**Federated Learning** (học liên kết) là một phương pháp học máy phân tán, cho phép những người nắm giữ dữ liệu hợp tác với nhau để đào tạo mô hình mà không cần chia sẻ dữ liệu 1 cách hợp lý , thay vào đó những người tham gia huấn luyện 1 bản sao của mô hình với dữ liệu cục bộ và chia sẻ lại các tham số cập nhật tham số kết quả qua 1 máy chủ tập trung. Qua các vòng huấn luyện mô hình trung tâm sẽ được cải thiện bằng cách kết hợp thông tin từ nhiều nguồn mà không cần tiết lộ dữ liệu người dùng.

**Adversarial attack** (tấn công chống đối) là một kỹ thuật được sử dụng để thay đổi đầu vào của mô hình trí tuệ nhân tạo (AI) nhằm gây ra lỗi hoặc đưa ra dự đoán sai lầm. Mục tiêu của tấn công chống đối là tìm ra những điểm yếu trong mô hình AI và khai thác chúng để tạo ra các đầu vào gây ra sự nhầm lẫn hoặc lừa đảo mô hình.

Tấn công chống đối thường được thực hiện bằng cách thêm nhiễu hoặc biến đổi đầu vào một cách tinh vi. Một dạng phổ biến của tấn công chống đối là tấn công chống đối trong hình ảnh, trong đó một hình ảnh gốc được thay đổi một cách nhỏ để đánh lừa mô hình nhận dạng hình ảnh.

Ví dụ, giả sử chúng ta có một mô hình nhận dạng hình ảnh chó và mèo. Mô hình này đã được huấn luyện trên một tập dữ liệu lớn chứa hình ảnh của chó và mèo để học cách phân biệt giữa hai loại động vật này. Tuy nhiên, nhà tấn công nhận thấy rằng chỉ cần thêm một nhiễu nhỏ vào hình ảnh chó, mô hình sẽ nhận diện nó là một con mèo.

Nhà tấn công tạo ra một đầu vào mới bằng cách thêm một lượng nhỏ nhiễu vào hình ảnh chó gốc. Những thay đổi này có thể không đáng kể với mắt người, nhưng lại đủ để gây nhầm lẫn cho mô hình nhận dạng. Khi hình ảnh chó bị biến đổi được đưa vào mô hình, nó sẽ không nhận ra đó là hình ảnh chó mà sẽ sai lầm là một con mèo.

Điều đáng ngạc nhiên là các tấn công chống đối có thể là không rõ ràng với con người, nhưng lại có thể gây ra sự nhầm lẫn lớn cho mô hình AI. Những thay đổi nhỏ như thêm nhiễu, biến đổi đặc trưng hay chèn dữ liệu độc hại đều có thể gây ra tác động lớn đến độ chính xác của mô hình.

Adversarial attack trong federated learning:

Tấn công vào quá trình tập trung dữ liệu (Data Poisoning Attack): Khi dữ liệu được gửi từ các thiết bị địa phương đến trung tâm, kẻ tấn công có thể chèn các dữ liệu giả mạo hoặc ô nhiễm vào dữ liệu huấn luyện. Điều này có thể làm sai lệch quá trình huấn luyện và làm hỏng mô hình toàn cục (global model) khi được cập nhật.

Tấn công vào quá trình học tập (Model Poisoning Attack): Kẻ tấn công có thể cố ý gửi các mô hình độc hại từ các thiết bị địa phương đến trung tâm để ảnh hưởng đến mô hình toàn cục. Các mô hình độc hại này có thể được thiết kế để gian lận hoặc giả mạo dữ liệu trong quá trình học tập.

Tấn công vào quá trình truyền dữ liệu (Communication Attack): Kẻ tấn công có thể ngăn chặn hoặc xâm nhập vào quá trình truyền dữ liệu giữa các thiết bị địa phương và trung tâm. Điều này có thể làm gián đoạn hoặc làm thay đổi dữ liệu gửi đi và nhận về, ảnh hưởng đến quá trình học tập và làm hỏng mô hình toàn cục.