Học sâu và ứng dụng phương pháp học sâu có đảm bảo tính riêng tư

Ngày nay, Trí tuệ nhân tạo (AI) hiện diện trong mọi lĩnh vực của đời sống con người, từ kinh tế, giáo dục, y khoa cho đến những công việc nhà, giải trí hay thậm chí là trong quân sự. Học máy là một ứng dụng của trí tuệ nhân tạo cung cấp cho các hệ thống khả năng tự động học hỏi và cải thiện từ kinh nghiệm mà không cần lập trình rõ ràng. Học máy tập trung vào việc phát triển các chương trình máy tính có thể truy cập dữ liệu và sử dụng nó để tự học. Do đó, vấn đề đảm bảo tính riêng tư trong ứng dụng phương pháp học sâu đang là một vấn đề được quan tâm hiện nay.

Lý thuyết về học sâu

Những năm gần đây, khi khả năng tính toán của các máy tính được nâng lên một tầm cao mới với lượng dữ liệu khổng lồ được thu thập thì học máy đã tiến thêm một bước dài, dẫn đến việc ra đời một lĩnh vực mới được gọi là học sâu.

Học sâu là một nhánh của ngành học máy dựa trên một tập hợp các thuật toán để cố gắng mô hình hóa dữ liệu trừu tượng ở mức cao bằng cách sử dụng nhiều lớp xử lý với cấu trúc phức tạp, hoặc bằng cách khác bao gồm nhiều biến đổi phi tuyến, được lấy cảm hứng từ cấu trúc và chức năng của bộ não được gọi là mạng thần kinh nhân tạo.

Một mạng thần kinh nhân tạo bao gồm ba lớp chính, đó là: lớp đầu vào, lớp ẩn và lớp đầu ra với một số mô hình học sâu (Mạng nơ ron tích chập (Convolutional Neural Network - CNN) và Mạng nơ ron hồi quy (Recurrent Neural Network - RNN)).

Đảm bảo tính riêng tư cho học sâu

Bài toán bảo vệ tính riêng tư cho học máy đã được nghiên cứu rộng rãi bởi cộng đồng khai thác dữ liệu trong những năm gần đây. Để đảm bảo tính riêng tư cho học máy nói chung và cho mô hình học sâu phân tán nói riêng có thể thực hiện theo các phương pháp khác nhau, mỗi phương pháp sẽ có những ưu, nhược điểm riêng của nó.

Tuy nhiên, các phương pháp này luôn tồn tại một sự đánh đổi cố hữu giữa tính đúng đắn của tính toán, tính riêng tư của những dữ liệu nhạy cảm và tính hiệu quả của giải pháp. Việc lựa chọn phương pháp nào phù hợp sẽ phụ thuộc vào mục tiêu của bài toán cần xử lý. Các giải pháp học máy đảm bảo tính riêng tư dựa trên tính toán bảo mật nhiều thành viên thường đảm bảo được độ chính xác và bảo vệ được các thông tin riêng tư, nhạy cảm trong dữ liệu của mỗi người dùng [1].

Tính toán bảo mật nhiều thành viên

Tính toán bảo mật (Sercure Computation - SC), tính toán nhiều bên (Multi-party Computation - MPC) hay tính toán bảo mật nhiều thành viên (Secure Multi-party Computation - SMC) là một lĩnh vực của mật mã với mục tiêu tạo ra các phương thức cho phép các bên cùng tính toán một hàm dựa trên các giá trị đầu vào của họ mà vẫn đảm bảo tính riêng tư của những giá trị đầu vào này.

Để thực hiện giao thức SMC chỉ cần mỗi bên tham gia có một máy tính đáng tin cậy để chạy phần giao thức của mình và cách (có thể không an toàn) để giao tiếp với các bên tham gia khác. Giao thức bao gồm một loạt các thông điệp được trao đổi giữa những bên tham gia và cuối cùng mỗi bên tham gia tìm hiểu đầu ra của giao thức. Bản thân giao thức là công khai, cho phép mỗi bên tham gia xác minh độc lập rằng phần mềm chạy trên máy của chính họ là hợp lệ [1].

Các giao thức tính toán bảo mật nhiều thành viên cho độ an toàn cao và đảm bảo được mức độ riêng tư mạnh. Tuy nhiên, những vấn đề về hiệu năng đang cản trở sự phát triển của các giao thức này.

