (DLG) đề xuất một thuật toán tối ưu hóa để trích xuất cả dữ liệu huấn luyện và nhãn [26]. Cuộc tấn công này mạnh hơn nhiều so với các cách tiếp cận trước đây. Nó có thể khôi phục chính xác các hình ảnh và văn bản thô được sử dụng để huấn luyện mô hình học sâu. Trong một nghiên cứu tiếp theo, một phương pháp phân tích được gọi là rò rỉ thông tin học sâu cải thiện từ DLG (Improved Deep Leakage from Gradient – iDLG) đã được đề xuất để trích xuất các nhãn dựa trên các gradient được chia sẻ và khám phá mối tương quan giữa các nhãn và các dấu của giá trị gradient. iDLG có thể được áp dụng để tấn công bất kỳ mô hình nào được huấn luyện với hàm mất mát entropy và nhãn được gán theo mô hình mã hóa one-hot, đây là một cài đặt điển hình cho các tác vụ phân loại.

Tóm lại, các cuộc tấn công suy luận thường giả định rằng người tấn công sở hữu các mô hình học phức tạp và tài nguyên tính toán không giới hạn. Hơn nữa, hầu hết các cuộc tấn công đều giả định rằng những người tham gia tấn công có thể được chọn (để cập nhật mô hình toàn cục) trong nhiều vòng của quá trình huấn luyện học liên kết. Trong học liên kết, các giả định này thường không thực tế trong các kịch bản H2C, nhưng có nhiều khả năng xảy ra hơn trong các kịch bản H2B. Các cuộc tấn công suy luận này nhấn mạnh nhu cầu bảo vệ gradient trong học liên kết, có thể thông qua các cơ chế bảo vệ quyền riêng tư vi phân [15, 17].

3.2. Phương pháp bảo vệ

Bảng 3. Các kỹ thuật bảo vệ tính riêng tư trong học liên kết

Các kỹ thuật bảo vệ tính riêng tư		Các nghiên cứu
Mã hóa đồng hình		[20]
Riêng tư vi phân	Riêng tư vi phân tập trung	[10]
	Riêng tư vi phân cục bộ	[12], [15], [14]
	Riêng tư vi phân phân bố	[16], [17], [24]
Tính toán bảo mật đa bên		[5], [9]

Trong khi việc bảo vệ quyền riêng tư đã được nghiên cứu rộng rãi trong cộng đồng học máy, việc bảo vệ quyền riêng tư trong học tập liên kết có thể khó khăn hơn do khả năng truy cập rời rạc vào mô hình và kết nối mạng, sự không đồng nhất trong dữ liệu, v.v. Các nghiên cứu hiện có trong học liên kết bảo vệ quyền riêng tư là hầu hết được phát triển dựa trên các kỹ thuật bảo vệ quyền riêng tư phổ biến, bao gồm: (1) mã hóa đồng hình (homomorphic encryption – HE), trong tài liệu nghiên cứu [4]; (2) Tính toán đa bên an toàn (Secure Multiparty Computation – SMC), chẳng hạn giao thức mã hóa an toàn [5] và chia sẻ tính toán bí mật [9]; và (3) quyền riêng tư vi phân (differential privacy – DP) [4, 6, 17]. Một bản tổng hợp về các kỹ thuật bảo vệ quyền riêng tư được liệt kê trong bảng 3.

A. Bảo toàn quyền riêng tư thông qua mã hóa đồng hình

Một lược đồ mã hóa đồng hình cho phép các phép toán số học được thực hiện trực tiếp trên các dữ liệu mã hóa, tương đương với một phép toán đại số tuyến tính cụ thể của bản thô. Các kỹ thuật mã hóa đồng hình hiện tại có thể được phân loại thành: 1) mã hóa đồng hình hoàn toàn tương đồng, 2) một số mã hóa đồng hình tương đồng và 3) mã hóa đồng hình bán phần. Mã hóa đồng hình hoàn toàn tương đồng có thể hỗ trợ tính toán tùy ý trên các bản mã, nhưng kém hiệu quả hơn. Mặt khác, một số mã hóa đồng hình tương đồng và mã hóa đồng hình bán phần hiệu quả hơn nhưng được chỉ định bởi một số phép toán hạn chế. Các lược đồ mã hóa tương đồng một phần được sử dụng rộng rãi hơn trong thực tế, bao gồm RSA [4], v.v. Các thuộc tính đồng hình được mô tả như:

$$E_{pk}(m_1 + m_2) = c_1 \; c_2 \tag{1}$$
$$E_{pk}(\alpha \cdot m_1) = \alpha \; c_1$$

Trong đó a là hằng số, m_1, m_2 là dữ liệu thô cần mã hóa, c_1, c_2 là dữ liệu mã hóa của m_1, m_2 tương ứng. Mã hóa đồng hình đang được sử dụng rộng rãi và đặc biệt hữu ích để đảm bảo quá trình huấn luyện bằng cách tính toán trên dữ liệu được mã hóa.

