Report from Dagstuhl Seminar 21342

Identifying Key Enablers in Edge Intelligence

Edited by

Aaron Ding[1], Ella Peltonen[2], Sasu Tarkoma[3], and Lars Wolf[4]

1 **TU Delft, NL,** aaron.ding@tudelft.nl
2 **University of Oulu, FI,** ella.peltonen@oulu.fi
3 **University of Helsinki, FI,** sasu.tarkoma@helsinki.fi
4 **TU Braunschweig, DE,** wolf@ibr.cs.tu-bs.de

Abstract

Edge computing, a key part of the 5G networks and beyond, promises to decentralize cloud applications while providing more bandwidth and reducing latencies. The 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. However, the current AI/ML methods assume computations are conducted in a powerful computational infrastructure, such as a homogeneous cloud with ample computing and data storage resources available. In this seminar, we discussed and developed presumptions for a comprehensive view of AI methods and capabilities in the context of edge computing, and provided a roadmap to bring together enablers and key aspects for edge computing and applied AI/ML fields.

Seminar August 22–25, 2021 – http://www.dagstuhl.de/21342
2012 ACM Subject Classification Computer systems organization → Distributed architectures; Computing methodologies → Artificial intelligence; Networks
Keywords and phrases artificial intelligence, communication networks, edge computing, intelligent networking
Digital Object Identifier 10.4230/DagRep.11.7.76
Edited in cooperation with Atakan Aral, atakan.aral@univie.ac.at

1 Executive Summary

Aaron Ding
Ella Peltonen
Sasu Tarkoma
Lars Wolf

Research Area

Edge computing, a key part of the upcoming 5G mobile networks and future 6G technologies, promises to decentralize cloud applications while providing more bandwidth and reducing latencies. The 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. The previous works have shown that edge computing devices are capable of executing computing tasks with high energy efficiency, and when combined with comparable computing power to server computers.

In stark contrast to the current edge-computing development, current artificial intelligence (AI) and in particular machine-learning (ML) methods assume computations are conducted in a powerful computational infrastructure, such as a homogeneous cloud with ample computational and data storage resources available. This model requires transmitting data from end-user devices to the cloud, requiring significant bandwidth and suffering from latency. Bringing computation close to the end-user devices would be essential for reducing latency and ensuring real-time response for applications and services. Currently, however, these benefits cannot be achieved as the perspective of "edge for AI", or even "communication for AI", has been understudied. Indeed, previous studies address AI only limitedly in different perspectives of the Internet of Things, edge computing, and networks.

Clear benefits can be identified from the interplay of ML/AI and edge computing. We divide this interplay into edge computing for AI and AI for edge computing. Distributed AI functionality can further be divided into edge computing for communication, platform control, security, privacy, and application or service-specific aspects. Edge computing for AI centres on the challenge of adapting the current centralized ML and autonomous decision-making algorithms to the intermittent connectivity and the distributed nature of edge computing. AI for edge computing, on the other hand, concentrates on using AI methods to improve the edge applications or the functionalities provided by the edge computing platform by enhancing connectivity, network orchestration, edge platform management, privacy or security, or providing autonomy and personalized intelligence on application level.

Previous studies address accommodating AI methods for different perspectives of IoT, edge computing and networks. However, there is still a need to understand the holistic view of AI methods and capabilities in the context of edge computing, comprising for example predictive data analysis, machine learning, reasoning, and autonomous agents with learning and cognitive capabilities. Further, the edge environment with its opportunistic nature, intermittent connectivity, and interplay of numerous stakeholders present a unique environment for deploying such applications based on computations units with different degrees of intelligence capabilities.

The AI methods used in edge computing can be further divided into learning and decision making. Learning refers to building, maintaining and making predictions with ML models, especially neural networks. Decision making is the business logic, that is, the process of acting upon the predictions. This is the domain of decision theory, control theory and game theory, whose solutions and equilibrium are now often estimated with data by reinforcement learning methods.

