**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KỲ**

**MÔN: AN TOÀN HỆ THỐNG THÔNG TIN**

**Đề tài: CÁC THUẬT TOÁN TỔNG HỢP TRONG FEDERATED LEARNING**

Nhóm sinh viên thực hiện:

1. Lê Thị Ngọc Lan MSHV:

2. Vũ Mai Huỳnh Loan MSHV:

3. Lê Bá Trực MSHV: 240202016

GVHD: TS. Nguyễn Tấn Cầm

Tp. Hồ Chí Minh, 10/2024

**THÔNG TIN CHUNG**

|  |  |  |  |
| --- | --- | --- | --- |
| Họ và tên | Lê Thị Ngọc Lan | Vũ Mai Huỳnh Loan | Lê Bá Trực |
| MSHV | 240104010 |  | 240202016 |
| Email | lanltn.19@grad.uit.edu.vn |  | [Truclb.19@grad.uit.edu.vn](mailto:Truclb.19@grad.uit.edu.vn) |
| Ngành học | Hệ Thống Thông Tin |  | An toàn thông tin |
| Tự đánh giá điểm |  |  |  |
| Hình cá nhân |  |  |  |

Mục Lục

Danh sách các bản báo cáo

Danh sách các hình ảnh

**Chương I**

**Federated Learning (FL) là gì?**

Federated Learning (FL), hay còn gọi là học liên kết, là một phương pháp học máy phân tán, trong đó các mô hình được huấn luyện trên nhiều thiết bị hoặc máy chủ mà không cần tập trung dữ liệu tại một nơi. Thay vì chia sẻ dữ liệu thô, các thiết bị gửi các bản cập nhật mô hình (model updates) sau quá trình huấn luyện cục bộ, giúp bảo vệ tính riêng tư của dữ liệu.

FL được giới thiệu bởi Google vào năm 2016, và từ đó, nó đã trở thành một trong những giải pháp tiềm năng để giải quyết các vấn đề về quyền riêng tư và bảo mật trong học máy.

**Cơ chế hoạt động của Federated Learning**

Cơ chế hoạt động của FL có thể được chia thành 4 giai đoạn chính:

**1. Khởi tạo mô hình toàn cục (Global Model Initialization)**

* Mô hình toàn cục ban đầu được tạo trên một máy chủ trung tâm. Đây có thể là một mô hình đơn giản hoặc một mô hình phức tạp được thiết kế để huấn luyện phân tán.
* Mô hình này sau đó được gửi đến tất cả các thiết bị hoặc client tham gia.

**2. Huấn luyện cục bộ tại client**

* Mỗi client nhận mô hình toàn cục và sử dụng dữ liệu cục bộ của mình để huấn luyện mô hình đó. Quá trình huấn luyện này thường diễn ra theo một hoặc nhiều epoch.
* Vì dữ liệu không rời khỏi client, quyền riêng tư của dữ liệu được bảo vệ tốt hơn.

**3. Cập nhật mô hình cục bộ (Model Update)**

* Sau khi huấn luyện xong, các client sẽ gửi các bản cập nhật mô hình (trọng số hoặc gradient) về máy chủ trung tâm.
* Chỉ có thông tin mô hình (model updates) được gửi, không có dữ liệu thô nào được chia sẻ.

**4. Tổng hợp mô hình toàn cục (Global Model Aggregation)**

* Máy chủ trung tâm nhận các bản cập nhật từ client và thực hiện quá trình tổng hợp, thường sử dụng thuật toán **FedAvg** (Federated Averaging) hoặc các chiến lược khác.
* Sau đó, mô hình toàn cục được cập nhật và phân phối lại cho các client trong vòng tiếp theo.

Quá trình này được lặp lại qua nhiều vòng (federated rounds) cho đến khi mô hình đạt được hiệu suất mong muốn.

**Ưu điểm của Federated Learning**

**1. Bảo vệ quyền riêng tư**

* Dữ liệu không bao giờ rời khỏi client, giúp giảm nguy cơ rò rỉ dữ liệu.
* Thích hợp cho các lĩnh vực nhạy cảm như chăm sóc sức khỏe, tài chính, và dữ liệu cá nhân.

