Report from Dagstuhl Seminar 23432

Edge-AI: Identifying Key Enablers in Edge Intelligence

Aaron Ding*[1], Eyal de Lara*[2], Schahram Dustdar*[3], Ella Peltonen*[4], and Tobias Meuser†[5]

1 **TU Delft, NL.** aaron.ding@tudelft.nl
2 **University of Toronto, CA.** delara@cs.toronto.edu
3 **TU Wien, AT.** dustdar@dsg.tuwien.ac.at
4 **University of Oulu, FI.** Ella.Peltonen@oulu.fi
5 **TU Darmstadt, DE.** tobias.meuser@kom.tu-darmstadt.de

Abstract

Edge computing promises to decentralize cloud applications while providing more bandwidth and reducing latency. Based on the discussion of our first Dagstuhl Seminar and the continuation work that took place after the seminar, we continued our work on identified challenges that need to be further addressed within the community. These challenges included 1) large-scale deployment of the edge-cloud continuum, 2) energy optimization and sustainability of such large-scale AI/ML learning and modelling, and 3) trustworthiness, security, and ethical questions related to the whole continuum. In this seminar, we discussed the current state of Edge Intelligence and shaped a holistic view of its challenges and applications. The main concerns were 1) the assessment and applicability of Edge Intelligence solutions, 2) energy consumption and sustainability, and 3) the new trend of Large-Language Models.

Seminar October 22–25, 2023 – http://www.dagstuhl.de/23432
2012 ACM Subject Classification Computer systems organization → Distributed architectures; Computing methodologies → Artificial intelligence; Networks
Keywords and phrases cloud computing, edge computing, edge intelligence
Digital Object Identifier 10.4230/DagRep.13.10.130

1 Executive Summary

Aaron Ding (TU Delft, NL)
Eyal de Lara (University of Toronto, CA)
Schahram Dustdar (TU Wien, AT)
Ella Peltonen (University of Oulu, FI)

Research Area

Edge computing promises to decentralize cloud applications while providing more bandwidth and reducing latency. These promises are delivered by moving application-specific computations between the cloud, the data-producing devices, and the network infrastructure components at the edges of wireless and fixed networks. Meanwhile, the current Artificial Intelligence (AI) and Machine Learning (ML) methods assume computations are conducted in a powerful computational infrastructure, such as data centres with ample computing and data storage resources available. To shed light on the fast-evolving domain that merges edge

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Edge-AI: Identifying Key Enablers in Edge Intelligence, *Dagstuhl Reports*, Vol. 13, Issue 10, pp. 130–138
Editors: Eyal de Lara, Aaron Ding, Schahram Dustdar, and Ella Peltonen
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

computing and AI/ML, referred to as Edge AI, the recent Dagstuhl Seminar 21342 gathered community inputs from a diverse range of experts. The results of our first iteration of the seminar were reflected in the ACM SIGCOMM CCR publication, focusing on three different angles of Edge AI: future networking, cloud computing, and AI/ML needs.

Along with three identified driving areas of 5G beyond (or so-called 6G), future cloud, and evolved AI/ML, the advancement of different technologies and the growing business interests will take Edge AI forward regarding hardware, software, service models, and data governance. Starting from the current state of play driven by cellular, cloud, and AI/ML service providers, the roadmap reflects five general phases: scalable framework, trustworthy co-design, sustainable and energy-efficient deployment, equal accessibility, and pervasive intelligent infrastructure. As changes can always occur, the sequence depicted in the roadmap could be switched or combined. Nonetheless, this Edge AI roadmap reflects the combined effects of technology enablers and non-tech demands, such as the socioeconomic transformation of user behaviors, purchasing power, and business interests.

Despite its promise and potential, Edge AI still faces major challenges in large-scale deployment, including energy optimization, trustworthiness, security, privacy, and ethical issues. As an important goal of sustainability, the energy consumption of Edge AI needs to be optimized. Energy efficiency is crucial for Edge AI embedded infrastructures (e.g., roadside units, micro base stations) to sustainably support advanced autonomous driving and Extended Reality (XR) services in the years to come. Through the pipeline of data acquisition, transfer, computation, and storage, there exists the possibility for Edge AI to trade accuracy with reduced power and less time consumed. For instance, noisy inputs from numerous sensors can be selectively processed and transferred in order to save energy.