B. Bảo mật tính riêng tư thông qua SMC

Tính toán đa bên an toàn (SMC-secure multiparty computation) cho phép những người tham gia khác nhau có dữ liệu đầu vào riêng tư thực hiện tính toán chung trên dữ liệu đầu vào huấn luyện của họ mà không tiết lộ cho nhau. Tài liệu nghiên cứu [5] đề xuất SecureML tiến hành huấn luyện mô hình đồng thời bảo vệ quyền riêng tư thông qua SMC, nơi chủ sở hữu dữ liệu cần xử lý, mã hóa và / hoặc chia sẻ bí mật dữ liệu của họ giữa hai máy chủ không thông đồng trong giai đoạn thiết lập ban đầu. SecureML cho phép chủ sở hữu dữ liệu đào tạo các mô hình khác nhau trên dữ liệu chung của họ mà không tiết lộ bất kỳ thông tin nào ngoài kết quả. Tuy nhiên, điều này đi kèm với chi phí tính toán và giao tiếp cao trong quá trình cộng tác huấn luyện mô hình. Bên cạnh đó một một giao thức an toàn, hiệu quả về giao tiếp và không bị lỗi dựa trên SMC để tổng hợp an toàn các gradient của người tham gia được đề xuất. Nó đảm bảo rằng thông tin dữ liệu cá nhân mà máy chủ muốn tìm hiểu và khai thác chỉ có thể thực hiện từ các kết quả tổng hợp. Tính bảo mật của giao thức được duy trì trong cả cài đặt trung thực gây tò mò và độc hại, ngay cả khi máy chủ và một nhóm nhỏ người dùng có hành động xấu, thông đồng. Có nghĩa là, không bên nào học được gì hơn là tổng các đầu vào của một tập hợp con những người dùng trung thực có số lượng quy mô người tham gia lớn [5].

Tóm lại, mã hóa đồng hình hoặc các phương pháp tiếp cận dựa trên SMC có thể áp dụng cho các tình huống học liên kết với quy mô lớn vì chúng phải chịu thêm chi phí giao tiếp và tính toán bổ sung đáng kể. Hơn nữa, các kỹ thuật dựa trên mã hóa cần được thiết kế và triển khai cẩn thận cho từng thao tác trong thuật toán học máy có mục tiêu [8]. Cuối cùng, tất cả các giao thức dựa trên mã hóa ngăn không cho bất kỳ ai kiểm tra các bản cập nhật của người tham gia đối với mô hình chung.

C. Bảo vệ tính riêng tư thông qua riêng tư vi phân

Riêng tư vi phân (DP-differential privacy) ban đầu được thiết kế cho cơ sở dữ liệu thống kê, trong đó đối với mỗi truy vấn được thực hiện, máy chủ sẽ trả lời truy vấn theo cách bảo vệ quyền riêng tư vi phân với kết quả chính xác của câu truy vấn cộng thêm sai số ngẫu nhiên phù hợp [8]. So với các phương pháp tiếp cận dựa trên mã hóa, riêng tư vi phân đánh đổi quyền riêng tư và độ chính xác bằng cách xáo trộn dữ liệu theo cách (i) hiệu quả về mặt tính toán, (ii) không cho phép kẻ tấn công khôi phục dữ liệu gốc và (iii) không làm ảnh hưởng đến độ hữu dụng của dữ liệu.

Khái niệm về riêng tư vi phân là ảnh hưởng của sự hiện diện có hoặc không có một bản ghi đối với kết quả truy vấn bị giới hạn bởi một sai số nhỏ ε. Định nghĩa riêng tư vi phân (ε; δ) – quyền riêng tư vi phân xấp xỉ [9] nới lỏng quyền riêng tư vi phân thuần túy bằng một thuật ngữ phụ δ, có nghĩa là các phản hồi truy vấn không chắc chắn không cần phải đáp ứng tiêu chí về quyền riêng tư vi phân.