Currently, AI's cloud-centric architecture requires transmitting raw data from the end-user devices to the cloud, introducing latencies, endangering privacy and consuming significant data transmission resources. The next step, currently under active research, is distributed or federated AI, which builds and maintains a central model in the cloud or on the edge but allows user devices to update the model and use it locally for predictions. We envision a fully decentralized AI which flattens the distributed hierarchy, with the joint model built and maintained by devices, edge nodes and cloud nodes with equal responsibility. The present challenges for AI in edge computing converge on 1) finding novel neural network architectures and their topological splits, with the associated training and inference algorithms with fast and reliable accuracy and 2) distributing and decentralized model building and sharing into the edge, by allowing local, fast-to-build personalized models and global, collaborative models, and information sharing. Finally, the novel methods need to be 3) integrated with key algorithmic solutions to be utilised in edge-native AI applications. The ground-breaking objectives and novel concepts edge-native artificial intelligence brings are:

- Edge-native AI can be used for obtaining higher quality data from massive Internet of Things, Web of Things, and other edge networks by filtering out large volumes of noise, context labelling, dynamic sampling, data cleaning, etc. High-quality data can thus be used to feed both edge inferencing and cloud-based data analysis systems, for example, training large-scale machine learning models.
- Edge computing provides low latency that is crucial especially for real-time applications, such as anything related to driving and smart mobility. AI applications on the edge and thus closer to the end-user will not only fasten existing applications but also provide opportunities for novel and completely new solutions.
- Edge-based computing provides data privacy when users are involved, and no need to share data to the cloud services but only the locally learned model.
- With edge-computing implemented for AI/ML model building, personalisation of such models can be done in local environments without unnecessary transmission overhead (when only local data is anyway considered for model building). Global models built in the cloud environment can be used to support these local models whenever a collaborative, large or more general model is requested.
- Edge-native AI/ML tasks provide mobility of the computation and cloudlet-like processing in the edge. In comparison to cloudlets, edge computing provides more flexibility and dynamic operations for load balancing, task management, distribution of the models, etc.
- Light-weight computation on the edge devices and local environments can enable energy savings.
- Ethical data management: edge-native AI can be used to keep data ownership control closer to the user, e.g., when computation is managed and task distribution controlled from the user's own devices, and suitable security and privacy protection methods are in use.

2 Table of Contents

3 Overview of Talks

3.1 Edge intelligence for Environmental Monitoring and Protection

Atakan Aral (Universität Wien, AT)

Joint work of EU CHIST-ERA SWAIN project partners

Contemporary machine learning algorithms are not limited by the training data available to them, but the computing power needed to process that data. This is particularly critical for the systems that learn from streaming big data and take time-critical decisions because storing the data and processing in batches is not an option.

In this talk, I introduced a use case for edge intelligence in environmental monitoring and disaster prediction. Water resource contamination substantially threatens the environment. Rapid identification of chemicals and their emission sources in watersheds is crucial for sustainable water resources management. Despite studies on the measurement of micropollutants in the water resources around Europe, efficient utilization of the data in decision-making to protect water resources from detrimental chemical pollution is currently not available. Novel Internet of Things technologies, coupled with advanced Artificial Intelligence and Edge Intelligence strategies, may provide faster and more efficient responses to these challenges in real-time reactions as well as long-term planning.

3.2 NSF AI Institute for Edge Computing Leveraging Next Generation Networks (Athena)

Yiran Chen (Duke University – Durham, US)

As the world is embracing the fifth generation of mobile networks (5G), mobile network infrastructures are facing a double whammy. On one hand, they have made bold performance promises, anticipating previously impossible mobile apps and services, and foretelling a mass deployment of Internet of Things (IoT) devices. On the other hand, their massive computation needs, conventionally carried by dedicated, specialized hardware at the base station, are projected to have to be met by multi-tenant datacenters at the edge and in the Cloud, as network operators are aggressively cutting cost, especially capital investment, and bracing for a future of cloud-native mobile networks. We seek to kindle and fuel revolution for wireless communications by tapping the ongoing revolution in Artificial Intelligence (AI). In doing so, we will develop AI-powered transformative technologies for next-generation mobile networks at the edge as well as new algorithmic and practical foundations of AI, such that the new functionalities, efficiency, scalability, security, privacy, and fairness of the AI solutions can be adopted in the next-generation wireless communications.

