netidee
STIPENDIEN

Communication and Energy Efficient Edge Artificial Intelligence Framework for Internet of Things

Endbericht | Call 18 | Stipendium ID 6801

Inhalt

1 Introduction

The research is focused on developing communication and energy efficient Edge-AI framework to overcome the limitations of centralized learning to pave the way for futuristic applications enabled by widespread adoption of smart but resource-constrained IoT devices. The literature review conducted during this work led to the discovery of the most critical issue namely the inability to cater to the needs geographically distributed clients faced by the current proposed solutions which consequently led to the development of a novel federated learning framework termed pHFedKD (personalized Hierarchical Federated Learning via Knowledge Distillations) to address such limitations. This document encompasses and presents the general information regarding the research conducted during this project, outcome of the project along with the planned future research activities, and last but not the least the set of guidelines for the interested parties including academics and the likes for extending or utilizing our work.

2 General

The proliferation of IoT sensors have given rise to a new set of futuristic applications such smart cities, smart environmental and precision medicine. These constellations of smart devices scattered across the globe generate massive amounts of data and hold potential to addressing some of the most pressing issues of the contemporary times such as climate change. Although the advancements in IoT and AI present new opportunities at the same time they pose significant challenges to traditional centralized computing due to the share volumes of the data being generated at the edge of network. It puts a massive strain on the network to transfer this much amount of data. Additionally, transporting data from remote places to a centralized server for data analysis raises service concerns about privacy and latency. This mandates developing efficient Edge-AI systems for resource-constrained devices to mitigate these challenges.

Federated learning (FL) allows distributed clients to collaboratively train a model without requiring them to share their private data. Each client is supposed to keep their data private instead only shares the weights of its locally trained model. After a few rounds of local training, clients share their updates with the global parameter server which aggregates these updates from the clients to generate the global model. FL works well in homogeneous settings in which clients' data are sampled from the same distribution. However, its performance deteriorates significantly in the face of data heterogeneity or more formally known as non-IID problem. As a direct consequence of this, FL struggles to be as effective in large environments with a large number of geographically dispersed clients. Moreover, frequent communication between the clients and single central server could incur huge communication costs.

The core objective of this work was to develop a framework leveraging the entire spectrum of computing continuum which addresses the aforementioned challenges. The research question that we set out to find an answer for was “can we leverage the client heterogeneity to our advantage instead of merely treating it as a problem? “. The way we do is first by quantifying the heterogeneity (diversity) among clients and later utilizing that for segregating homogeneous (similar) clients into clusters. Subsequently, the complex task of learning a single global model is divided into smaller tasks by assigning each cluster to an edge server and learning cluster-specific edge models. Hence, we develop a framework, which leverages client heterogeneity to learn individual client, cluster, and global model simultaneously.

3 Results

We introduce a hierarchical federated learning framework called pHFedKD which addresses client heterogeneity, concept drift, and scalability in FL for decentralized environments with diverse client data distributions. The proposed framework combines dynamic clustering, multi-level hierarchical aggregation, and multi-teacher knowledge distillation to optimize performance of the client, cluster, and global models.

Our key contributions made to the domain through this work are listed below:

- We introduce federated geospatial clustering (FGC) based on a novel client affinity (CA) metric to quantify data and client similarity. This dynamic clustering method adapts to spatio-temporal client mobility and concept drift by reassigning clients during training.
- We introduce a hierarchical federated learning framework which spans three computing continuum levels; client, edge, and cloud. At level 0, clients train local models using their private data. At level 1, edge server aggregate local models to generate cluster-specific models. At level 2, cloud server aggregate cluster models to yield the global model.
- We introduce multi-teacher knowledge distillation to enable inter-cluster knowledge sharing and enhance global and edge models. This ensures that clusters are able to retain their unique characteristics while simultaneously benefiting from the shared global knowledge.

We conduct through evaluations of our proposed framework using real-world CityScapes dataset for semantic segmentation in both static and dynamic settings and compare it to the state-of-the-art baselines such as FedAvg, FedProx, and HierFavg. pHFedKD outperforms baselines under both static and dynamic conditions, achieving substantial gains in model accuracy and communication efficiency. More specifically, it achieves up to

19% improvement in client-level accuracy and 24.7% in edge-level accuracy. Furthermore, it reduces communications costs up to 50% transmitting only lightweight logit vectors instead of the entire weight vectors.

4 Planned Follow-up Activities

The finished work is currently submitted for publication and under review at the IEEE International Conference on Parallel and Distributed Processing Symposium (IEEE IPDPS). The developed framework was evaluated on an autonomous vehicle urban scene understanding use case; however, it can easily be extended to other domains. Validation of its effectiveness on additional use cases is ongoing. Currently, we are adapting the framework for an intelligent farming application focused on tracking animals in remote alpine regions.

In parallel, the preparation of the dissertation is underway. The next planned steps include:

- Submitting the complete draft for internal review and incorporating feedback.
- Finalizing the dissertation and submitting it formally for evaluation.
- Coordinating with the examination committee to schedule the defense.
- Preparing for the oral defense through presentation rehearsals and potential Q&A simulations.

These steps are aimed at completing the dissertation process and achieving a successful defense in the coming months.

5 Suggestions for Continuation by Third Parties

The proposed framework can be applied in diverse real-world applications such as autonomous systems, smart cities, and precision agriculture, where heterogeneity, mobility, and data privacy and real-time requirements are critical. Furthermore, pHFedKD offers potential for interoperability studies to explore integration with existing federated learning frameworks and edge computing infrastructures.

Third parties could investigate incorporating advanced clustering methods, such as meta learning-driven adaptive clustering, to further optimize dynamic grouping. Additionally, integrating pHFedKD with privacy-preserving mechanisms, such as differential privacy or secure multi-party computation, could expand its applicability in domains with stringent regulatory requirements, like healthcare or finance. Finally, pHFedKD can be studied alongside resource optimization techniques to explore trade-offs between computational cost and personalization, particularly in energy-constrained environments like IoT devices.