2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT) | 978-1-6654-6090-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDCAT56447.2022.00031

Communication and Energy Efficient Edge Intelligence

Sabtain Ahmad
Institute of Information Systems Engineering
Vienna University of Technology
Vienna, Austria
sabtain.ahmad@tuwien.ac.at

***Abstract*—Driven by the rapid advances in Artificial Intelligence of Things (AIoT), billions of mobile and IoT devices are connected to the internet, generating huge quantities of data at the network edge. Meanwhile, traditional analytics approaches such as cloud computing and centralized AI are unable to manage these massively distributed heterogeneous data primarily because 1) moving a tremendous amount of data across the network poses severe challenges to network capacity 2) cloud-based analytics can result in prohibitively high transmission delays 3) transporting data containing private information over the network poses serious concerns for the privacy and may not even be possible due to regulations like GDPR. Accelerated by the success of AI and IoT technologies, there is an urgent need to push AI to the network edge to tap the full potential of big data.**

***Index Terms*—edge intelligence, machine learning, federated learning, collaborative learning**

I. Background

The fusion of AI and edge computing has resulted in a new paradigm called Edge Intelligence (EI) or Edge AI; it pushes the intelligence closer to where the data is being produced by enabling the deployment of machine learning algorithms to edge devices. On the one hand, edge computing defines the process of extending computing, storage, and communication resources from a centralized cloud server to the edge of the network with the aim of improving the quality of service (QoS) for mission-critical applications such as connected vehicles, smart grids, and intelligent traffic management [1]. On the other hand, the widespread use of AI has revolutionized a number of application areas, such as autonomous cars, security and surveillance, time-series forecasting, and language modeling. The confluence of AI and edge computing has resulted in the emergence of EI.

II. Problem Statement

Critical applications such as smart health, internet of vehicles (IoVs), and smart environmental monitoring (SEM) may require performance and privacy assurances, low latency and faster inference, bandwidth and energy efficiency, and scalability. In smart environmental monitoring systems, near real-time analytics are used for condition monitoring, such as water quality monitoring, helping such systems react promptly to hazardous materials. In these types of applications, failure to provide timely decisions will bring catastrophic outcomes; for instance, in the case of water quality monitoring, delayed response to the presence of a hazardous chemical may have severe health and ecological consequences. Another pressing issue that is core to such applications is privacy; for example, multiple hospitals that might be interested in collaboratively training a model may not be inclined or even allowed to share their private data. The edge AI can respond to these low latency and strict privacy requirements due to its geographically distributed nature in close proximity to users.

III. Related Work

FedAvg algorithm [2] provides a simple yet effective approach to keep the training data private since only global model θ_t and local model θ^k is allowed to be communicated between the server and the clients, all the training data are kept on the user devices without being accessed either by the server or other clients. However, the back-and-forth communication between clients and the server means the network can quickly become an overhead. Client updates sampling [3], model compression [4], and quantization [5] are some of the most common strategies for reducing the communication size.

IV. Overview of Proposed Work

We identify several scenarios that could benefit from the federation of entities/devices to collaboratively train a model, such as text-to-speech synthesis, condition monitoring, virtual personal assistant, navigator, and image/video recognition. To illustrate with an example, smart environmental monitoring requires deploying tens of hundreds of IoT sensors to monitor an area of interest. Let us assume that none of these resource-constrained sensors have the capacity to train a model independently. Additionally, training a model per device might lead to overfitting. The processing time would be reduced significantly since the devices will only be communicating aggregates over the network rather than transferring the raw data to the server.