3.3 Edge Intelligence

Schahram Dustdar (TU Wien, AT)

In this talk I discuss the challenges ahead when researching the confluence of Internet of Things, Edge Computing, Fog Computing, and Cloud Computing. In particular we discuss the topics related to research issues in the area of AI and Edge Computing.

Introduction

Today's systems infrastructure is composed out of three building blocks: People, Software Services, and Things. There are two fundamental approaches to discuss such infrastructures. The first one is the cloud centric perspective. This basically views the Cloud as the center of the world, and everything else is connected to the cloud. Some people call it the brain. In this case, everything is centralized and the brain is the most important thing. In this view, IoT is always connected to the Cloud as all machine learning and decision making is done on the Cloud, nothing goes unnoticed by the Cloud. The second perspective is the Internet-centric view. Here, decision making, learning, model building etc. is also done on the edge of a network; partially consolidated and transferred in a federated fashion to Cloud systems, if needed. In this talk we propose to look at the whole compute continuum and utilize IoT, Edge, Fog, and Cloud in all our systems development and engineering efforts. 5G or 6G base stations in the future, they are typically general-purpose computing infrastructures, sometimes they are telecom operator-controlled pieces of equipment, sometimes they are belonging to organizations or even to individuals. Currently, the fog infrastructure base stations, for example, is one example for that. The Cloud has essentially unlimited compute and storage resources, with the full spectrum of cloud services with high availability and lower costs is another important part of the compute continuum. Now, we know that there is a new family of applications, which require extremely low latency and different levels of privacy and security that actually cannot work only with the Cloud; autonomous driving is a good example for that. One needs to have extremely low latencies. Another example is telemedicine in real time surgery. Hence, this means that you have to make sense out of the whole spectrum from the IoT via the Edge, Fog, to the Cloud.

Compute Continuum

In this talk I suggest a software-intensive Edge systems focus. We completely rethink the design and the operation of such an environment. Why is that necessary? The main reason is that we have fundamentally conflicting factors concerning the system requirements, which need to be resolved. So on the one hand side, we have latency. There is this an inherent traditional division of the Cloud and IoT, which has different time factors and performance factors, which we can manage better when we look at it from a software intensive side. Secondly, we have computation as an edge resource, which basically means that we need to use edge infrastructures similarly to cloud infrastructures to perform complex infrastructure tasks, such as safety and security. And thirdly, we have the question of locality and mobility, where we can introduce novel solutions to privacy software configuration and system evolution. So the question is, which characteristics of edge computing systems then should be abstracted as first class citizens to the underpinning model?

Elastic Diffusion

In this talk we will first understand this from a hypothetical perspective. Is it possible to move the computation and decision making the model creation, etc, closer to where the data is actually being created? In other words, to take proximity, context or capability and energy more into account. That is definitely some an important area of research that people are working on. In this talk, we will focus on what I call the main principle, which is elastic diffusion. We will break it down into essentially two points. The first one is elasticity. Elasticity is a property that we know from physics, and it is basically a property, which says something about the resilience. We know elasticity from physics, it's a property of returning to initial form or a state following some deformation. In other words, you put force on a material, it changes its shape. When you take away that force, it goes back to its initial form. That is the principle of elasticity. In, in science in neuroscience for the brain, they call it plasticity. So you learn something and something changes its shape, so to speak. The second principle we discuss is Osmotic Computing. Osmosis is a principle from Chemistry. Molecules flow from higher to lower concentration. Similarly, we aim at mimicking the flow of microservices (functionality) from Cloud, Fog, and Edge devices from and to each other. In this talk we will discuss what Elasticity and Osmosis mean for the domain of Edge Intelligence and what the fundamental research question entail.