Để làm rõ độ an toàn và mức độ đảm bảo riêng tư mạnh của phương pháp này, tác giả trình bày một giao thức học sâu có đảm bảo tính riêng tư hiệu quả dựa trên phương pháp tính toán bảo mật nhiều thành viên dựa trên giao thức tính tổng bảo mật cho bài toán an toàn thông tin phát hiện thư rác và tiến hành thử nghiệm.

Giao thức học sâu có đảm bảo tính riêng tư hiệu quả dựa trên tính toán bảo mật nhiều thành viên

Trong mô hình huấn luyện mạng học sâu phân tán, cần định nghĩa bài toán đảm bảo tính riêng tư cho mô hình này.

Có N bên $\mathcal{P} = \{P1, P2, \ldots, PN\}$ tham gia huấn luyện mô hình, trong đó mỗi bên sở hữu một bộ dữ liệu huấn luyện riêng tư tương ứng $D1, D2, \ldots, DN$. Các bên này muốn kết hợp để thực hiện việc huấn luyện một mô hình chung tổng quát mà không tiết lộ các thông tin cục bộ của mình bao gồm:

- Các thuộc tính của dữ liệu đầu vào;
- Nhãn dữ liệu đầu ra hoặc phản hồi của mô hình với dữ liệu;
- Kiến trúc chi tiết của mô hình bao gồm: kiến trúc mạng, tham số, các hàm mất mát;
- Thông tin định danh về đóng góp của một bên dữ liệu đối với một bản ghi cụ thể.

Để làm được điều này, các bên cần xây dựng và thực thi một giao thức an toàn π. Trong bài toán đặt ra, tác giả trình bày giao thức huấn luyện mạng học sâu phân tán sử dụng giao thức tính tổng bảo mật an toàn.

Giao thức tính tổng bảo mật cho vector số thực

Như đã biết, hầu hết các giao thức tính tổng hoặc rộng hơn là các giao thức tính toán bảo mật nhiều thành viên sử dụng mật mã đều có độ phức tạp rất lớn và chỉ làm việc với số nguyên lớn trên các trường hữu hạn.

Để giải quyết bài toán này, bài báo trình bày một giao thức tính tổng bảo mật cho vector với các thành phần số thực. Giao thức này có khả năng tính tổng bảo mật cho vector số thực mà không làm tiết lộ các thành phần của các vector của các bên tham gia.

Giao thức dựa trên độ khó của bài toán logarithm rời rạc. Các tham số của giao thức yêu cầu bao gồm:

Các tham số công khai: H, p, g là các tham số công khai được tất cả các bên tham gia biết.

Các tham số bí mật: Mỗi bên P_i sở hữu các vector bí mật W_i đại diện cho tất cả các tham số của mô hình cục bộ. Bên Pi này cũng chọn một ma trận ngẫu nhiên $r_i \in Z$ có các thành phần tương ứng $r_i^{(jk)}$để làm ma trận mặt nạ sẽ được loại bỏ ở quá trình tổng hợp tại máy chủ.

Giao thức gồm hai pha thực thi chính:

Pha khởi tạo

Trước khi bắt đầu thực hiện giao thức, các bên tham gia $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$ gửi các ma trận khóa công khai: $M_i = \{X_i, Y_i\}$ lên máy chủ tổng hợp S. Sau đó, máy chủ tổng hợp S tính trước các ma trận X, Y có các thành phần $X^{(jk)}$, $Y^{(jk)}$ tương ứng:

$$X^{(jk)} = \prod_{i=1}^{N} X_i^{(jk)}; Y = \prod_{i=1}^{N} Y_i^{(jk)} \quad (1)$$

và gửi lại ma trận công khai $M = \{X, Y\}$ tới toàn bộ các bên tham gia.

Pha tính tổng bảo mật

Sau pha khởi tạo, tất cả các bên tham gia đã có đủ các tham số cần thiết để thực hiện giao thức.

Bước 1. Mỗi bên tham gia thực hiện tính giá trị mặt nạ công khai R_i dựa trên mặt nạ bí mật r_i;

Bước 2. Mỗi bên tham gia tính ma trận $T_i = V_i H + r_i$;

Bước 3. Gửi $M = (R_i, T_i)$ đến máy chủ tổng hợp;

Bước 4. Tại máy chủ tổng hợp, máy chủ thực hiện việc tính ma trận M_s. Từ ma trận này, máy chủ tổng hợp sẽ tính được ma trận tổng của mặt nạ Q sử dụng thuật toán Bước lớn bước nhỏ [6];

Bước 5. Máy chủ tổng hợp tính ma trận $T = \sum_{i=1}^{N} T_i$;

Bước 6. Máy chủ tính ma trận kết quả $V = \sum_{i=1}^{N} V_i = (T - Q)H^{-1}$.