A set of applications would be satisfied with an ‘acceptable’ accuracy instead of exact and absolutely correct results. By introducing this new dimension of accuracy to the optimization design, energy efficiency can be further improved. Concerning trustworthiness, Edge AI benefits from its close proximity to the end devices. However, due to the distributed deployment with deep insights into a personal context, the safety and perceived trustworthiness of Edge AI services are raising concerns among the stakeholders (e.g., end-users, public sectors, ISP). To achieve trustworthy Edge AI, critical building blocks are needed, including verification and validation mechanisms that ensure transparency and explainability, especially in the training and deployment of Edge AI in decentralized, uncontrolled environments. The trustworthiness of Edge AI is a stepping stone to establishing appropriate governance and regulatory framework on which the promise of Edge AI can be built.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

Panel discussions

3 Overview of Talks

3.1 Future of Communications: Why we need Edge AI and more

Susan Bayhan (University of Twente – Enschede, NL)

In the era of the climate crisis, it is essential to optimize the operation of many sectors, from agriculture to health with the help of smart sensing and data analytics. While the data is essential for better understanding and consequently for better decision-making, it has to be moved, stored, computed, secured, and interpreted, all steps bringing their own challenges. According to the Decadal Plan by Semiconductors (`https://www.src.org/about/decadal-plan/`), there are five seismic shifts that need to be considered for the future of ICT: 1) Analog hardware for generating smarter world-machine interfaces that can sense, perceive, and reason, 2. Radically new memory and storage solutions. 3. New research directions to close the gap between communication capacity vs. data-generation rates. 4. security challenges in highly interconnected systems and AI, and 5. New computing paradigms for higher energy efficiency. The talk argues that edge AI can help with some of these challenges, such as by decreasing the need to communicate huge volumes of data with the cloud, however for sustainable communications systems of the future, more is needed at all levels of the system stack as well as design, production, and use stages of all infrastructures.

3.2 Increasing AI Sustainability with Symbolic Data Representation on the Edge

Ivona Brandic (Technische Universität Wien, AT)

Joint work of Daniel Hofstätter, Shashikant Ilager, Ivan Lujic, Ivona Brandic
Main reference Daniel Hofstätter, Shashikant Ilager, Ivan Lujic, Ivona Brandic: "SymED: Adaptive and Online Symbolic Representation of Data on the Edge", in Proc. of the Euro-Par 2023: Parallel Processing – 29th International Conference on Parallel and Distributed Computing, Limassol, Cyprus, August 28 – September 1, 2023, Proceedings, Lecture Notes in Computer Science, Vol. 14100, pp. 411–425, Springer, 2023.
URL https://doi.org//10.1007/978-3-031-39698-4_28

The Edge Computing model is beneficial for analyzing and processing data generated by the Internet of Things (IoT) in the proximity to its source. However, the transfer, storage, and processing of this rapidly increasing data volume is challenging on edge devices with limited resources. Symbolic Representation (SR) algorithms show promise in reducing data size by transforming raw data into symbols. They also enable direct data analytics on symbols, such as anomaly detection and trend prediction, which is advantageous for many edge applications. Nonetheless, the current SR algorithms are mainly centralized and operate offline with batch data, making them unsuitable for real-time scenarios. In this talk, SymED – Symbolic Edge Data representation method is introduced. This method is an online, adaptive, and distributed approach to symbolic data representation at the edge. SymED utilizes Adaptive Brownian Bridge-based Aggregation (ABBA) and involves low-powered IoT devices sending and performing the low-cost initial data compression while the more capable edge devices perform the more demanding symbolic conversion.

3.3 Edge Enabled Autonomous Driving and Mobility Services

Liam Pedersen (Nissan North America – Santa Clara, US)

Connectivity, smart infrastructure and autonomous driving software hosted on edge computing will transform the ways in which we use and drive cars, the freeway system and the electrical grid. In this talk, examples of smart cloud-based services are presented for managing freeway congestion, low-cost autonomous valet parking and EV integration with the grid.

3.4 The economics of edge AI don't look great – or why edge computing may always be the future

Henning Schulzrinne (Columbia University – New York, US)

Edge computing for AI encompasses a wide variety of architectures. "Easy" cases include embedded systems, e.g., in most modern vehicles, or dedicated processing in sensors, e.g., cameras performing image segmentation and basic object recognition. The harder, interesting architectures provide generic computing capabilities to paying customers, i.e., cloud-like arrangements.

Even for edge AI systems, the edge computing element will likely not store large volumes of data as the computing need may be transient. Thus, the latency advantages of edge systems may be reduced since the edge system will need to contact the regional or trans-regional cloud to query databases or access API-based microservices functionality. This split functionality may negate the latency advantages of edge computing for real-world systems and needs to be part of any system evaluation.