Riêng tư vi phân (ε; δ) [9]: đối với các đại lượng vô hướng $\varepsilon > 0$ và $0 \leq \delta < 1$, kỹ thuật M được cho là bảo toàn (ε; δ) riêng tư vi phân nếu đối với tất cả các tập dữ liệu liền kề D; $D' \in D^n$ và các giá trị trong tập S được tạo ra với kỹ thuật M. Để tránh trường hợp xấu nhất luôn vi phạm quyền riêng tư do chọn δ quá nhỏ, giới hạn chuẩn là chọn $\delta << 1 / |D|$, trong đó $|D|$ là kích thước của cơ sở dữ liệu.

Riêng tư vi phân thường được phân loại thành ba phần theo các giả định tin cậy và nguồn làm nhiễu khác nhau: riêng tư vi phân tập trung (Centralized differential privacy – CDP), riêng tư vi phân cục bộ (Local differential privacy – LDP) và riêng tư vi phân phân tán (Distributed differential privacy – DDP) trong Bảng 4.

Riêng tư vi phân tập trung (CDP) ban đầu được thiết kế một máy chủ cơ sở dữ liệu đáng tin cậy, và có quyền xem tất cả dữ liệu của người tham gia một cách rõ ràng, đồng thời trả lời các truy vấn hoặc công khai các thống kê trong cơ sở dữ liệu theo cách bảo vệ quyền riêng tư bằng cách ngẫu nhiên hóa các kết quả truy vấn [10]. Khi CDP được sử dụng trong mô hình liên kết, CDP giả định một bộ tổng hợp đáng tin cậy, và chịu trách nhiệm thêm nhiễu vào các

gradient cục bộ tổng hợp để đảm bảo quyền riêng tư ở mức bản ghi đối với toàn bộ dữ liệu của tất cả những người tham gia [17]. Tuy nhiên, CDP hướng đến việc giải quyết hàng nghìn người dùng tham gia huấn luyện để đạt được sự hội tụ và sự cân bằng có thể chấp nhận giữa quyền riêng tư và độ chính xác. Dẫn đến, CDP chỉ có thể đạt được độ chính xác chấp nhận được với một số lượng lớn người tham gia, do đó không thể áp dụng cho H2B có số lượng người tham gia tương đối nhỏ.

Trong khi đó, giả định về một máy chủ đáng tin cậy trong CDP là không phù hợp trong nhiều ứng dụng vì nó tạo thành một điểm lỗi duy nhất cho các vi phạm dữ liệu và đảm bảo cho người quản lý đáng tin cậy các nghĩa vụ pháp lý và đạo đức để giữ an toàn cho dữ liệu của người dùng. Khi bộ tổng hợp ở máy chủ không đáng tin cậy thường xảy ra trong các kịch bản phân tán dữ liệu, thì cần có riêng tư vi phân cục bộ (LDP) [12] hoặc riêng tư vi phân phân tán (DDP) [16] để bảo vệ quyền riêng tư của người tham gia.

Bảng 4. So sánh giữa riêng tư vi phân tập trung, riêng tư vi phân cục bộ, và riêng tư vi phân phân bố

Loại riêng tư vi phân	Máy chủ tin cậy	Đối tượng thêm nhiễu	Mức độ đảm bảo riêng tư
Riêng tư vi phân tập trung	Có	Máy chủ	Tham số đã tổng hợp
Riêng tư vi phân cục bộ	Không	Người tham gia	Tham số cục bộ
Riêng tư vi phân phân bố	Không	Người tham gia	Tham số đã tổng hợp

Quyền riêng tư vi phân cục bộ (LDP) [12] cung cấp đảm bảo quyền riêng tư mạnh mẽ hơn cho chủ sở hữu dữ liệu bằng cách xáo trộn thông tin cá nhân của họ để đáp ứng riêng tư vi phân cục bộ trước khi gửi kết quả cho máy chủ không đáng tin cậy [14, 15].

Riêng tư vi phân cục bộ ($\sum$; δ): thuật toán ngẫu nhiên M thỏa mãn riêng tư vi phân ($\sum$; δ) nếu và chỉ nếu khi với bất kỳ đầu vào v và v' chúng ta đều có:

$$\Pr\{M(v) = o\} \leq \exp(\varepsilon) \cdot \Pr\{M(v') = o\} + \delta \qquad (2)$$

Cho $\forall o \in$ Range (M), trong đó Range (M) biểu diễn tập hợp tất cả các đầu ra có thể có của thuật toán M. Hơn nữa M được cho là bảo toàn ε-LDP nếu điều kiện giữ cho $\delta = 0$.