Phân tích giao thức

Việc phân tích giao thức, bài báo thực hiện đánh giá và chứng minh tính đúng đắn của giao thức đối với bài báo tính tổng bảo mật của ma trận số thực, đồng thời đưa ra một số đánh giá về tính an toàn cũng như hiệu năng của giao thức đề xuất về mặt lý thuyết.

Chứng minh tính đúng đắn

Bài báo chứng minh rằng kết quả đầu ra cuối cùng V sau khi thực hiện giao thức sẽ đúng là tổng của các ma trận riêng tư của các bên tham gia bán trung thực.

$$T = \sum_{i=1}^{N} T_i = \sum_{i=1}^{N} (V_i H + r_i) = \sum_{i=1}^{N} (V_i H) + \sum_{i=1}^{N} r_i \quad (2)$$

Thực hiện chứng minh $Q = \sum_{i=1}^{N} r_i$.

$$M_s = \prod_{i=1}^{N} R_i = g^{\sum_{i=1}^{N} r_i}$$

nếu tìm được $Q^{(jk)}$ thỏa mãn $g^{Q^{(jk)}} mod\ p = M_s^{(jk)}$ thì $Q^{(jk)} = \sum_{i=1}^{N} r_i^{(jk)}$, hay nói cách khác, $Q = \sum_{i=1}^{N} r_i$

Từ công thức 3.2:

$$T = \sum_{i=1}^{N} (V_i H) + Q = \left(\sum_{i=1}^{N} V_i \right) H + Q \quad (3)$$

Từ công thức 3.3:

$$V = (T - Q)H^{-1} = \left(\sum_{i=1}^{N} V_i \right) H H^{-1} = \sum_{i=1}^{N} V_i \quad (4)$$

Cần phải chứng minh ma trận kết quả là tổng của các ma trận bí mật thành phần hay giao thức thực hiện đúng mục tiêu đề ra.

Phân tích tính an toàn

Bài báo đã giả sử các bên tham gia là các bên bán tin cậy và tuân thủ đầy đủ giao thức, cũng như các điều kiện giả thiết ở trên được thỏa mãn.

Mệnh đề 3.1. Giao thức trên là an toàn với bất kỳ bên bán tin cậy nào trên mạng công khai.

Chứng minh: Trong pha khởi tạo, mỗi bên i gửi các ma trận khóa công khai X_i và Y_i với mỗi phần tử được tính từ các khóa bí mật được chọn ngẫu nhiên. Trong pha tính tổng bảo mật, mỗi bên chỉ gửi các thông điệp R_i, T_i.

Do tính chất và giả thiết về độ khó của bài toán logarith rời rạc, sẽ rất khó để tìm lại được các khóa bí mật từ các khóa công khai tương ứng này và các giá trị riêng lẻ trong ma trận mặt nạ là an toàn và không thể suy ngược lại được từ giá trị công khai gửi đi.

Tương tự do r_i là ma trận bí mật, do đó việc tìm lại ma trận V_i tương đương với việc giải hệ phương trình tuyến tính $2 \times d \times d$ ẩn và $d \times d$ phương trình. Với d đủ lớn và không gian lựa chọn k đủ lớn thì gần như không có cách nào để thu lại được đúng ma trận r_i để đưa ra dự đoán về ma trận bí mật đầu vào V_i. Có nghĩa là từ giá trị công khai được chia sẻ, không có khả năng tìm lại được giá trị của ma trận bí mật V_i.

Tóm lại, trong tất cả các trường hợp, từ giá trị công khai được chia sẻ, kẻ tấn công bị động hoặc bất cứ một bên bán tin cậy nào cũng không thể khôi phục lại được

các ma trận khóa bí mật x_i, y_i, ma trận nhiễu mặt nạ r_i, ma trận đầu vào bí mật V_i của bất cứ một bên tham gia P_i nào.

Mệnh đề 3.2. Giao thức tính tổng vector thực bí mật trên bảo vệ cả dữ liệu bí mật của một bên tham gia bất kỳ ngay cả khi tồn tại (n - 2) bên thông đồng (và thông đồng với máy chủ $\mathcal{S}$).

Chứng minh: Không mất tính tổng quát, giả sử hai bên P_1 và P_2 là không thông đồng và các bên khác là thông đồng với nhau và thông đồng với máy chủ tổng hợp S.