**2. Giảm chi phí truyền dữ liệu**

* Thay vì truyền tải toàn bộ dữ liệu đến một trung tâm, chỉ các bản cập nhật mô hình (trọng số hoặc gradient) được gửi đi, tiết kiệm băng thông mạng.

**3. Tận dụng dữ liệu phi tập trung**

* FL tận dụng dữ liệu cục bộ trên các thiết bị hoặc client, làm tăng tính đa dạng và độ bao phủ của dữ liệu.
* Có thể huấn luyện mô hình trên các thiết bị như điện thoại di động, IoT, hoặc hệ thống doanh nghiệp.

**4. Phù hợp với các quy định pháp lý**

* FL giúp tuân thủ các quy định nghiêm ngặt về bảo vệ dữ liệu, như GDPR ở châu Âu hoặc HIPAA trong ngành y tế.

**Nhược điểm của Federated Learning**

**1. Sự không đồng nhất trong dữ liệu (Non-IID)**

* Dữ liệu trên các client có thể không đồng nhất (non-IID), gây khó khăn cho việc huấn luyện và làm giảm hiệu suất mô hình toàn cục.

**2. Tài nguyên không đồng đều**

* Các client có thể khác nhau về tài nguyên tính toán, bộ nhớ và băng thông, dẫn đến sự không đồng nhất trong quá trình huấn luyện.

**3. Độ trễ cao**

* FL phụ thuộc vào kết nối mạng giữa máy chủ và client. Các client có băng thông yếu hoặc mất kết nối có thể làm chậm quá trình huấn luyện.

**4. Tấn công và bảo mật**

* Client hoặc máy chủ trung tâm có thể trở thành mục tiêu của các cuộc tấn công như **poisoning attack** (đầu độc dữ liệu), **model inversion attack** (tái tạo dữ liệu từ mô hình), và **free-rider attack** (client không đóng góp nhưng hưởng lợi từ mô hình).

**5. Chi phí tổng hợp mô hình**

* Việc tổng hợp mô hình (aggregation) đòi hỏi sức mạnh tính toán cao ở máy chủ trung tâm, đặc biệt với các mô hình phức tạp và số lượng lớn client.

**Ứng dụng của Federated Learning**

**1. Chăm sóc sức khỏe**

* Các bệnh viện sử dụng FL để huấn luyện mô hình dự đoán bệnh dựa trên dữ liệu của từng bệnh viện mà không cần chia sẻ dữ liệu bệnh nhân.

**2. Cá nhân hóa trên thiết bị**

* Google sử dụng FL để cải thiện tính năng gõ dự đoán trên bàn phím (Gboard) mà không cần tải dữ liệu người dùng lên máy chủ.

**3. Phát hiện gian lận**

* FL có thể được dùng để phát hiện gian lận trong các hệ thống tài chính mà không cần chia sẻ dữ liệu nhạy cảm giữa các tổ chức.

**4. IoT và Smart City**

* FL hỗ trợ các thiết bị IoT huấn luyện mô hình nhận diện hành vi, tối ưu hóa năng lượng, hoặc dự đoán giao thông.

**Chương II**

1. Triển khai Federated trong An toàn hệ thống thông tin  
   1. Vòng đời Mô hình trong Federated Learning cho Hệ thống Phát hiện Xâm nhập (IDS)

Federated Learning (FL) là một hướng tiếp cận tiên tiến trong việc huấn luyện mô hình học máy, đặc biệt là trong các hệ thống có dữ liệu phân tán và yêu cầu cao về quyền riêng tư. Trong lĩnh vực phát hiện xâm nhập (Intrusion Detection Systems - IDS), FL mang lại giải pháp hiệu quả, vừa đảm bảo sự an toàn dữ liệu, vừa nâng cao hiệu suất của mô hình phát hiện. Dưới đây là phân tích chi tiết về các giai đoạn trong vòng đời của một mô hình FL được áp dụng cho IDS, từ xác định vấn đề, thu thập dữ liệu, đến triển khai và bảo trì.