Mặc dù phản hồi ngẫu nhiên [10] và các biến thể của nó [11] đã được sử dụng rộng rãi để cung cấp LDP khi các cá nhân tiết lộ thông tin cá nhân của họ. Tất cả các cơ chế ngẫu nhiên được sử dụng cho CDP, chẳng hạn như cơ chế Laplace và cơ chế Gaussian [9], có thể được sử dụng riêng lẻ bởi từng người tham gia để đảm bảo LDP. Tuy nhiên, trong kịch bản phân tán, không có sự trợ giúp của các kỹ thuật mã hóa, mỗi người tham gia phải thêm đủ nhiễu vào dữ liệu đã hiệu chỉnh để đảm bảo LDP trước khi gửi về máy chủ. Do đó, các thuộc tính quyền riêng tư của LDP đi kèm với sự suy giảm độ hữu dụng rất lớn, đặc biệt là với mô hình có số lượng hàng tỷ người tham gia.

Một số nghiên cứu trước đây đã cố gắng áp dụng LDP cho học liên kết. Ví dụ, Shokri [19] lần đầu tiên áp dụng LDP cho học liên kết, trong đó mỗi người tham gia thêm nhiễu vào các gradient của nó trước khi gửi đến máy chủ, do đó đảm bảo riêng tư vi phân cục bộ. Tuy nhiên, các giới hạn về quyền riêng tư của họ được quy định cho mỗi tham số, số lượng lớn các tham số ngăn cản phương pháp của họ cung cấp một đảm bảo quyền riêng tư có ý nghĩa [20]. Các cách tiếp cận khác cũng được coi là áp dụng LDP cho học liên kết chỉ có thể hỗ trợ các mô hình hồi quy logistic và chỉ tập trung vào các nhiệm vụ và bộ dữ liệu đơn giản [16, 17]. Melis đã trình bày một cách tiếp cận khả thi để huấn luyện mô hình riêng tư cục bộ với quy mô lớn [25]. Tác giả thực hiện thực nghiệm với hơn 200 vòng lặp huấn luyện và phải chịu chi phí bảo mật cao hơn nhiều, với bộ dữ liệu MNIST ($\varepsilon = 500$) và CIFAR-10 ($\varepsilon = 5000$). Kết quả

thu được của họ cho thấy rằng việc sử dụng các cơ chế LDP vẫn có thể bảo vệ tốt chống lại việc tái xây dựng mô hình.

Ngoài việc rò rỉ tính riêng tư trong học liên kết với các kiến trúc đồng nhất, học liên kết với các kiến trúc không đồng nhất cũng gặp phải các vấn đề về quyền riêng tư tương tự. Trong học liên kết với các kiến trúc đồng nhất, các dự đoán từ các mô hình cục bộ cũng chứa các thông tin nhạy cảm và có thể làm rò rỉ thông tin riêng tư [21, 23]. Hiện tại, không có gì đảm bảo về mặt lý thuyết rằng việc chia sẻ dự đoán là riêng tư và an toàn [23]. Để giải quyết vấn đề này, một cách tiếp cận đơn giản là thêm nhiễu ngẫu nhiên riêng tư vi phân cục bộ vào các dự đoán như các nghiên cứu trước. Mặc dù mối quan tâm về quyền riêng tư được giảm thiểu với sự thêm nhiễu ngẫu nhiên, nhưng nó lại mang đến một vấn đề mới với sự cân bằng đáng kể giữa bảo mật và độ hữu dụng của mô hình. Wang [23] đã đưa ra giải pháp giải quyết vấn đề bằng cách đề xuất một mô hình mới có tên FEDMD-NFDP, mô hình này tích hợp một cơ chế bảo mật riêng tư vi phân với việc tùy chỉnh nhiễu (Noise-Free Differential Privacy – NFDP) vào quá trình huấn luyện mô hình liên kết. Sự đảm bảo của LDP về nguồn gốc NFDP trong quá trình lấy mẫu dữ liệu cục bộ, giúp loại bỏ rõ ràng việc thêm nhiễu và các vấn đề bùng nổ chi phí bảo mật trong các nghiên cứu trước đó.

