ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Nghiên cứu và đề xuất mô hình học liên kết hướng người dùng

ĐOÀN NGỌC KHÁNH
khanh.dn180110@sis.hust.edu.vn

Ngành: Công nghệ thông tin

Giảng viên hướng dẫn: TS. Nguyễn Phi Lê _____

Chữ kí GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ thông tin và Truyền thông

HÀ NỘI, 03/2023

LỜI CẢM ƠN

Hơn bốn năm học đại học trôi qua thật nhanh mà cũng thật chậm. Chậm cho những giây phút đợi chờ, nhanh cho những khoảng khắc đáng nhớ mà ta đã không kịp trân trọng đúng mực. Bách Khoa đã cho tôi thật nhiều, cũng lấy đi thật nhiều, nhưng giúp tôi trưởng thành và khôn lớn. Bách Khoa là nền tảng, là bước ngoặt trong cuộc đời tôi.

Trước hết, con xin cảm ơn cha mẹ, chị gái, gia đình, và đặc biệt cảm ơn TS. Nguyễn Phi Lê. Mọi người đã luôn quan tâm, chăm sóc cho con, luôn giúp đỡ con những lúc khó khăn, lúc con đáng trách nhất.

Em xin cảm ơn cô Nguyễn Phi Lê và mọi anh chị em trong tập thể AIoT Lab. Cô là người tận tâm, chuyên cần, tận tình và thấu đáo. Cô đã chỉ dạy cho em rất nhiều, tạo cho em nhiều cơ hội học tập, tiếp xúc tốt đẹp. Được làm việc trong lab là được trao đổi với mọi người, giúp em xây dựng được hướng đi, cách nghĩ, cách làm việc.

Cảm ơn Bá Tân, Đức Long, Quang Điện, Hồng Sang, Việt Hoàng, Tiến Long, Phi Phúc, Trần Anh đã sát cánh trong suốt 4 năm đại học, cùng nhau vượt qua những khó khăn, những ngày vui và bao chuyện đáng nhớ. Cảm ơn những người bạn cấp ba vẫn luôn gắn bó với tôi.

Cuối cùng, xin được cảm ơn Bách Khoa.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Cách mạng công nghiệp lần thứ tư đem lại nhiều thay đổi lớn đối với nhân loại, đó là sự phổ biến của Internet kết nối vạn vật (IoT), dữ liệu lớn (BigData) và trí tuệ nhân tạo (AI). Các thiết bị di động với khả năng tính toán tương đối tốt ngày một phổ biến. Dữ liệu gia tăng với tốc độ hàm mũ, được lưu trữ, số hóa và trở thành tài nguyên phát triển các mô hình học máy, học sâu của trí tuệ nhân tạo. Việc huấn luyện các mô hình học sâu thông thường cần yêu cầu tập hợp dữ liệu lại trên cùng một thiết bị. Tuy nhiên, bên cạnh các nguồn dữ liệu mở, dữ liệu mang tính cá nhân, bảo mật chiếm đa số. **Học liên kết (federated learning)** ra đời như một giải pháp xây dựng mô hình học sâu từ nhiều thiết bị riêng biệt với bộ dữ liệu độc lập để đảm bảo hiệu quả tốt trên tập dữ liệu kiểm thử. **Học liên kết hướng người dùng (personalized federated learning)** là một ngữ cảnh khác khi mà mỗi thiết bị tham gia có một tập dữ liệu kiểm thử riêng bên cạnh việc có một tập dữ liệu huấn luyện riêng. Mục tiêu khi này là xây dựng cho mỗi thiết bị một mô hình riêng đạt hiệu quả cao trên bộ dữ liệu kiểm thử của chính nó. Trong những nghiên cứu gần đây về bài toán này, có một nghiên cứu mới nổi là **Personalized Federated Learning through Local Memorization** [1] đem lại hiệu quả tốt dựa trên kỹ thuật tương đối đơn giản đó là sử dụng kết hợp mô hình huấn luyện chung giữa các thiết bị với thuật toán K **láng giềng gần nhất (**K**NN)** khi thực hiện dự đoán. Trong đồ án này, em đề xuất phương pháp huấn luyện sử dụng **học đối lập (contrastive learning)** khi huấn luyện mô hình nhằm cải thiện hiệu quả của phương pháp K**NN**, qua đó nâng cao hiệu quả của phương pháp trên.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	**Ý nghĩa**
KNN	Thuật toán K láng giềng gần nhất (KNN)
KNN-Per	Nghiên cứu **Personalized Federated Learning through Local Memorization**
FedAvg	Thuật toán FedAvg trong học liên kết
FedProx	Thuật toán FedProx trong học liên kết
SAM	Phương pháp tối ưu tránh điểm cực tiểu có bề mặt nhọn (Sharpness Aware Minimization)