1. Từ xác định vấn đề đến thu thập dữ liệu

Quá trình phát triển một mô hình FL bắt đầu từ việc xác định rõ ràng mục tiêu cụ thể mà hệ thống IDS cần đạt được. Điều này bao gồm việc trả lời các câu hỏi: Hệ thống này cần phát hiện những dạng tấn công nào? Độ chính xác mục tiêu là bao nhiêu? Có yêu cầu gì về thời gian phản hồi hoặc tỷ lệ cảnh báo sai không?

Một khi mục tiêu đã rõ ràng, các thiết bị khách trong hệ thống, chẳng hạn như tường lửa, router, hoặc thiết bị đầu cuối, được thiết lập để thu thập dữ liệu cục bộ. Dữ liệu này bao gồm nhật ký lưu lượng mạng, sự kiện bảo mật, hoặc hành vi hệ thống. Do dữ liệu nhạy cảm không thể chuyển ra khỏi thiết bị, FL đóng vai trò quan trọng trong việc đảm bảo rằng dữ liệu này vẫn nằm trong giới hạn bảo mật trong khi được sử dụng để huấn luyện mô hình.

1. Huấn luyện mô hình phân tán

Giai đoạn huấn luyện là trọng tâm của FL. Tại đây, các thiết bị khách thực hiện huấn luyện cục bộ trên dữ liệu riêng của chúng. Điều này đảm bảo rằng dữ liệu thô không bị lộ ra ngoài mà vẫn đóng góp vào việc cải thiện mô hình toàn cục.

Quá trình huấn luyện diễn ra như sau:

Khởi tạo: Một mô hình khởi đầu được gửi từ máy chủ trung tâm đến các thiết bị khách.

Huấn luyện cục bộ: Mỗi thiết bị thực hiện tối ưu hóa mô hình dựa trên dữ liệu của mình. Điều này có thể bao gồm việc tính toán gradient hoặc cập nhật trọng số.

Tổng hợp mô hình: Các thiết bị gửi bản cập nhật mô hình (model updates) – thường là gradient hoặc trọng số – về máy chủ trung tâm. Máy chủ này sẽ tổng hợp các bản cập nhật để tạo ra một mô hình toàn cục mới. Quá trình tổng hợp phổ biến nhất là Federated Averaging (FedAvg), trong đó các bản cập nhật được kết hợp dựa trên trọng số của số lượng dữ liệu từng thiết bị.

Lợi thế lớn nhất của phương pháp này là nó tận dụng được sự đa dạng của dữ liệu từ các nguồn khác nhau, giúp mô hình học hỏi từ nhiều dạng lưu lượng mạng và hành vi tấn công khác nhau.

1. Đánh giá và lựa chọn mô hình

Sau khi huấn luyện, bước tiếp theo là đánh giá hiệu suất mô hình. Có hai phương pháp chính để đánh giá:

Kiểm tra với dữ liệu chuẩn: Mô hình được đánh giá trên một tập dữ liệu chuẩn trong trung tâm dữ liệu, thường được sử dụng để so sánh các mô hình khác nhau.

Kiểm tra trên các thiết bị thực tế: Mô hình được triển khai thử nghiệm trên các thiết bị cục bộ chưa tham gia vào quá trình huấn luyện để đánh giá khả năng phát hiện các hành vi bất thường hoặc xâm nhập thực tế.

Các chỉ số đánh giá quan trọng bao gồm:

Độ chính xác (Accuracy): Đo lường khả năng phát hiện đúng các hành vi xâm nhập.

Tỷ lệ cảnh báo sai (False Positive Rate - FPR): Xác định tần suất mô hình báo động sai.

Hiệu quả tính toán: Kiểm tra xem mô hình có đáp ứng được yêu cầu thời gian thực trong môi trường sản xuất không.

Khi quá trình đánh giá hoàn tất, mô hình tốt nhất sẽ được chọn dựa trên hiệu suất tổng thể và khả năng đáp ứng các yêu cầu cụ thể của hệ thống IDS.

1. Triển khai mô hình

Giai đoạn triển khai bao gồm tích hợp mô hình vào môi trường mạng thực tế. Quá trình này không chỉ đơn thuần là đưa mô hình vào hoạt động, mà còn yêu cầu:

Kiểm tra chất lượng: Đảm bảo rằng mô hình hoạt động ổn định và không gây gián đoạn cho hệ thống mạng.