Riêng tư vi phân phân tán (DDP) nằm giữa LDP và CDP, đồng thời đảm bảo quyền riêng tư của mỗi cá nhân bằng cách kết hợp với các giao thức mật mã [17], [2]. Do đó, DDP dùng cho trường hợp không đặt niềm tin vào bất kỳ máy chủ nào và cung cấp độ hữu dụng tốt hơn LDP. Về mặt lý thuyết, DDP cung cấp độ hữu dụng tương tự như CDP, vì tổng lượng nhiễu là như nhau. Khái niệm DDP phản ánh thực tế là nhiễu cần thiết trong thống kê mục tiêu được lấy từ nhiều người tham gia [17]. Các phương pháp tiếp cận DDP thực hiện một cơ chế nhiễu thêm vào tổng thể bằng cách tổng hợp cùng một cơ chế thực hiện ở mỗi người tham gia (thường là ít nhiễu hơn) các kỹ thuật cần thiết có phân phối ổn định và mã hóa để ẩn tất cả nhưng kết quả cuối cùng từ những người tham gia [2, 10, 17]. Các phân phối ổn định bao gồm phân phối Gaussian, phân phối nhị thức, v.v., tức là tổng các biến ngẫu nhiên Gaussian vẫn tuân theo phân phối Gauss và tổng các biến ngẫu nhiên nhị thức vẫn tuân theo phân phối nhị thức. DDP sử dụng sự ổn định này để cho phép mỗi người tham gia thêm nhiễu ngẫu nhiên vào các thống kê cục bộ của mình ở một mức độ thấp hơn so với LDP. Tuy nhiên, trong DDP, tổng thống kê của tất cả riêng tư vi phân thỏa mãn riêng tư vi phân (ε; δ) chứ không phải thông tin riêng tư vi phân của từng đối tượng tham gia, tức là Σr_i là thỏa mãn mức độ riêng tư vi phân, nhưng nhiễu của từng cá nhân thì không thỏa mãn, do đó $x_i + r_i$ không thể được gửi trực tiếp đến máy chủ. Ở đây x_i chỉ ra dữ liệu thô và r_i là thông tin nhiễu được thêm vào. Do đó, DDP cần có sự trợ giúp của SMC để duy trì độ hữu dụng và đảm bảo tính không biết của bộ tổng hợp, được chứng minh trong [2, 17].

Một công việc song song khác để huấn luyện mô hình phân tán bảo vệ quyền riêng tư là chuyển kiến thức về tổng thể của nhiều mô hình sang mô hình *student* [13, 17, 18]. Ví dụ, Fang [11] lần đầu tiên tạo dữ liệu có nhãn từ dữ liệu phụ không gán nhãn, sau đó sử dụng dữ liệu phụ được gán nhãn để tìm bộ giảm thiểu rủi ro theo thực nghiệm, cuối cùng đã phát hành bộ phân loại thỏa mãn riêng tư vi phân sử dụng nhiễu đầu ra [7]. Tương tự, Papernot [18] đề xuất Private Aggregation of Teacher Ensembles (PATE), mô hình này đầu tiên huấn luyện một nhóm *teacher* trên các tập dữ liệu con cá nhân rời rạc, sau đó điều chỉnh mô hình của nhóm *teacher* bằng cách thêm nhiễu vào lựa chọn mô hình ở giáo viên tổng hợp trước khi chuyển mô hình cho *student*. Cuối cùng, một mô hình *student* được huấn luyện dựa trên đầu ra tổng hợp của nhóm *teacher*. PATE yêu cầu nhiều người tham gia đạt được độ chính xác hợp lý và mỗi người tham gia cần có đủ dữ liệu để huấn luyện một mô hình chính xác. Mô hình này không được lưu trong hệ thống học liên kết, nơi mà việc phân phối dữ liệu của những người tham gia có thể bị mất cân bằng, làm cho đề xuất PATE không phù hợp với hệ thống học liên kết.

3.3. Thảo luận và hướng nghiên cứu trong tương lai

Vẫn còn những lỗ hổng tiềm ẩn cần được giải quyết để cải thiện tính riêng tư và tính bảo mật của hệ thống học liên kết. Hơn nữa, có nhiều mục tiêu thiết kế quan trọng như nhau về tính riêng tư và tính bảo mật, do đó cần được xem xét đồng thời trong học liên kết. Trong phần này, các hướng nghiên cứu có triển vọng được giới thiệu.