Trong giao thức được đề xuất, các bên tham gia chỉ gửi các ma trận T_i, R_i, X_i, Y_i tới máy chủ tổng hợp trong đó các ma trận R_i, X_i, Y_i là các ma trận ngẫu nhiên.

Từ các ma trận R_i, không thể tính được R_1 hoặc R_2. Tương tự, từ các ma trận T_i cũng không thể tính được các ma trận T_1 và T_2 do chúng ta không đủ thông tin cần thiết. Các T_i, R_i, X_i, Y_i là độc lập với nhau

Lựa chọn tham số an toàn

Các giao thức đề xuất có các siêu tham số g, G, p là các tham số tương ứng với bài toán logarith rời rạc. Do đó, các tham số này được chọn theo các khuyến nghị với các hệ mật dựa trên bài toán logarith rời rạc. Ngoài ra các tham số bí mật x_i, y_i được chọn ngẫu nhiên sao cho $x_i \neq y_i$. Đối với r_i, chúng ta cũng chọn r_i ngẫu nhiên nằm trong một khoảng giá trị đủ lớn nhưng không quá lớn để thuận tiện cho quá trình tìm ma trận nhiễu mặt nạ tổng.

Giao thức học sâu có đảm bảo tính riêng tư dựa trên giao thức tính tổng bảo mật

Phần trên bài báo đã xem xét giao thức tính tổng bảo mật vector an toàn. Giao thức này sẽ được sử dụng để đảm bảo an toàn cho quá trình trao đổi và tổng hợp tham số trong quá trình huấn luyện. Giao thức huấn luyện mạng học sâu phân tán sử dụng giao thức tính tổng bảo mật an toàn thể hiện trong Thuật toán trong giai đoạn huấn luyện.

Algorithm	Thuật toán huấn luyện mạng học sâu phân tán có đảm bảo tính riêng tư dựa trên giao thức tính tổng vector bảo mật
Input	N bên P_i, mỗi bên sở hữu m_i mẫu dữ liệu và các ma trận khóa bí mật X_i, Y_i; Số lượng máy khách được chọn trong mỗi vòng huấn luyện K; Kích thước batch cục bộ B; Số vòng huấn luyện cục bộ E; Learning rate η; Bộ tham số khởi tạo chung W_0
Output	Bộ tham số của mô hình tổng hợp W

for *mỗi vòng t = 1, 2, ...* do

Máy chủ S chọn ra k thành viên S_t huấn luyện trong vòng t và gửi bộ tham số thu được từ vòng trước đó đến k thành viên này

$M \leftarrow \sum_{i \in S_t} m_i$

for *mỗi thành viên* $P_i \in S_t$ do

end

end

return W_t

SecureNodeUpdate *(i, W)*

end

end

return

Mã hóa theo thuật toán tính tổng bảo mật vector và gửi đến S

Bài báo tiến hành thực nghiệm phương pháp học sâu có đảm bảo tính riêng tư cho bài toán thông tin phát hiện thư rác dựa vào phương pháp tính tổng bảo mật đã phân tích ở trên.

Ứng dụng phương pháp học sâu có đảm bảo tính riêng tư cho bài toán an toàn thông tin phát hiện thư rác

Để đánh giá mô hình huấn luyện đề xuất, bài báo đã thực hiện thử nghiệm bằng cách sử dụng tập dữ liệu UCI SMS Spam. Phần này trình bày chi tiết về các đánh giá thử nghiệm về hiệu quả của mô hình với phân bố dữ liệu dạng IID, Non-IID.

Bảng 1. Kích cỡ bộ dữ liệu cho huấn luyện và kiểm thử

	SMS SPAM
Training size	4457
Testing size	1115

Bộ dữ liệu

Để đánh giá mô hình, nghiên cứu sử dụng 2 bộ dữ liệu phổ biến được dùng trong các đánh giá về các mô hình học sâu. Bộ dữ liệu đầu tiên là bộ dữ liệu ảnh MNIST [2] gồm các ảnh của chữ số viết tay có kích thước 32x32, được chuẩn hóa và đưa các chữ số về tâm ảnh và biến đổi kích thước về dạng 28x28 pixel.

Bộ dữ liệu thứ hai được sử dụng là bộ dữ liệu tin nhắn rác UCI SMS Spam Collection [3]. Đây là một bộ dữ liệu không cân bằng, chứa một tập các tin nhắn SMS được đánh dấu là Spam hoặc tin nhắn thường được thu thập để phục vụ cho mục đích nghiên cứu. Bộ dữ liệu này chứa 5.572 tin nhắn SMS tiếng Anh, trong đó có 746 mẫu được gắn nhãn là spam và 4.826 tin nhắn được gắn nhãn là bình thường.