Edge AI systems are further challenged by economic considerations. The cost of computing can be divided into capital costs, typically the initial investment into computing and layer-0 infrastructure such as data centers, and the ongoing operational costs. AI-suitable GPUs may have shorter effective lifespans than other CPUs, but even server CPUs are typically only used for five years before being retired. Thus, edge AI systems with low utilization may increase the amortized costs on a per-task basis. For the same probability of task rejection, smaller edge systems need to provide more computational reserves by standard Erlang-C considerations.

Electricity makes up about 60-70% of the operational cost. Edge AI can only be competitive if the energy costs are similar to those in large-scale data centers, i.e. if Edge AI can draw on cheap renewable energy. (However, their intermittency may increase the amortized cost of capital, as noted above.) For example, grid electricity in New York costs roughly $0.21/kWh, while data centers aim for $0.05/kWh. As of June 2021, the levelized cost of energy (LCOE) for utility-scale photovoltaic systems ranges from $0.03 to $0.05 and is thus competitive.

Other operational costs, not further discussed here, include security costs if edge systems impose a security premium, development and DevOps costs, as well as failure risk trade-offs.

Thus, edge computing for AI may be roughly divided into "cheap" and "dependable." The former may only offer batch-style processing with intermittent availability, while the latter is willing to tolerate higher costs than traditional cloud computing.

Trust for edge AI deserves more careful consideration. Unless the entity running the computation owns and operates the hardware, they still have to trust the edge computing provider, which may well be smaller than typical cloud providers.

As cloud services now offer a wide range of computing hardware, from traditional i386 to ARM and GPU-based processors as well as specialized ML engines, edge computing will struggle to compete, given the smaller rack count.

Smaller edge installations may also find it more difficult to provide physical security and uninterrupted power. In summary, given the uncertainties of economic competitiveness, security, and reliability, edge AI requires careful feasibility analysis, where non-technical considerations may outweigh technical feasibility or advantages. Resource discovery needs to take cost and reliability needs into account. Resources may well be mediated and aggregated to relieve application developers from maintaining and creating relationships with hundreds of edge computing service providers. To facilitate computational roaming, systems have to provide appropriate AAA capabilities.

3.5 Enabling data spaces: existing developments and challenges

Gürkan Solmaz (NEC Laboratories Europe – Heidelberg, DE)

This talk at Dagstuhl includes a short introduction to the concept of Data Spaces based on the recent developments of IDSA, Gaia-X, and FIWARE, as well as the challenges of data interoperability and data value. The recent work from the Data Ecosystems and Standards (DES) group at NEC Laboratories Europe focuses on solving those challenges using real-world sensor data and geographic data from the case studies of the City Liveability Index (CLI) of SALTED project, Smart Campus Murcia, and Humanitarian Landmine project. The data enrichment and contextualization platform is utilized in the case studies through technologies such as TrioNet, FIWARE Scorpio Broker, FIWARE FogFlow and AI/machine learning for predictions and transfer learning.

4 Working groups

4.1 Definition and Usecases of Edge AI

Dewant Katare (TU Delft, NL), Eyal de Lara (University of Toronto, CA), Aaron Ding (TU Delft, NL), Schahram Dustdar (TU Wien, AT), Tobias Meuser (TU Darmstadt, DE), Shishir Girishkumar Patil (University of California – Berkeley, US), and Ella Peltonen (University of Oulu, FI)

In this working group, the definition and use cases of Edge AI were discussed. The discussions highlighted the complexity of clearly defining edge, highlighting some key points:

- **Application-specific Definition**: Depending on the considered application, the definition of Edge varies drastically, ranging from small data centers close to the user to end devices under the control of the user.

- **Business Models**: The impact of business models on Edge AI's popularity and development is acknowledged. Cloud-based models, especially those driven by advertisement, have been given more attention, as the management is much easier and availability is much higher.
- **Real-world Examples and Use-cases**: Examples from Amazon, like Greengrass and Sagemaker, demonstrate Edge AI applications, but these are not always orchestrated for end-users. Edge solutions often complement cloud solutions, providing benefits like reduced latency and enhanced privacy without replacing cloud infrastructure.
- **Cloud Definition and Academic Consensus**: Similar to Edge AI, the definition of "Cloud" also varies widely, ranging from proximity (in milliseconds) to computing resources. There is still no universally agreed-upon definition of Edge in academia, leading to inconsistencies.
- **Perspective-Dependent Success**: The success or failure of Edge AI is contingent on the specific definition of Edge used and the viewpoint from which it is considered.

