Efficient Machine Learning Systems in Edge Cloud Environments

In-Young Ko[1], Michael Mrissa[2,3], Juan Manuel Murillo[4,5] and Abhishek Srivastava[6]

[1]*Korea Advanced Institute of Science and Technology, South Korea*
[2]*InnoRenew CoE, Slovenia*
[3]*University of Primorska, Slovenia*
[4]*University of Extremadura, Spain*
[5]*COMPUTAEX Foundation, Spain*
[6]*Indian Institute of Technology Indore, India*
E-mail: iko@kaist.ac.kr; michael.mrissa@innorenew.eu; juanmamu@unex.es; asrivastava@iiti.ac.in

The international workshop on Big Data-Driven Edge Cloud Services (BECS) aims to provide a platform for scholars and practitioners to share their experiences and present ongoing work in developing data-driven AI applications and services within distributed computing environments, commonly referred to as the edge cloud.

The fourth edition of the workshop (BECS 2024)[1] was held in conjunction with the 24th International Conference on Web Engineering (ICWE 2024)[2], which took place in Tampere, Finland, from 17–20 June 2024. This special issue of the Journal of Web Engineering focuses on enhancing the efficiency of machine learning (ML)-based systems by leveraging the unique features of distributed edge cloud environments. For this issue, we selected papers from BECS 2024 that propose conceptual frameworks to improve the performance and privacy of ML-based systems and explore distributed ML-based solutions for addressing real-world challenges.

[1]https://becs.kaist.ac.kr/iwbecs2024/
[2]https://icwe2024.webengineering.org/

Federated learning (FL) has emerged as a promising approach for enhancing the performance and privacy of ML-based systems. FL enables collaborative model training without directly sharing data, making it particularly well-suited to the distributed nature of edge cloud environments.

The first article, titled "Personalized User Models in Real-world Edge Computing Environments: A Peer-to-peer Federated Learning Framework," by Xiangchi Song et al., highlights the limitations of traditional FL in large-scale edge cloud collaborative Internet of Things (IoT) systems, particularly in terms of computational and communication efficiency. To address these challenges, the authors propose a novel FL framework that enables not only vertical edge cloud collaboration but also horizontal peer-to-peer (P2P) collaboration. Using a real-world dataset, they evaluated their proposed framework and demonstrated that the P2P-extended approach significantly improves communication efficiency for FL in edge-cloud environments.

In their article titled "Privacy and Performance in Virtual Reality: The Advantages of Federated Learning in Collaborative Environments," Daniel Flores-Martin *et al.* explore the effective integration of FL with virtual reality (VR). The authors investigated the accuracy and efficiency of training FL models directly on VR devices, leveraging user data to develop personalized models. Through experiments conducted on various VR devices, they demonstrated that the proposed FL approach is both efficient in execution time and optimized for battery consumption, making it well-suited for VR environments.

Infrastructure-as-code (IaC) is a paradigm that automates infrastructure management through machine-readable configuration files. It enables seamless scaling and reproducibility of services deployed in edge-cloud environments. In their article, "Code Smell-guided Prompting for LLM-based Defect Prediction in Ansible Scripts," Hyunsun Hong *et al.* leverage large language models (LLMs) to predict defects in Ansible IaC scripts. They introduce a code smell-guided prompting (CSP) method to enhance LLM performance in defect prediction. The authors evaluated various prompting strategies, including zero-shot, one-shot, and chain of thought CSP (CoT-CSP), demonstrating how incorporating code smell information can improve defect detection.

A wireless sensor network (WSN) is a specialized type of edge cloud environment consisting of small devices, known as motes, distributed to form an ad-hoc network over a designated region of interest. In their article "Overcoming Terrain Challenges with Edge Computing Solutions: Optimizing WSN Deployments Over Obstacle Clad, Irregular Terrains," Shekhar

Tyagi and Abhishek Srivastava address the challenges of deploying motes in irregular terrains with various obstacles. They propose a novel approach that involves extracting satellite images of the region of interest (RoI) and processing them locally at the edge. This edge computing solution enables real-time decision-making at the network edge, allowing for deterministic mote deployment in diverse terrain conditions. The authors evaluated their approach by implementing deployments in real-world environmental conditions and comparing their results with existing techniques.

The transportation and automotive industry is undergoing a significant shift toward software-defined vehicles (SDVs), which continuously interact with software platforms built on edge cloud environments. In the article "Software Practice and Experience on a Smart Mobility Digital Twin in the Transportation and Automotive Industry: Toward an SDV-empowered Digital Twin through an EV Edge Cloud and AutoML," Jonggu Kang examines the development and application of a smart mobility digital twin. This digital twin is based on data collected from electric vehicles (EVs), which are evolving into SDVs. The author presents a case study showcasing the creation of an EV edge cloud data platform. This platform continuously collects and analyzes data generated by EVs and leverages automated machine learning (AutoML) for predictive modeling.

The previously discussed articles explore various ways to enhance the efficiency of ML-based systems in edge cloud environments. They offer not only theoretical and technical advancements but also valuable insights and real-world examples drawn from practical industry experiences.

Acknowledgment

The BECS 2024 workshop was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2024-2020-0-01795) supervised by the Institute of Information & Communications Technology Planning Evaluation (IITP).

Michael Mrissa gratefully acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Regional Development Fund).