Thử nghiệm A/B: So sánh hiệu suất của mô hình mới với các mô hình hoặc phương pháp hiện có.

Triển khai dần (Staged Rollout): Đưa mô hình vào hoạt động theo từng giai đoạn, từ nhỏ đến lớn, để giảm thiểu rủi ro.

Khi được triển khai, mô hình sẽ liên tục giám sát lưu lượng mạng, phân tích nhật ký và phát hiện các mối đe dọa trong thời gian thực. Hệ thống cũng cần có cơ chế phản hồi, cho phép tự động hóa một phần hoặc toàn bộ các hành động ứng phó với xâm nhập.

1. Bảo trì và cập nhật mô hình

Môi trường an ninh mạng luôn thay đổi với các mối đe dọa mới xuất hiện liên tục. Do đó, một mô hình FL cho IDS cần được bảo trì thường xuyên để đảm bảo hiệu suất tối ưu.

Các hoạt động bảo trì bao gồm:

Giám sát hiệu suất: Theo dõi độ chính xác và khả năng phát hiện của mô hình qua thời gian.

Phát hiện vấn đề: Xác định các kịch bản mà mô hình hoạt động kém, chẳng hạn như tỷ lệ cảnh báo sai tăng cao hoặc bỏ sót các dạng tấn công mới.

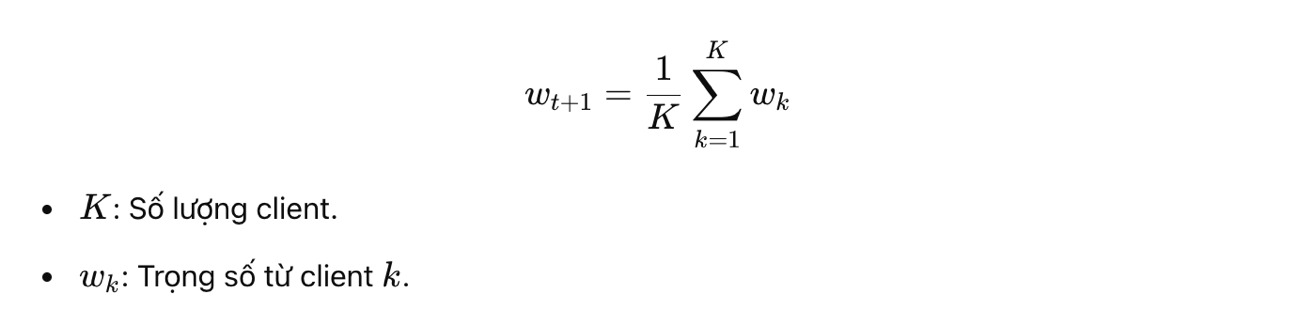
Huấn luyện lại định kỳ: Mô hình được huấn luyện lại với dữ liệu mới từ các thiết bị khách. Điều này đảm bảo rằng mô hình học hỏi được từ các xu hướng tấn công mới và duy trì hiệu quả trong phát hiện xâm nhập.

Quá trình bảo trì không chỉ cải thiện hiệu suất của mô hình mà còn tăng cường độ tin cậy của hệ thống IDS trong việc bảo vệ mạng.

**Chương 3:** So sánh các thuật toán tổng hợp: Federated Averaging, Weghted Federated Averaging, Median-based Federated Averaging, Trimmed Mean, Krum (K-Center)

**Federated Averaging (FedAvg)**

* Mô tả:
  + Trung bình các mô hình từ các client mà không sử dụng trọng số.
  + Mỗi client có ảnh hưởng ngang nhau trong việc tổng hợp mô hình toàn cục.
* Công thức:



* Ưu điểm:
  + Đơn giản, dễ triển khai.
  + Tính toán nhanh, không yêu cầu thêm thông tin như kích thước dữ liệu.
* Nhược điểm:
  + Không hiệu quả nếu dữ liệu không cân bằng giữa các client (client có ít dữ liệu cũng đóng góp ngang bằng client có nhiều dữ liệu).