Các cuộc tấn công về quyền riêng tư hiện tại có một số điểm yếu cố hữu có thể hạn chế khả năng áp dụng của chúng trong học liên kết [16]. Ví dụ, tấn công GAN giả định rằng toàn bộ kho dữ liệu huấn luyện cho một lớp nhất định đến từ một người tham gia duy nhất và chỉ trong trường hợp đặc biệt khi tất cả các thành viên trong lớp đều giống nhau, các đại diện do GAN xây dựng tương tự với dữ liệu huấn luyện. Những giả định này có thể ít thực tế hơn trong học liên kết. Đối với DLG và iDLG [26], cả hai đều hoạt động: (1) áp dụng phương pháp tối ưu hóa bậc hai gọi là Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), tốn kém hơn về mặt tính toán so với các phương pháp tối ưu hóa bậc một; (2) chỉ áp dụng cho các gradient được tính toán trên các tập dữ liệu nhỏ, tức là tối đa số lượng bó là 8 trong DLG và số lượng bó là 1 trong iDLG, không phải là trường hợp thực tế đối với FL, trong đó gradient thường được chia sẻ sau khi ít nhất 1 lần huấn luyện tại người tham gia; (3) sử dụng mô hình chưa được huấn luyện, bỏ qua các gradient trong nhiều vòng giao tiếp. Tấn công hệ thống học liên kết theo cách hiệu quả hơn và trong các cài đặt thực tế phần lớn vẫn chưa được khám phá.

Xem xét lại các biện pháp phòng thủ hiện tại: học liên kết với tổng hợp an toàn cho mục đích riêng tư dễ bị tấn công nhiễu độc hơn vì không thể kiểm tra các bản cập nhật riêng lẻ. Tương tự, vẫn chưa rõ liệu huấn luyện mô hình tấn công, một phương pháp phòng thủ hiện đại chống lại các cuộc tấn công của đối phương trong học máy thông thường [23], có thể được điều chỉnh cho phù hợp với học liên kết hay không, vì quá trình huấn luyện tấn công đã được phát triển chủ yếu cho dữ liệu phân bố định danh và độc lập (IID) và vẫn chưa rõ ràng về hiệu suất của nó trong các cài đặt dữ liệu phân bố không định danh và độc lập (Non-IID). Hơn nữa, huấn luyện mô hình tấn công rất tốn kém về mặt tính toán và có thể ảnh hưởng đến hiệu suất [24], điều này có thể không khả thi đối với kịch bản H2C. Xét về các phương pháp dựa trên riêng tư vi phân [20], riêng tư vi phân cung cấp cơ chế giới hạn sự thành công của suy luận thành viên, nhưng không ngăn cản tấn công suy luận thuộc tính cho một nhóm huấn luyện [18]. Mặt khác, riêng tư vi phân ở cấp độ người tham gia được hướng tới làm việc với hàng nghìn người dùng để huấn luyện nhằm hội tụ và đạt được sự đánh đổi có thể chấp nhận được giữa quyền riêng tư và độ chính xác. Mô hình học liên kết không hội tụ được với một số lượng nhỏ người tham gia, khiến nó không phù hợp với các kịch bản H2B. Hơn nữa, riêng tư vi phân có thể làm ảnh hưởng đến độ chính xác của mô hình đã học [19]. Cần phải tìm hiểu thêm để xem xét liệu riêng tư vi phân ở cấp độ người tham gia có thể bảo vệ các hệ thống học liên kết với ít người tham gia hay không.

Quyền riêng tư giai đoạn suy diễn trong học liên kết: các cuộc tấn công và phòng thủ trong giai đoạn huấn luyện trong học liên kết được tập trung giới thiệu, xem xét các khả năng tấn công nhiều hơn được mở ra bởi thuộc tính huấn luyện phân tán của các hệ thống học liên kết. Thực tế, học liên kết cũng dễ bị tấn công cả về tính riêng tư và bảo mật trong giai đoạn suy luận bởi người dùng cuối của mô hình học liên kết khi được triển khai như một dịch vụ.

