An Infrastructure for Collaborative Machine Learning on Resource-Constrained Heterogeneous Environments

Name: Kundjanasith Thonglek

Laboratory' s name: Software Design and Analysis Laboratory

Supervisor' s name: Professor Hajimu Iida

Abstract

Collaboration has been vital for the rapid and successful growth of the software industry. Software development infrastructures, such as GitHub for source codes and DockHub for container images, allow individuals from diverse backgrounds and organizations to work together and create complex and large-scale software that even big tech companies find it challenging to develop and maintain. However, such collaborative infrastructure is not yet available for machine learning models. This presents an opportunity to introduce LiberatAI into computer science for removing the barrier to the collaborative development of machine learning models from the limitation of data privacy and existing resource constraints.

In this dissertation, I propose LiberatAI, an infrastructure for collaboratively developing machine learning models that allow researchers to work together and potentially build better models than big companies can. LiberatAI applies federated learning to train the models while preserving data privacy. LiberatAI allows individuals to collaboratively train models on their environments, which are usually heterogeneous. Three modules in LiberatAI support training a model on diverse storage, computing, and communication resources. (1) Compressor module is proposed to reduce the model size to fit the storage capacity of the environment. (2) Aggregator module is proposed to aggregate the models trained on heterogeneous computing resources. (3) Sparsifier module is proposed to sparsify the model for exchanging the model between a server and clients. LiberatAI was evaluated using state-of-the-art neural network models to detect COVID-19 cases from chest X-ray images. COVID-19 detection is one of the most popular machine learning applications for privacy-sensitive data. As a result, the ensemble model with heterogeneous structures on six different hardware environments from LiberatAI produces accuracy higher than a trained single COVID-NET by 5.39%.