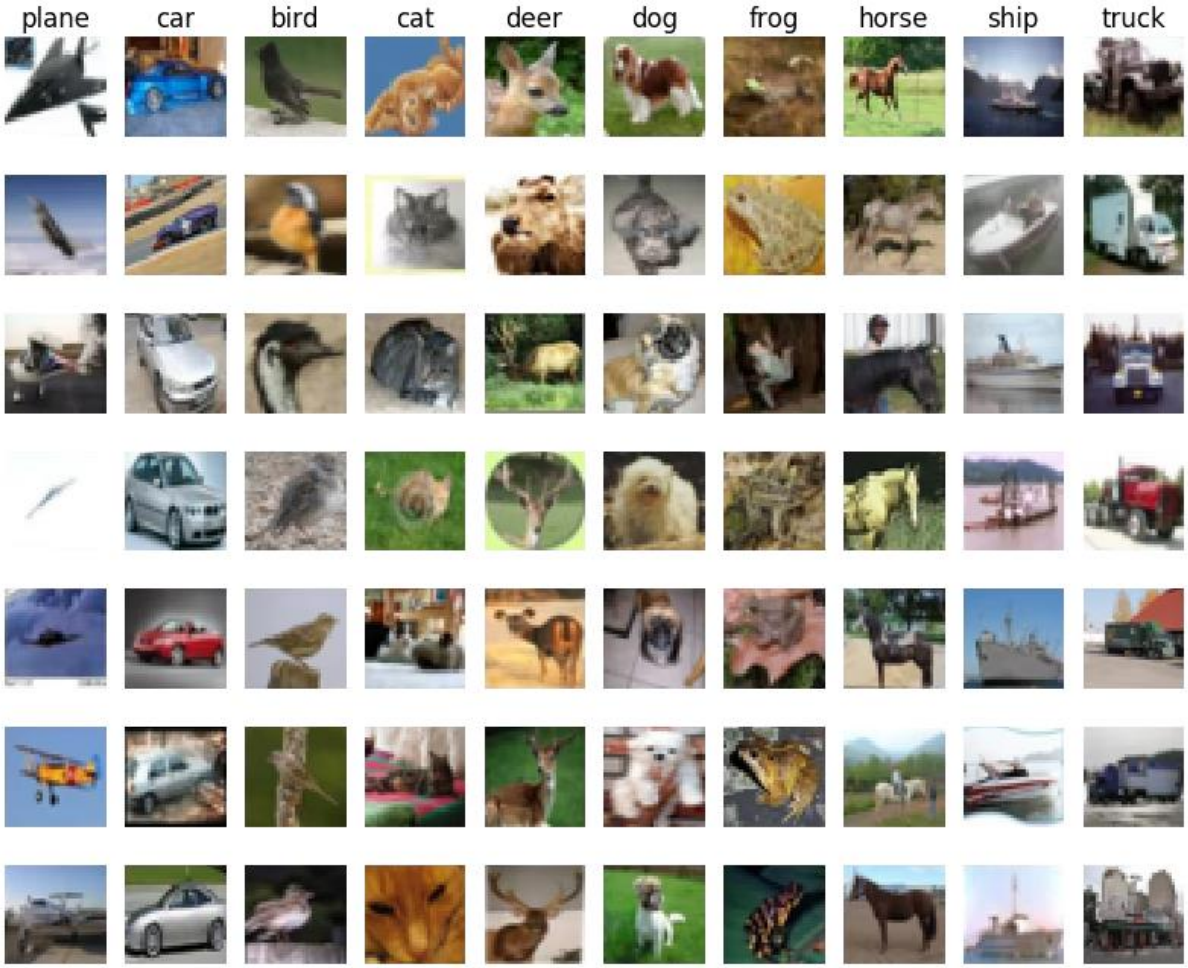
LAB 2 – HỌC LIÊN KẾT (FEDERATE LEARNING) VÀ BẢO MẬT TẠI SERVER

1. Tập dữ liệu CIFAR10

1.1. Giới thiệu

- CIFAR10 là tập các dữ liệu hình ảnh được sử dụng rộng rãi cho các thuật toán máy học và thị giác máy tính. CIFAR10 bao gồm 60.000 hình ảnh màu 32x32 pixel trong 10 lớp khác nhau.
- 10 lớp khác nhau đại diện cho máy bay, ô tô, chim, mèo, hươu, chó, ếch, ngựa, tàu và xe tải.
- Mỗi lớp có 6.000 hình, gồm 5.000 hình ảnh cho huấn luyện và 1.000 hình ảnh cho thử nghiệm.



1.2. Tải tập dữ liệu

Chúng ta có thể tải tập dữ liệu CIFAR10 bằng nhiều cách:

- Tải trực tiếp từ trang CIFAR10: https://www.cs.toronto.edu/~kriz/cifar-10-python.tar.gz
- Cài đặt thư viện torch and torchvision.