4.2 Ecosystem: Software and Hardware Problems

Dewant Katare (TU Delft, NL), Eyal de Lara (University of Toronto, CA), Aaron Ding (TU Delft, NL), Schahram Dustdar (TU Wien, AT), Nitinder Mohan (TU München, DE), Shishir Girishkumar Patil (University of California – Berkeley, US), and Ella Peltonen (University of Oulu, FI)

One of the major problems discussed in this working group is the issues associated with trust in Edge Intelligence. While the group acknowledged that edge computing is promising for handling data in public spaces, trust remains a critical issue. This is evident in the disagreement among unions over video cameras, indicating that edge computing alone is not sufficient to establish trust. One issue is the possibility of reconfiguring and reprogramming edge devices such that any functionality can be added at any point. While privacy or trust is not the sole argument for the usage of edge intelligence, a key argument is the volume of data that needs to be processed or transmitted. Edge computing allows data processing closer to where it is generated, reducing the need for data transmission and potentially enhancing privacy and efficiency.

Overall, the group highlighted the potential of edge computing in various domains, emphasizing its role in data volume management, privacy preservation, and compliance with legal and regulatory constraints. However, trust needs to be built and programmable devices might need to be regulated to prevent arbitrary reconfiguration that ultimately harms trust in these devices.

4.3 Measure what matters

Dewant Katare (TU Delft, NL), Eyal de Lara (University of Toronto, CA), Aaron Ding (TU Delft, NL), Schahram Dustdar (TU Wien, AT), Tobias Meuser (TU Darmstadt, DE), Shishir Girishkumar Patil (University of California – Berkeley, US), and Ella Peltonen (University of Oulu, FI)

This working group discussed the challenges and considerations in measuring and managing energy consumption in Edge Intelligence, with a focus on data centers and end devices.

There have been major discussions on the energy consumption of Edge Intelligence. The exponential growth in energy consumption for IT is widely criticized and revised. This is associated with inaccuracies in technological forecasts, such as those by MIT for self-driving technology. The concept of distinguishing between fungible and non-fungible energy is introduced, suggesting that sometimes it might be better to turn off an energy source or consume it, depending on its nature. There is optimism with new benchmarks emerging, which help in understanding the carbon footprint of various technologies. However, there is a need to precisely define and break down what is being measured in terms of energy consumption. As this varies across industries, the lack of clear benchmarks or standards is a problem. Without these proper benchmarks, the discussions are not scientifically robust.

In summary, the discussions emphasise the complexities in measuring and managing energy and carbon footprint in Edge Intelligence, emphasizing the need for precise benchmarks, consideration of regional differences, and the importance of trend analysis in the face of challenging measurements.

5 Panel discussions

5.1 What's Next after Edge AI

Henning Schulzrinne (Columbia University – New York, US), Shishir Girishkumar Patil (University of California – Berkeley, US), Liam Pedersen (Nissan North America – Santa Clara, US), and Jan Rellermeyer (Leibniz Universität Hannover, DE)

In this panel, the future of Edge AI has been discussed. One discussion point was the vagueness of the term Edge in the research community, which sometimes leads to confusion among researchers. However, this does not limit the success of some applications of Edge and Edge Intelligence. A major point in these discussions has been the role of Large-Language-Models (LLMs) in Edge devices. This also led to the discussion of how LLMs could be used as Operating Systems and their future role in research.

Participants

- Atakan Aral
Universität Wien, AT
- Susan Bayhan
University of Twente – Enschede, NL
- Christian Becker
Universität Stuttgart, DE
- Monowar Bhuyan
University of Umeå, SE
- Ivona Brandic
Technische Universität Wien, AT
- Eyal de Lara
University of Toronto, CA
- Kemal A. Delic
The Open University – Milton Keynes, GB
- Aaron Ding
TU Delft, NL
- Schahram Dustdar
TU Wien, AT
- Janick Edinger
Universität Hamburg, DE
- James Gross
KTH Royal Institute of Technology – Kista, SE
- Volker Hilt
Nokia Bell Labs – Stuttgart, DE
- Dewant Katare
TU Delft, NL
- Lauri Lovén
University of Oulu, FI
- Tobias Meuser
TU Darmstadt, DE
- Nitinder Mohan
TU München, DE
- Shishir Girishkumar Patil
University of California – Berkeley, US
- Liam Pedersen
Nissan North America – Santa Clara, US
- Andy D. Pimentel
University of Amsterdam, NL
- Jan Rellermeyer
Leibniz Universität Hannover, DE
- Tina Rezaei
University of Twente – Enschede, NL
- Etienne Rivière
UC Louvain, BE
- Henning Schulzrinne
Columbia University – New York, US
- Stephan Sigg
Aalto University, FI
- Pieter Simoens
Ghent University, BE
- Gürkan Solmaz
NEC Laboratories Europe – Heidelberg, DE
- Michael Welzl
University of Oslo, NO
- Lars Wolf
TU Braunschweig, DE