Kiến trúc mạng

Trong phần thử nghiệm, bài báo sử dụng 2 kiến trúc mạng học sâu phổ biến là Mạng nơ ron tích chập cho bài toán phân loại ảnh MNIST và Mạng bộ nhớ dài-ngắn (Long-Short Term Memory - LSTM) cho bài toán phân loại tin nhắn rác trên bộ dữ liệu UCI SMS Spam. CNN là kiến trúc đặc biệt của mạng nơ ron đa lớp với kết nối thưa [4] và LSTM là kiến trúc đặc biệt của RNN [5].

Kích thước của dữ liệu đầu vào đối với bộ dữ liệu ảnh MNIST là 28 x 28 = 784 bằng với kích thước của lớp đầu vào của mạng CNN. Mục tiêu phân lớp là 2 lớp spam và tin nhắn thường, do đó kích thước lớp đầu ra là 2.

Mô hình kiến trúc của các mạng chi tiết được sử dụng được chỉ ra trong hình dưới đây. Số lượng tham số với mô hình CNN là 600810 và LSTM là 337761.

Kết quả thử nghiệm

Sau 100 vòng huấn luyện, mô hình huấn luyện đề xuất đạt độ chính xác lên tới 97,6% trong trường hợp phân phối đều và 93% trong trường hợp dữ liệu phân bố không đều. Mặc dù có những sự thay đổi trong các siêu tham số (B và E) tại các máy trạm thì kết quả huấn luyện đều cho độ chính xác rất cao.

Trong thực nghiệm này, bài báo sử dụng K = 100 tức là toàn bộ các bên sở hữu dữ liệu đều tham gia quá trình huấn luyện. Thử nghiệm diễn ra trên kích thước batch sử dụng cục bộ là 10 và 50 và số vòng huấn luyện cục bộ lần lượt là 1, 5 và 20. Các chỉ số đánh giá được thực hiện đo đạc tại các mốc 5, 10, 20, 50 và 100 vòng huấn luyện. Với kích thước batch là 50 và sử dụng 1 vòng huấn luyện cục bộ thì độ chính xác ban đầu đối với trường hợp dữ liệu phân bố đều và không đều lần lượt là 85% và 76% tại vòng huấn luyện toàn cục thứ 5. Sau 20 vòng huấn luyện, có sự biến đổi lớn về độ chính xác trong cả hai trường hợp phân phối dữ liệu.

Với trường hợp IID, báo cáo đánh giá độ chính xác của mô hình toàn cục tại một vòng nhất định và chọn đường cơ sở có độ chính xác là 97% vì độ chính xác này cao hơn nhiều so với độ chính xác nào của bất kỳ mô hình cục bộ độc lập nào có thể đạt được (tức là 95%). Kết quả cho thấy rằng mỗi bên thực hiện càng nhiều tính toán trên mỗi vòng, thì chi phí truyền thông để đạt đến được độ chính xác cơ sở càng giảm. Gần như trong mọi trường hợp độ chính xác đều có thể vượt qua độ chính xác cơ sở 97 % trong ít hơn 100 vòng huấn luyện.

Trong dữ liệu phân phối không đều, sau 100 vòng chỉ đạt đến độ chính xác cơ sở là 92 % với mọi trường hợp thay vì 97% như trường hợp phân phối đều. Nhưng trong trường hợp này, các siêu tham số cục bộ dường như không ảnh hưởng nhiều đến độ chính xác của mô hình khi số vòng truyền thông tăng lên. Vì vậy, đây là một bằng chứng mạnh mẽ minh chứng cho tính hiệu quả của mô hình đề xuất.

Kết luận

Vấn đề bảo đảm tính riêng tư cho quá trình học máy hiện nay đang trở thành 1 trong 3 chủ đề nóng nhất trong lĩnh vực trí tuệ nhân tạo. Kỷ nguyên số đang ngày càng đặt ra những yêu cầu khắt khe về tính riêng tư cho dữ liệu của người dùng. Mặc dù đã có rất nhiều giải pháp và hướng tiếp cận để giải quyết vấn đề về tính riêng tư cho quá trình học máy, tuy vậy những giải pháp này mới chỉ giải quyết được một phần nhỏ của vấn đề.

Các kết quả trong bài báo này là cơ sở để phát triển các nghiên cứu sâu hơn về những ứng dụng của đảm bảo tính riêng tư cho quá trình học sâu trong tương lai.