3.4 AWS Wavelength and Verizon 5G Edge

Janick Edinger (Universität Hamburg, DE)

Modern applications generate complex tasks that must be executed promptly and reliably. Edge computing in combination with 5G networks offer computing capabilities that can be accessed with ultralow latencies. However, unlike cloud computing, edge computing introduces a new type of infrastructure that needs to be installed, configured, and maintained. AWS Wavelength and Verizon 5G Edge provide a standard cloud infrastructure at the speed of the edge. Applications with strict latency and high bandwidth requirements benefit from this deployment, among them AR/VR rendering, 360-degree video streaming, real-time monitoring of connected cars, as well as the detection of quality issues on fast-moving assembly lines in smart factories.

3.5 Future Cloud Computing View – A Perspective from LRZ

Dieter Kranzlmüller (LMU München, DE)

The Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities is the IT service provider for science in Munich, Bavaria, Germany and Europe and thus a partner in many scientific projects. Recently, more and more demand is raised from Edge devices, which work in combination with LRZ´s own cloud infrastructure. Several examples demonstrate the workflow from the senders, through edge and network into the

cloud. The crucial question is where to place the respective functionality , starting from processing and storing. The answer on this question depends on the capabilities of components along the edge-cloud continuum. Clouds serve as a means of accessoring largescale (federated) and "always-on" resources. In summary, clouds will continue to grow in terms of capacities and capabilities, limited only by availability of funding and power provisioning, which offering more and more heterogenous resources, from special AI devices to future QC accelerators. In any case, applications will benefit from using the entire spectrum corresponding to their specific needs and characteristics.

3.6 Performance and Security in Edge Video Analytics

Ling Liu (Georgia Institute of Technology – Atlanta, US)

The rapid growth of wireless mobile broadband communication networks has fueled new capabilities in scalable device-to-edge-to-cloud continuum, ranging from increased data rates, ultra-low latencies, larger coverage with massive number of devices connected 24x7. These advances have enabled new edge assisted applications, such as Augmented Reality/Virtual Reality (AR/VR) and video analytics. In this invited talk for Edge AI at Dagstuhl Seminar 21342, I will describe research challenges for performance and security in edge video analytics with dual goals. First, I will advocate high degree of resilience against systemic and adversarial disruptions for scalable video analytics on heterogeneous edge devices. Second, I will advocate combining multiple innovative techniques synergistically to provide the end-to-end resilience for next generation intelligent systems.

3.7 Towards Ubiquitous Intelligence in 6G

Sasu Tarkoma (University of Helsinki, FI)

Main reference Xiang Su, Xiaoli Liu, Naser Hossein Motlagh, Jacky Cao, Peifeng Su, Petri Pellikka, Yongchun Liu, Tuukka Petäjä, Markku Kulmala, Pan Hui, Sasu Tarkoma: "Intelligent and Scalable Air Quality Monitoring With 5G Edge", IEEE Internet Comput., Vol. 25(2), pp. 35–44, 2021.
URL http://dx.doi.org/10.1109/MIC.2021.3059189

The presentation focused on the motivation and development of ubiquitous Intelligence for beyond 5G systems towards 6G. Ubiquitous Intelligence pertains to the fusion and integration of sensing, AI, and connectivity in a hyper-local context. This emerging paradigm is expected to support many current and new vertical application areas, such as holographic interaction, tactile Internet, and massive-scale intelligent city services. Ubiquitous Intelligence requires the real-time discovery and interconnection of sensing components, context gathering, and storage with the algorithmic elements such as context estimation and prediction, positioning, and general AI algorithms. Due to the diversity of use cases, ubiquitous Intelligence requires asynchronous data-driven communications, serverless and function-based operation, and separation of concerns between the service types, such as communications, sensing, and AI.

The key aim of the paradigm is to support networks and applications that operate in the ubiquitous computing environment. Ubiquitous Intelligence aims to support a cognitive network architecture capable of accommodating the requirements of the verticals. In addition, it seeks to facilitate the design and deployment of vertical applications that can utilize the resources of the programmable world.