Fig. 1 depicts the proposed Edge AI architecture for the smart environmental monitoring use case. The first step requires the deployment of IoT sensors at the most critical and informative locations. The next step entails the deployment of static or mobile gateways for transferring the measurements from sensors to edge nodes. The final step involves training and deploying machine learning models on edge nodes through the edge-cloud collaboration for processing the collected data

978-1-6654-6090-3/22/$31.00 ©2022 IEEE
DOI 10.1109/BDCAT56447.2022.00031

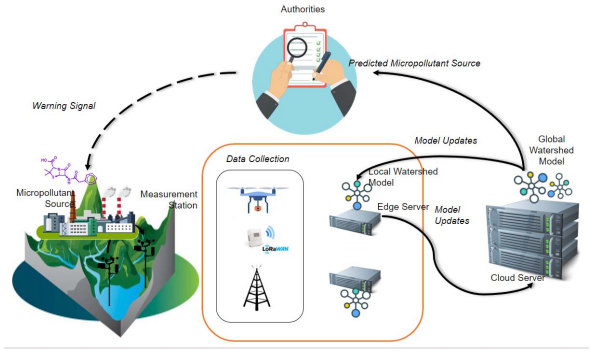

Fig. 1. Smart environmental monitoring through Edge AI.

and detecting the presence of micropollutants in order to facilitate the authorities to take swift actions.

V. Preliminary Results

A. Sustainable Environmental Monitoring via Energy and Information Efficient Multi-Node Placement

Sustainable deployment of sensors for environmental sampling is a challenging task because of the spatial and temporal variation of the environmental attributes to be monitored, the lack of the infrastructure to power the sensors for uninterrupted monitoring, and the large continuous target environment despite the sparse and limited sampling locations.

In this work, we present an environment monitoring framework that deploys a network of sensors and gateways connected through low-power, long-range networking to perform reliable data collection. The three objectives correspond to the optimization of information quality, communication capacity, and sustainability. Therefore, the proposed environment monitoring framework consists of three main components: (i) to maximize the information collected, we propose an optimal sensor placement method based on QR decomposition that deploys sensors at information and communication-critical locations; (ii) to facilitate the transfer of big streaming data and alleviate the network bottleneck caused by low bandwidth, we develop a gateway configuration method with the aim to reduce the deployment and communication costs; and (iii) to allow sustainable environmental monitoring, an energy-aware optimization component is introduced. We validate our method by presenting a case study for monitoring the water quality of the Ergene River in Turkey. Detailed experiments subject to real-world data show that the proposed method is both accurate and efficient in monitoring a large environment and catching up with dynamic changes.

B. FedCD: Personalized Federated Learning via Collaborative Distillation [6]

Federated learning enables the creation of a centralized global model by aggregating updates from the locally trained models from multiple clients. While powerful, such an architecture is limited to applications where the needs of heterogeneous clients can be served by a single global model. It does not cater to the scenarios where each client independently designs its own model. Task and data heterogeneity inherent to such scenarios demand each client to specialize in the local setting while still being able to collaborate and transfer the acquired knowledge to the rest of the federation without sharing the data or the model.

In this work, we utilize ensemble and collaborative learning to design a framework that enables the training of personalized models for heterogeneous clients with different learning capacities using federated learning. Empirical evaluations performed on the CIFAR100 dataset demonstrate that our framework is able to consistently improve the performance of all the participating models and outperform the independently trained models on the complete training set without collaboration. We analyze that all participants benefit from collaborative distillation and boast an average 1.4% increase in performance. Moreover, a comparison with the state-of-the-art approaches demonstrates that our framework outperforms the Federated Learning and Federation Distillation methods by up to a 2× increase in the average global accuracy.

VI. Planned Work

The path to edge intelligence is filled with numerous obstacles, such as i) restrictions regarding privacy and latency render centralized model training methods essentially ineffective, ii) increased energy consumption exacerbated by decentralized model training, data transmission, and storage of big data, and iii) system and data heterogeneity inherent to most IoT applications deteriorate the performance of the system. We plan to implement novel and autonomic solutions to the aforementioned challenges.

Acknowledgments

The work presented in this paper has been supported by the CHIST-ERA grant CHIST-ERA-19-CES-005 and by the Austrian Science Fund (FWF): I 5201-N.

References

[1] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1508–1532, 2019.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[4] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[5] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[6] S. Ahmad and A. Aral, "FedCD: Personalized federated learning via collaborative distillation," in *2nd Workshop on Distributed Machine Learning for the Intelligent Computing Continuum (DML-ICC)*, 2022, (accepted).