1.3. Yêu cầu xử lý dữ liệu (1đ)

- Mô hình logistic không thể hoạt động tốt trên tập dữ liệu phức tạp như CIFAR10. Vì hồi quy logistic là bộ phân loại tuyến tính và hoạt động tốt nhất với dữ liệu có thể tách biệt tuyến tính, nhưng dữ liệu CIFAR10 lại có tính phi tuyến cao.
- Yêu cầu: xử lý dữ liệu CIFAR10 để có thể chạy được với mô hình hồi quy logistic.

2. Học liên kết

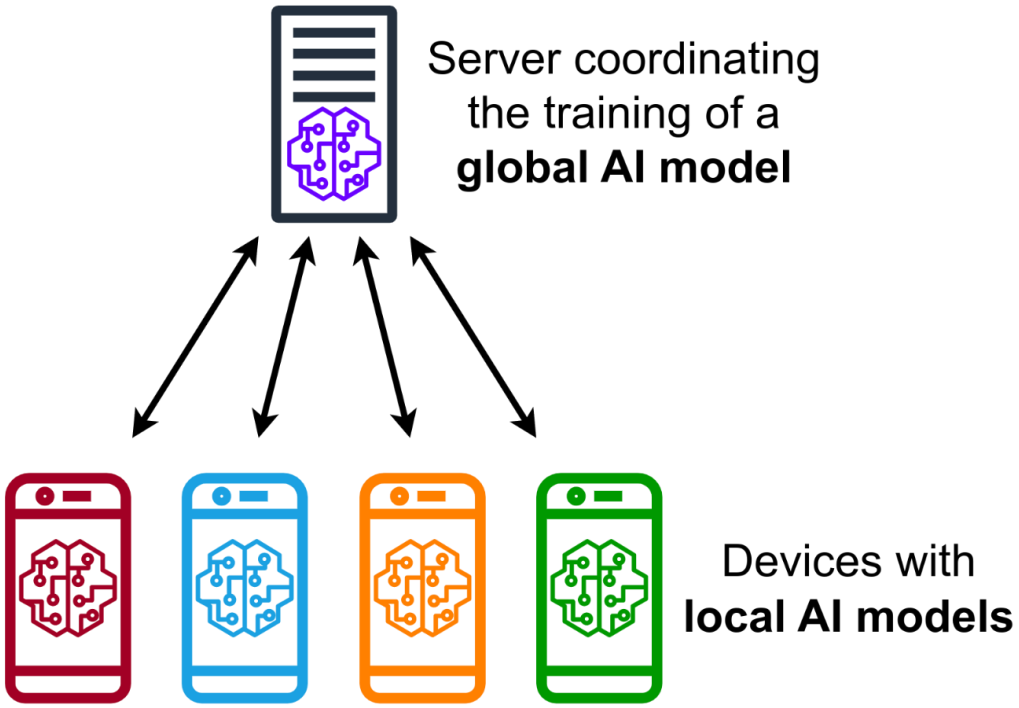
2.1. Hồi quy Logistic

Hồi quy logistic là một kỹ thuật thống kê và học máy được sử dụng để mô hình hóa mối quan hệ giữa một biến độc lập (hoặc nhiều biến độc lập) và một biến phụ thuộc nhị phân (binary). Nó được sử dụng rộng rãi trong các bài toán phân loại, đặc biệt là những bài toán mà đầu ra chỉ có hai giá trị, chẳng hạn như "có" hoặc "không", "1" hoặc "0".

2.2. Học liên kết

Federated learning: là một kỹ thuật học máy phân tán, nơi mà các mô hình được huấn luyện trên nhiều thiết bị hoặc máy chủ (clients) cục bộ mà không cần phải gửi dữ liệu lên một máy chủ trung tâm. Thay vì chia sẻ dữ liệu thô giữa các thiết bị, mỗi thiết bị sẽ huấn luyện mô hình cục bộ dựa trên dữ liệu của mình, sau đó chỉ chia sẻ các thông số mô hình đã cập nhật (như trọng số) với một máy chủ trung tâm.

Máy chủ trung tâm sẽ nhận các bản cập nhật mô hình từ nhiều thiết bị khác nhau, kết hợp chúng để tạo ra một mô hình toàn cục mà không cần truy cập trực tiếp vào dữ liệu cục bộ của các thiết bị. Điều này giúp đảm bảo tính riêng tư của dữ liệu, đồng thời giảm thiểu việc truyền tải dữ liệu lớn qua mạng.



2.3 Yêu cầu (6đ)

Cài đặt mô hình hồi quy logistic với tập dữ liệu CIFAR10.

Cài đặt mô hình 1 server, 3 client. Client setup huấn luyện với tập dữ liệu CIFAR10.