Về lỗ hổng bảo mật, mô hình toàn cục được huấn luyện có thể tiết lộ thông tin nhạy cảm từ các dự đoán của mô hình khi được triển khai dưới dạng dịch vụ, gây rò rỉ quyền riêng tư. Trong thiết lập này, người tấn công không có quyền truy cập trực tiếp vào các thông số mô hình, nhưng có thể xem các cặp đầu vào-đầu ra. Các nghiên cứu trước đây đã chỉ ra một loạt vụ rò rỉ quyền riêng tư chỉ cho phép truy cập mô hình học theo hộp đen vào các mô hình đã được huấn luyện, chẳng hạn như (1) các cuộc tấn công ăn cắp mô hình trong đó các tham số mô hình có thể được xây dựng lại bởi người tấn công chỉ có quyền truy cập vào API suy luận

/ dự đoán dựa trên các tham số [24]; (2) các cuộc tấn công suy luận thành viên nhằm xác định xem một bản ghi cụ thể có được sử dụng để huấn luyện mô hình hay không [22]. Các mô hình học liên kết phải đối mặt với tình trạng tương tự trong quá trình triển khai mô hình cho mục đích suy luận. Sự phát triển của các biện pháp bảo vệ chống lại sự rò rỉ quyền riêng tư trong quá trình triển khai mô hình đòi hỏi cần các nghiên cứu kỹ hơn.

4. KẾT LUẬN

Mặc dù học liên kết vẫn còn sơ khai, nó sẽ tiếp tục phát triển mạnh và sẽ là một lĩnh vực nghiên cứu tích cực và quan trọng trong tương lai. Khi học liên kết phát triển, các mối đe dọa về quyền riêng tư và tính bảo mật đối với học liên kết cũng vậy. Điều quan trọng là cung cấp một cái nhìn tổng thể về các cuộc tấn công và phòng thủ hiện tại trên học liên kết để các nhà thiết kế hệ thống học liên kết nhận thức rõ về các lỗ hổng tiềm ẩn trong các thiết kế hiện tại và giúp họ hiểu và thực hiện việc triển khai học liên kết trong thế giới thực được thuận lợi hơn. Cuộc khảo sát này đóng vai trò là một tổng quan ngắn gọn và dễ tiếp cận về chủ đề bảo vệ tính riêng tư trong mô hình học liên kết, đồng thời sẽ giúp ích rất nhiều cho sự hiểu biết về bối cảnh tấn công và phòng thủ quyền riêng tư và tính bảo mật trong học liên kết. Mục tiêu cuối cùng của việc phát triển một cơ chế bảo vệ học liên kết có mục đích chống lại các cuộc tấn công khác nhau mà không làm giảm hiệu suất của mô hình sẽ đòi hỏi nỗ từ nhiều nghiên cứu ở các góc độ và khía cạnh khác nhau.

TÀI LIỆU THAM KHẢO

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L. - Deep learning with differential privacy, Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (2016) 308-318.
2. Agarwal, N., Suresh, A.T., Yu, F.X.X., Kumar, S. and McMahan, B. - cpSGD: Communication-efficient and differentially-private distributed SGD, Advances in Neural Information Processing Systems **31** (2018) 7575–7586
3. Andrew, G., Thakkar, O., McMahan, B. and Ramaswamy, S. - Differentially private learning with adaptive clipping, Advances in Neural Information Processing Systems **34** (2021) 17455-17466
4. Aono, Y., Hayashi, T., Wang, L. and Moriai, S. - Privacy-preserving deep learning via additively homomorphic encryption, IEEE Transactions on Information Forensics and Security, 13(5) (2017) 1333-1345.
5. Barreno, M., Nelson, B., Sears, R., Joseph, A.D. and Tygar, J.D. - Can machine learning be secure?. Proceedings of the ACM Symposium on Information, computer and communications security (2006) 16-25.
6. Chaudhuri, K., Monteleoni, C. and Sarwate, A.D. - Differentially private empirical risk minimization, Journal of Machine Learning Research **12** (3) (2011) 1069-1109.
7. Gentry, C. - Fully homomorphic encryption using ideal lattices. In Proceedings of the 41st annual ACM symposium on Theory of computing (2009) 169-178.
8. Dwork, C. and Roth, A. - The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. **9** (3-4) (2014) 211-407.
9. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M. - Our data, ourselves: Privacy via distributed noise generation, Annual international conference on the theory and applications of cryptographic techniques (2006) 486-503.

