Design and Implementation of Computation and Communication Efficient Distributed Machine Learning Frameworks for Edge AI

Contact
Dr Rehmat Ullah: rehmat.ullah@newcastle.ac.uk Personal website: www.rehmatkhan.com

Research project
Standard machine learning (ML) techniques require centralising the training data on cloud data centres due to the high computing resources (e.g., GPUs) available on the cloud, so that large amounts of data can be analyzed to obtain useful information for the detection, classification, and prediction of future events with high accuracy However, the proliferation of mobile devices, ranging from smartphones to autonomous vehicles, drones, and various Internet-of-things (IoT) devices such as wearable sensors and surveillance cameras, has resulted in a vast amount of data being generated. Data from all of these devices is collected in a distributed manner and sent to a central server, where it is used to train a powerful ML model. Due to network bandwidth, latency and data privacy concerns, sending all of the data to a remote cloud is impractical and often unnecessary. Furthermore, in many applications, user data contains sensitive personal information, raising privacy concerns as another reason to avoid offloading data to a centralized server. Instead of sending raw data to distant clouds, it should be processed and stored locally leveraging edge computing paradigm.

Edge artificial intelligence (Edge AI) emerged as an evolution of the edge computing paradigm, deploying AI algorithms and models directly on edge devices. Within this context the concept of federated learning (FL) provides privacy by design in an ML technique, enabling collaboratively learning across multiple distributed devices without sending raw data to a central server while processing data locally on devices. However, given the limited availability of resources on many devices, performing FL on such devices is impractical due to increased training times. During ML training, the model generates significant traffic, leading to increased communication time, which may result in longer completion times, higher training loss, and increased energy consumption. Specifically, the end devices used in FL are predominantly wireless and typically operate with limited bandwidth, such as 2G, 3G, or Wi Fi. The global and local FL model parameters use uplink and downlink transmission which depends on the bandwidth resource block allocation, fading, and interference from others. Exchanging model parameters over such lossy network may results in challenges such as transmission delay which impact the convergence time of the model and packet losses which affect the model’s accuracy. Therefore, trimming the communication time for ML training is of upmost importance for fast convergence of ML models.

FL especially Split FL generates bursty traffic which may lead to increased completion time, training loss, energy consumption and reliability. In the literature, the communication cost is reduced by reducing the size of the parameters using techniques, such as quantization, compression, pruning and sparsification under the assumption that the network is perfect and free of errors. However, packet loss and recovery remain significant challenges in realistic network settings, particularly in wireless scenarios.

Highly motivated student is expected to develop efficient distributed machine learning frameworks at the edge, focusing on securing user data with FL techniques, reducing computation overhead, and addressing communication challenges through techniques like semantic communication, quantization, compression, and sparsification. Topics include (but not limited to) algorithms for distributed ML, adaptive techniques for changing network conditions, edge AI resilience, benchmarking and running generative models at the edge, optimizing neural networks, semantic communications, asynchronous FL training, and empirical evaluations using real testbeds.

Supervision Environment

Student will be supervised by researchers at the National Edge AI Hub at Newcastle University. It is the only national research centre in the UK, funded by the UK government, that addresses critical challenges in deploying AI at the edge and its continuum. Students benefit from working alongside world-renowned distributed systems and AI experts, contributing to impactful research projects. The Hub's strong partnerships with academia and industry provide access to state-of-the-art technologies and resources, enabling students to test and implement their solutions in high-impact, real-world scenarios.

Applicant skills/background

This project involves working at the intersection of machine learning, cloud-edge computing, and IoT. The ideal candidate should have strong programming and analytical skills, particularly in Python, and experience with machine learning algorithms suited for cloud-edge, mobile, or IoT environments. Experience in prototyping and testbed development for the cloud-edge-device continuum is a plus.

References

1. D.Wu,R.Ullah,P.Rodgers,P.Kilpatrick,I.SpenceandB.Varghese,"EcoFed: Efficient Communication for DNN Partitioning-Based Federated Learning," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 35, no. 3, pp. 377- 390, March 2024

2. R.Ullah,D.Wu,P.Harvey,P.Kilpatrick,I.SpenceandB.Varghese,"FedFly: Toward Migration in Edge-Based Distributed Federated Learning," in *IEEE Communications Magazine,* vol. 60, no. 11, pp. 42-48, November 2022

3. D.Wu,R.Ullah,P.Harvey,P.Kilpatrick,I.SpenceandB.Varghese,"FedAdapt: Adaptive Offloading for IoT Devices in Federated Learning," in *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 20889-20901, 1 Nov.1, 2022

4. G.Cleland,D.Wu,R.UllahandB.Varghese,"FedComm:Understanding Communication Protocols for Edge-based Federated Learning," *IEEE/ACM 15th International Conference on Utility and Cloud Computing (UCC)*, Vancouver, WA, USA, 2022