- Client 1 dữ liệu gồm có các lớp: máy bay, ô tô, chim, mèo, hươu.
- Client 2 dữ liệu gồm có các lớp: ô tô, chim, mèo, hươu, chó, ếch, ngựa, tàu và xe tải.
- Client 3: dữ liệu đầy đủ các lớp.

Server tổng hợp dữ liệu từ 3 client sử dụng trung bình cộng các trọng số của các model được gửi lên từ client.

Link github tham khảo: https://github.com/coMindOrg/federated-averaging-tutorials

3. Attacks (Tấn công)

3.1. Black-box attacks

Black-box attacks là một loại tấn công trong lĩnh vực an ninh mạng và học máy, trong đó kẻ tấn công không có quyền truy cập trực tiếp vào mô hình hoặc dữ liệu huấn luyện của hệ thống mà họ muốn tấn công. Thay vào đó, họ chỉ có thể tương tác với mô hình thông qua đầu vào và đầu ra.

Black-box attacks xảy ra khi kẻ tấn công cố gắng khai thác một mô hình học máy mà không có thông tin nội bộ về cấu trúc, tham số hay dữ liệu huấn luyện của mô hình đó. Kẻ tấn công chỉ có thể gửi yêu cầu (input) đến mô hình và nhận phản hồi (output).

3.2. Các Kỹ Thuật Tấn Công Black-Box Phổ Biến:

- **Adversarial Examples:** Đây là những đầu vào được chế tạo một cách cẩn thận nhằm mục đích bị phân loại sai bởi mô hình, mặc dù chúng có vẻ tương tự như các đầu vào hợp lệ đối với con người. Ví dụ đối kháng có thể được tạo ra bằng nhiều kỹ thuật khác nhau, chẳng hạn như tối ưu hóa theo độ dốc (gradient ascent), thuật toán di truyền, và mạng đối kháng sâu (deep adversarial networks).
- **Membership Inference Attacks:** Những cuộc tấn công này cố gắng xác định xem một điểm dữ liệu cụ thể có được sử dụng để huấn luyện một mô hình hay không. Điều này có thể đạt được bằng cách phân tích các dự đoán của mô hình trên điểm dữ liệu mục tiêu và so sánh chúng với các dự đoán trên các điểm dữ liệu khác.
- **Model Stealing Attacks:** Những cuộc tấn công này nhằm mục đích sao chép một mô hình mục tiêu bằng cách truy vấn nó với các đầu vào được chọn cẩn thận và phân tích các đầu ra. Điều này có thể được thực hiện bằng cách sử dụng các kỹ thuật như suy diễn tham gia hoặc bằng cách huấn luyện một mô hình mới trên cùng một tập dữ liệu với mô hình mục tiêu.

3.3. Yêu cầu (3đ)

- Viết báo cáo về các mô hình tấn công theo thứ tự:

Nhóm 1 - 7: Adversarial Examples

Nhóm 8 – 14: Membership Inference Attacks

Các nhóm còn lại: Model Stealing Attacks

- Lập trình mô hình tấn công (optional).

4. Các yêu cầu khác

- Ngôn ngữ sử dụng bắt buộc là Python, không được phép sử dụng ngôn ngữ khác.
- Không giới hạn thư viện được sử dụng trong Python.
- Các nhóm cần kiểm tra mã nguồn trước khi nọp. Nếu mã nguồn không chayuj được mà không phải do nguyên nhân khách quan (thiếu thư viện, lỗi do thư viện gây ra, sử dụng thư viện sai phiên bản, …) thì sẽ bị 0 điểm đồ án.

- Bài nộp gồm có 3 phần:
    - Report: chứ các file báo cáo.
    - Source: Chứa các file mã nguồn.
    - Presentation: Chứa các file dùng để thuyết trình.
- Trong file nộp, nhóm cần ghi rõ thông tin các thành viên gồm họ tên và MSSV. Riêng đối với mã nguồn, nhóm có thể ghi thông tin trên dưới dạng comment trong code của nhóm.
- Bài nộp sẽ được đặt trong thư mục có tên MSSV01[_MSSV02_[MSSV03[…]]] và được nén lại bằng định dạng ZIP với cùng tên như trên. Ví dụ đặt tên nhóm 1 có 1 sinh viên là MSSV01 , nhóm có 2 sinh viên là MSSV01_MSSV02.
- Nghiêm cấm hành vi gian lận, không trung thực trong học tập như sao chép bài làm giữa các nhóm với nhau, sao chép bài làm của các nhóm khóa trước hoặc các nhóm lớp khác trường khác, nhờ người làm hộ. Nếu phát hiện các hành vi thì cả nhóm sẽ bị 0 điểm và xử lý theo quy định của Khoa và Trường.