10. Erlingsson, Ú., Pihur, V. and Korolova, A. - Rappor: Randomized aggregatable privacy-preserving ordinal response, Proceedings of the 21st ACM SIGSAC conference on computer and communications security (2014) 1054-1067.
11. Fang, M., Cao, X., Jia, J. and Gong, N. - Local Model Poisoning Attacks to Byzantine-Robust Federated Learning, In the 29th USENIX Security Symposium (USENIX Security 20) (2020) 1605-1622.
12. Ha, T., Dang, T.K., Dang, T.T., Truong, T.A. and Nguyen, M.T. - Differential privacy in deep learning: an overview, Proceeding of the International Conference on Advanced Computing and Applications (ACOMP) (2019) 97-102.
13. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R. and D'Oliveira, R.G. - Advances and open problems in federated learning, Foundations and Trends® in Machine Learning **14** (1–2) (2021) 1-210.
14. Liu, Y., Kang, Y., Xing, C., Chen, T. and Yang, Q. - A secure federated transfer learning framework, IEEE Intelligent Systems **35** (4) (2020) 70-82.
15. Lyu, L. - Lightweight crypto-assisted distributed differential privacy for privacy-preserving distributed learning, International Joint Conference on Neural Networks (IJCNN) (2020) 1-8.
16. Melis, L., Song, C., De Cristofaro, E. and Shmatikov, V. - Exploiting unintended feature leakage in collaborative learning, IEEE Symposium on Security and Privacy (SP) (2019) 691-706.
17. Papernot, N., Abadi, M., Erlingsson, Ú., Goodfellow, I. and Talwar, K. - Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data, International Conference on Learning Representations (2017) 3-19.
18. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. and Erlingsson, Ú. - Scalable private learning with pate, International Conference on Learning Representations (2018) 24-58.
19. Shokri, R., Stronati, M., Song, C. and Shmatikov, V. - Membership inference attacks against machine learning models, IEEE symposium on security and privacy (SP) (2017) 3-18.
20. Sun, L. and Lyu, L. - Federated model distillation with noise-free differential privacy. The proceedings of the 30th International Joint Conference on Artificial Intelligence (2020) 1563-1570.
21. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K. and Ristenpart, T. - Stealing Machine Learning Models via Prediction APIs. In 25th USENIX security symposium (2016) 601-618.
22. Truex, S., Liu, L., Chow, K.H., Gursoy, M.E. and Wei, W. - LDP-Fed: Federated learning with local differential privacy, Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking (2020) 61-66.
23. Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B. and Gu, Q. - On the Convergence and Robustness of Adversarial Training, Proceeding of the 36th International Conference on Machine Learning (2019) 6586-6595
24. Yin, D., Chen, Y., Kannan, R. and Bartlett, P. - Byzantine-robust distributed learning: Towards optimal statistical rates, International Conference on Machine Learning (2018) 5650-5659.

25. Zhao, Y., Zhao, J., Yang, M., Wang, T., Wang, N., Lyu, L., Niyato, D. and Lam, K.Y. - Local differential privacy-based federated learning for internet of things, IEEE Internet of Things Journal **8** (11) (2020) 8836-8853.
26. Zhu, L., Liu, Z. and Han, S. - Deep leakage from gradients. Advances in Neural Information Processing Systems **32** (2019) 14774-14784.

ABSTRACT

THREATS AND DEFENSES OF PRIVACY IN THE FEDERATED LEARNING

Ha Le Hoai Trung[1], Dang Tran Khanh[2*]
[1]*University of Information Technology, VNU-HCM, Vietnam*
[2]*Ho Chi Minh City University of Food Industry, Vietnam*
[*]Email: *khanh@hufi.edu.vn*

With scientific development, people have a more modern and comfortable life, and also create more data. This data is stored in different devices and application domains and society is becoming more and more aware of data privacy issues. The traditional centralized training or traditional artificial intelligence (AI) models are facing efficiency and privacy challenges. In recent years, federated learning has emerged as an alternative solution and continues to thrive in the field of artificial intelligence for responding to the demands of everyday life. Existing federated learning models have been shown to be vulnerable to attackers within or outside of the system, affecting data privacy and system security. Besides training global models, it is of paramount importance to design federated learning systems that have privacy guarantees and are resistant to different types of attacks. In this study, a comprehensive survey on privacy in federated learning is presented. Through a brief introduction to the concept of federated learning, its classification includes: 1) threats models; 2) privacy attacks and defenses. Key techniques and basic assumptions adopted by various attacks and defenses in federated learning are also introduced to help better understand the nature and conditions of attacks. Finally, future research directions to protect privacy in federated learning models are discussed in detail.

Keywords: Federated learning, generative adversarial network, attack, defense, privacy.