**Weighted Federated Averaging**

* Mô tả:
  + Mở rộng từ FedAvg, sử dụng trọng số dựa trên kích thước dữ liệu của từng client.
  + Các client có nhiều dữ liệu hơn sẽ đóng góp nhiều hơn vào mô hình toàn cục.
* Công thức:

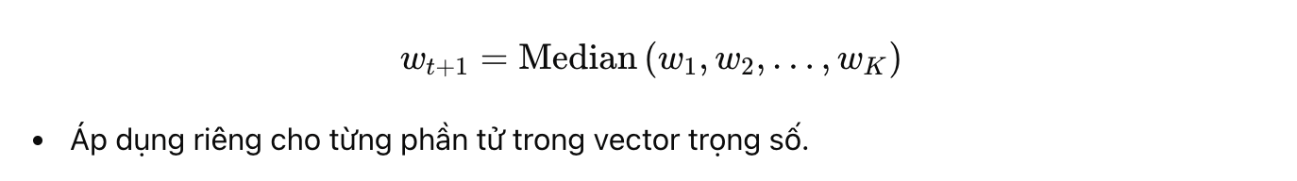
A black and white math equations

Description automatically generated with medium confidence

* Ưu điểm:
  + Giải quyết vấn đề dữ liệu không cân bằng giữa các client.
  + Cải thiện hiệu năng tổng hợp mô hình toàn cục.
* Nhược điểm:
  + Yêu cầu thông tin về kích thước dữ liệu của từng client, có thể làm giảm tính bảo mật.

**Median-based Federated Averaging**

* Mô tả:
  + Thay vì trung bình, sử dụng trung vị (median) để tổng hợp mô hình.
  + Loại bỏ ảnh hưởng của các giá trị ngoại lai (outliers).
* Công thức:



* Ưu điểm:
  + Chống lại các mô hình bất thường hoặc độc hại từ các client.
  + Đơn giản và dễ triển khai.
* Nhược điểm:
  + Hiệu năng kém trong trường hợp tất cả các client đều hợp tác và không có ngoại lệ.

**Trimmed Mean**

* Mô tả:
  + Loại bỏ một số lượng nhỏ các giá trị lớn nhất và nhỏ nhất từ các trọng số trước khi tính trung bình.
  + Bảo vệ mô hình toàn cục khỏi các outliers.
* Công thức:

A white background with black text

Description automatically generated

* Ưu điểm:
  + Tốt hơn trung vị (median) trong việc giảm thiểu ảnh hưởng của outliers, đặc biệt khi dữ liệu gần Gaussian.
* Nhược điểm:
  + Yêu cầu tham số betaβ, đòi hỏi điều chỉnh tùy thuộc vào dữ liệu.

**Krum (K-Center)**

* Mô tả:
  + Chọn một trọng số từ các client làm trọng số toàn cục sao cho nó gần với đa số các trọng số khác nhất.
  + Được thiết kế để chống lại các tấn công Byzantine (các client gửi thông tin sai lệch).
* Công thức:

A white background with black text

Description automatically generated

* Ưu điểm:
  + Hiệu quả trong việc bảo vệ mô hình khỏi các client độc hại.
  + Được thiết kế cho môi trường không tin cậy.
* Nhược điểm:
  + Tốn kém về tính toán do cần đo khoảng cách giữa các trọng số.
  + Không hiệu quả nếu tất cả client hợp tác.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Thuật toán | Độ đơn giản | Chống ngoại lai | Chống client độc hại | Hiệu quả với dữ liệu không cân bằng | Tính toán phức tạp |
| Federated Averaging | Cao | Thấp | Thấp | Thấp | Thấp |
| Weighted Federated Averaging | Trung Bình | Thấp | Thấp | Cao | Trung Bình |
| Median-based FedAvg | Trung Bình | Cao | Trung Bình | Trung Bình | Trung Bình |
| Trimmed mean | Trung Bình | Cao | Trung Bình | Trung Bình | Trung Bình |
| Krum | Thấp | Cao | Cao | Thấp | Cao |

*Table 01: Bảng so sánh các thuật toán tổng hợp trọng số*

Chương III Thiết kế thực nghiệm và Demo

Chương IV Kết quả chạy demo

Chương V Kết Luận