We envisage that the ubiquitous Intelligence environment is hierarchical in terms of geography and capabilities. The end-to-end environment consists of endpoints with opportunistic interactions through fog computing and various edge and core cloud processing tiers. In typical interactions, mobile devices and industrial devices utilize nearby capabilities for processing first and then send content for processing at more distant cloud levels. Learning and inference are distributed with partial models being generated and aggregated in the end-to-end environment.

3.8 Energy-efficient Energy-efficiency Calculations using Edge Intelligence

Michael Welzl (University of Oslo, NO)

Machine Learning (ML) methods can learn traffic patterns to make better decisions for energy efficiency in communications, e.g. for wireless devices. However, ML itself is quite energy-hungry. We can solve this dilemma by assigning the ML task to an edge device that is powered by renewable energy, e.g. via a solar panel.

4 Panel discussions

4.1 Breakout Session: Future Cloud View

Tobias Meuser (TU Darmstadt, DE)

Joint work of Based on the discussions between the remote attendees and the two panelists: Dieter Kranzlmüller and Schahram Dustdar

In the session on the future cloud perspective, we discussed technical and non-technical aspects of the interaction between edge and cloud. From the non-technical view, cloud and edge should aim to collaborate to improve the overall service quality. From the technical view, the resource constraints, energy efficiency, trust, and privacy have been discussed and considered as future research challenges.

4.2 Breakout Session: Beyond 5G View

Nitinder Mohan (TU München, DE)

Joint work of Based on the discussions between the remote attendees and the two panelists: Jari Arkko and Sasu Tarkoma

"Edge computing" is still a diffused term as the definition and possibilities of the "edge" are varied – which makes it more widespread and opens room for innovation. While academics have considered utilizing and deploying edge in devices (and servers) in home and office environments, the industry has proposed engraving the edge compute capabilities within their managed network infrastructure, e.g. cellular backbones, on-premise devices (e.g., home gateways), industry building (servers), enterprise, or smart city infrastructure (e.g., roadside units). Despite the "exact" availability, the benefits of the edge, and its synergy with supporting complex AI-based applications, are only possible if the cellular/networking fabric seamlessly connects users/sensors to such servers. Take, for example, the case of autonomous vehicles within a smart city environment. The most optimal operation of the vehicle is only possible if the AI models in the vehicle and the smart city are in sync with each other, i.e. the city learns of the driving destination and requirements from the vehicle to dynamically adjust its smart traffic control. Simultaneously, the vehicle feeds in data about congestion and road conditions to tweak its speed, lane, etc. Such an interaction is possible if the network interconnecting the two is itself self-learning and is able to adapt to dynamic environmental changes.

In the seminar, the participants discussed not just how to achieve such a vision within the up-and-coming 5G connectivity but also how to go (above and) beyond it. With future cellular access technologies, three pertinent opportunities can be explored. Firstly, the improved communication speeds along with reliable wireless communication will only enable novel IoT sensors and actuators that can help improve the granularity of data and control in AI applications. Secondly, in future cellular standards, there can now be a possibility to integrate AI within inherent control decisions (using smart switch and router hardware) that transparently integrates data generated in different ingress ports to improve not only the Quality-of-Experience of AI applications but also the Quality-of-Service of the network itself. Finally, there is a possibility to tailor computations in the network to be "human-driven" which integrates specific characteristics of humans (e.g. mobility patterns) into training its ML models. This way, for future cellular technologies, AI and edge computing will not just enable "human-in-the-loop" applications but also "human-centric" applications. Here, the participants agreed that a unified edge-in-the-network model should be worked out in the near future to account for the high cost of operation and management for supporting such an extensive and pervasive network of compute servers integrated deeply within the network fabric.

4.3 Breakout Session: AI/ML View on Edge Intelligence

Gürkan Solmaz (NEC Laboratories Europe – Heidelberg, DE)

Joint work of Based on the discussions between the remote attendees and the two panelists: Prof. Yiran Chen and Prof. Ling Liu

The remote breakout session on artificial intelligence (AI) and machine learning (ML) view focused on two main topics: 1) Edge infrastructure for distributed intelligence, 2) Distributed and federated learning models.