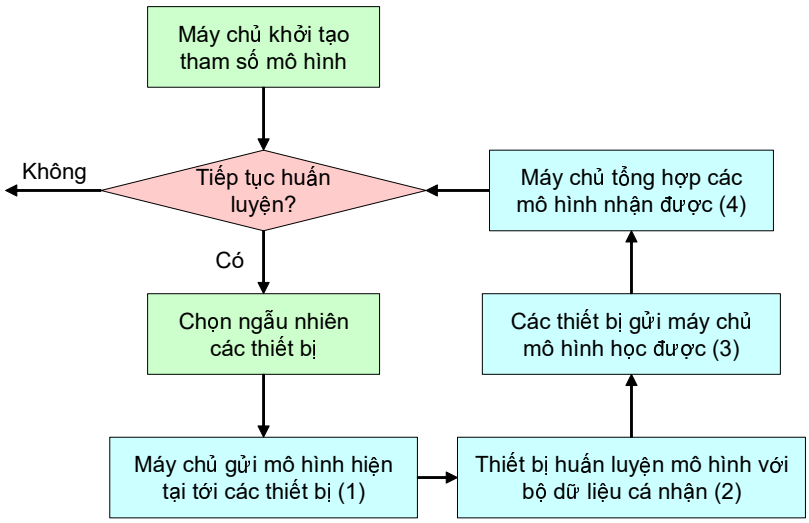
CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

1.1.1 Học liên kết (Federated Learning)

Học liên kết được Google giới thiệu vào năm 2017. Với hướng tiếp cận xây dựng các mô hình học máy trước đó, chúng ta cần tập trung dữ liệu lại trên một thiết bị hay trung tâm dữ liệu, và bài toán hướng đến là tập trung nâng cao hiệu quả mô hình. Học liên kết tồn tại một máy chủ (server), cho phép nhiều thiết bị kết nối với máy chủ, như điện thoại di động với khả năng tính toán tương đối tốt, tham gia huấn luyện một mô hình học sâu chung bằng những bộ dữ liệu trên từng thiết bị đó.

Thuật toán đầu tiên được đề xuất giải quyết bài toán học liên kết đó là **FedAvg** [2]. Đây cũng là hình mẫu chung cho nhiều thuật toán về sau phát triển. Như học sâu thông thường, ban đầu máy chủ sẽ khởi tạo một bộ tham số mô hình. Quá trình học mô hình sẽ tồn tại nhiều vòng, gọi là vòng kết nối (communication round). Tại mỗi vòng, máy chủ sẽ chọn một hay một vài thiết bị ngẫu nhiên, gửi bộ tham số mô hình hiện tại về cho các thiết bị đó. Các thiết bị này sẽ tiến hành huấn luyện mô hình dữ trên bộ dữ liệu của chúng với một số ít vòng lặp, sau đó gửi lại máy chủ mô hình thu được. Máy chủ sẽ tổng hợp lại các mô hình này theo phương pháp trung bình cộng có trọng số, trong đó trọng số tỷ lệ thuận với lượng dữ liệu huấn luyện của các thiết bị tham gia. Đây sẽ là bộ tham số mới của mô hình tại vòng này. Vòng mới sẽ được bắt đầu tương tự cho đến khi mô hình hội tụ.



Hình 1.1: Sơ đồ hoạt động thuật toán FedAvg

Tồn tại bốn thách thức lớn đối với học liên kết [3]: