UNIVERSITÄT BIELEFELD
Faculty of Technology

Master Thesis

Federated Learning: Client Clustering with Soft Clusters

Semantic Computing Group
Christoph Düsing
cduesing@techfak.uni-bielefeld.de

Federated Learning (FL) is a novel, privacy-preserving approach towards distributed machine learning. As such, it allows participating parties, so-called clients, to jointly train an arbitrary machine learning model without having to share data among each other. To do so, a central server orchestrates several rounds of training, each consisting of a local training performed by each client and a subsequent aggregation of all model updates. During several rounds of training, a global model is achieved that incorporates knowledge derived from all clients' data silos without directly accessing them.

In spite of its potential, the performance of FL has been repeatedly shown to suffer from data imbalance. Here, data imbalance refers to client-specific characteristics of data that differ from the federation in terms of labels, features, and quantity held. In order to overcome this issue, a vast variety of research has been conducted to improve FL in face of data imbalance. One such approach is client clustering. Instead of training a single FL model among all clients, clustering clients according to their data and than performing FL within individual clusters turns out to improve the performance for most involved clients.

However, assigning each client to a single cluster only might cause other clusters to loose access to valuable data held by some clients. Therefore, the offered thesis aims to identify to what extent using soft instead of hard client clustering improves the overall performance within the federation of clients. Therefore, existing client clustering approaches will be extended in order to allow clients to be assigned multiple clusters. The newly developed approach is compared to those baselines afterwards.

The offered thesis does not require prior knowledge of FL. However, familiarity with traditional machine learning, data preprocessing, and python are highly recommended.

Related literature

1. McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.
2. Ruan, Y., & Joe-Wong, C. (2022, June). Fedsoft: Soft clustered federated learning with proximal local updating. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 7, pp. 8124-8131).

The Semantic Computing Group researches and develops methods that enable machines to acquire relevant knowledge as well as linguistic capabilities. Using methods from *natural language understanding* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading unstructured textual data. In particular, the group focuses on methods for information extraction, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at: http://sc.cit-ec.uni-bielefeld.de.

Interested? @mail to `cduesing@techfak.uni-bielefeld.de`