Currently, AI accelerators have the problem of applying AI designed for specialized hardware on another hardware. For instance, training an AI model on hardware can be fast, whereas training the same AI model would take longer on other machines. To satisfy requirements of distributed intelligence, it would be beneficial to have programmable devices instead of "black-box" devices and re-use existing techniques developed in edge and cloud computing. In recent years, GPUs have become larger and more expensive, whereas many small and embedded devices have become available. The new requirements for edge infrastructures may create new opportunities for industry stakeholders such as internet service providers, cloud providers, and local providers of hardware/software.

Distributed and federated learning has been of interest to academic and industrial research, especially for motivating edge computing scenarios such as autonomous driving. For distributed/federated learning, it is not straightforward to run advanced AI on embedded devices. A solution to this might be training lower granularity AI models on embedded devices and leveraging edge AI outcomes as features for more advanced models. Furthermore, researchers need to explore the trade-offs between accuracy vs. fairness and generalization vs. performance of AI/ML for edge intelligence. Moreover, it is not clear how different AI models can be adapted, re-used, and integrated. There has been ongoing research on multi-task learning and multi-modal data sources; on the other hand, there is a need for a unified AI library that implements joint loss functions for different AI models.

Participants

- Christian Becker
Universität Mannheim, DE
- Schahram Dustdar
TU Wien, AT
- Janick Edinger
Universität Hamburg, DE
- Lauri Lovén
University of Oulu, FI
- Tri Nguyen
University of Oulu, FI
- Ella Peltonen
University of Oulu, FI
- Jan Rellermeyer
TU Delft, NL
- Martijn Warnier
TU Delft, NL
- Lars Wolf
TU Braunschweig, DE

Remote Participants

- Atakan Aral
Universität Wien, AT
- Jari Arkko
Ericsson – Jorvas, FI
- Yiran Chen
Duke University – Durham, US
- Eyal De Lara
University of Toronto, CA
- Aaron Ding
TU Delft, NL
- Fred Douglis
Peraton Labs – Basking Bridge, US
- Diego Ferran
Telefónica Research – Barcelona, ES
- Thomas Hiessl
Siemens AG – Wien, AT
- Dewant Katare
TU Delft, NL
- Dieter Kranzlmüller
LMU München, DE
- Ling Liu
Georgia Institute of Technology – Atlanta, US
- Madhusanka Liyanage
University College Dublin, IE
- Ivan Lujic
TU Wien, AT
- Setareh Maghsudi
Universität Tübingen, DE
- Nitinder Mohan
TU München, DE
- Iqbal Mohomed
Samsung AI Research – Toronto, CA
- Roberto Morabito
Ericsson – Jorvas, FI
- Petteri Nurmi
University of Helsinki, FI
- Jörg Ott
TU München, DE
- Francesco Regazzoni
University of Amsterdam, NL & Università della Svizzera italiana – Lugano, CH
- Olga Saukh
TU Graz, AT
- Stefan Schulte
TU Hamburg, DE
- Henning Schulzrinne
Columbia University – New York, US
- Maarten Sierhuis
Nissan Research Center – Sunnyvale, US
- Stephan Sigg
Aalto University, FI
- Pieter Simoens
Ghent University, BE
- Gürkan Solmaz
NEC Laboratories Europe – Heidelberg, DE
- Sasu Tarkoma
University of Helsinki, FI

- Wiebke Toussaint
TU Delft, NL
- Antero Vainio
University of Helsinki, FI
- Marten Van Dijk
CWI – Amsterdam, NL
- Maarten van Steen
University of Twente, NL
- Blesson Varghese
Queen's University of Belfast, GB
- Shiqiang Wang
IBM TJ Watson Research Center – Yorktown Heights, US
- Klaus Wehrle
RWTH Aachen, DE
- Michael Welzl
University of Oslo, NO
- Chenren Xu
Peking University, CN