**Federated Learning** là một phương pháp học máy phân tán, nơi quá trình huấn luyện mô hình diễn ra trên các thiết bị đầu cuối phân tán, mà không cần truyền dữ liệu về một máy chủ trung tâm. Thay vì gửi dữ liệu từ các thiết bị đến máy chủ trung tâm để huấn luyện, mô hình được gửi đến các thiết bị và huấn luyện trên chúng. Sau đó, thông tin được tóm tắt từ các mô hình địa phương được tổng hợp và cập nhật mô hình trung tâm.

Quá trình huấn luyện Federated Learning có thể được mô tả như sau:

* Khởi tạo mô hình trung tâm: Một mô hình trung tâm ban đầu được khởi tạo trên máy chủ trung tâm.
* Phân phối mô hình đến các thiết bị đầu cuối: Mô hình được gửi từ máy chủ trung tâm đến các thiết bị cá nhân để huấn luyện.
* Huấn luyện mô hình trên các thiết bị cá nhân: Các thiết bị đầu cuối sử dụng dữ liệu địa phương của chúng để huấn luyện mô hình. Việc huấn luyện này có thể diễn ra nhiều vòng lặp để cải thiện mô hình.
* Tổng hợp thông tin mô hình: Thông tin tóm tắt về các thay đổi của mô hình trên các thiết bị địa phương được gửi về máy chủ trung tâm.
* Cập nhật mô hình trung tâm: Máy chủ trung tâm sử dụng thông tin tóm tắt từ các thiết bị địa phương để cập nhật mô hình chung.

Qua các vòng lặp tiếp theo của quá trình này, mô hình trung tâm sẽ được cải thiện bằng cách kết hợp thông tin từ tất cả các thiết bị đầu cuối 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.

Tấn công chống đối có thể gây nguy hiểm trong nhiều ngữ cảnh, bao gồm hệ thống an ninh, y tế và xe tự hành. Vì vậy, việc nghiên cứu và phát triển các biện pháp phòng ngừa và phát hiện tấn công chống đối là rất quan trọng để bảo vệ tính toàn vẹn và đáng tin cậy của các mô hình trí tuệ nhân tạo.