EdgeAI: A Vision for Distributed, Edge-native Artificial Intelligence in Future 6G Networks

Lauri Lovén, Teemu Leppänen, Ella Peltonen, Juha Partala, Erkki Harjula,
Pawani Porambage, Mika Ylianttila, Jukka Riekki
first.last at oulu.fi, University of Oulu, Finland

Edge computing, a key part of the upcoming **5G mobile networks** and future **6G technologies**, promises to distribute cloud applications while providing more bandwidth and reducing latencies [1]. 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. In stark contrast, current **artificial intelligence** (AI) and in particular **machine learning** (ML) methods assume computations are conducted in a homogeneous cloud with ample computational and data storage resources available. Currently, AI's cloud-centric architectural model requires transmitting data from the end-user devices to the cloud, consuming significant data transmission resources and introducing latencies.

Previous studies address AI in different perspectives of IoT, edge computing and networks [2], [3], [4], [5]. However, we provide a holistic view of AI methods and capabilities in the context of edge computing. In our vision, a holistic view of AI methods for edge computing comprises the well-known paradigms, such as 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, presents a unique environment for deploying such applications based on computations units with different degrees of intelligence capabilities.

Joint consideration of edge computing and AI methods, *EdgeAI*, improves both fields in a variety of aspects. We aim to **identify the challenges and detail the potential benefits of AI at the edge**, building a coherent and overarching vision of what distributed artificial intelligence means in the context of edge computing. Further, we aim to **find the methods realizing those benefits**, testing hypotheses in a real-world setting on the edge platform atop the 5G test network (http://5gtn.fi). Our vision will be realized within the 8-year span of the Academy of Finland 6Genesis Flagship.

I. Vision

Bringing edge computing and AI together is challenging due to the fundamental difference in the premises of AI and edge computing. Whereas edge computing by design distributes computational units across system architecture layers and decentralizes computations vertically and horizontally, modern AI methods are only beginning to allow for distributed, let alone decentralized, computations.

Yet, clear benefits can be identified from the interplay of AI and edge computing. Following Park et al. [3], we divide this interplay into *edge computing for AI* (Edge4AI) and *AI for edge computing* (AI4Edge). Fig.1 illustrates the expected benefits for distributed AI functionality, further dividing edge computing to communication, platform control, security, privacy, and application or service specific aspects.

Indeed, EdgeAI may improve communication networks in many ways. Scalability and platform KPIs can be improved, especially in relation to connectivity, data transmission, computation offloading, and capabilities for reactivity and proactivity. This contributes towards robustness, reliability and scalability, resulting in improved Quality of Experience (QoE) for edge applications.

Further, in contrast to the current vertical application silos, we emphasise horizontal connectivity and interoperability between contributing devices. Such devices, e.g. user devices and network infrastructure components, share their communication and computation resources within and across layers.

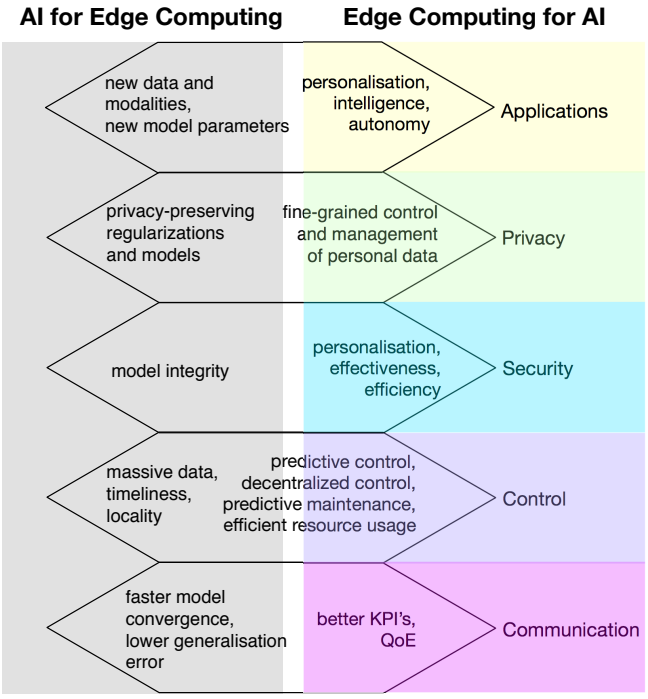

Fig. 1. Benefits provided by EdgeAI for both AI and edge computing.

AI agents can address network issues such as low throughput and intermittent connectivity by sharing connections between devices locally. Such operation requires autonomous, context-aware operation of the participating components, requiring AI capabilities both on the user devices and wireless routers. Conversely, AI models converge faster with a lower generalization error when the architectures of the underlying communication network are built to support AI workflows.

Management of massive-scale IoT systems is difficult due to device mobility and communications latencies across system architecture layers. Edge computing provides a platform for decentralized IoT system control. An edge platform lies close to the managed devices, allowing real-time context-awareness for intelligent platform management. The main challenge is the tradeoff between communication and computation resource usage for simultaneous applications, each with their own requirements and dynamically available resources.

AI methods on the edge can base their ML models on the massive amounts of timely system control and application execution data available. AI can provide new capabilities, such as predictive algorithms for resource sharing and decentralized dynamic system orchestration and control. However, as user devices have a degree of autonomy, edge components need understanding of application requirements, available resources in the connected devices, and the application-specific contexts. Such knowledge requires partitioning of application execution according to the opportunistic resource availability. However, intermittent connectivity may corrupt or slow down intelligent operation such as the distributed training of ML models [3]. Further, due to application requirements and contextual understanding in different scales, based on massive amounts of control and application data, the complexity of the models is to increase significantly.

AI improves edge security with personalized, shareable, location-aware security systems adapted to each individual user context with fine-grained control. Further, intrusion prevention stands to gain from ML-based pattern and anomaly recognition in the network and application operation. ML models, on the other hand, are ensured of data integrity. Further, while AI methods may improve user privacy with personal privacy guards, fine-grained control on consent management and data ownership, or decentralizing trust with for example distributed ledgers, ML models benefit from novel privacy-preserving regularisations and data-generating unsupervised or semi-supervised models.

Finally, on one hand, the edge-native AI methods provide applications with unprecedented access to personalized and localized prediction and control models, enhancing application intelligence and autonomy. On the other hand, applications themselves provide AI models with new data and modalities as well as new model parameters. General improvements can be seen in optimizing the edge applications execution in opportunistic environment while improving connectivity, reactivity, adaptivity and proactivity.

This vision defines the building blocks for targeting **edge-native AI integration in the future 6G networks**. The biggest paradigm shift lies in the growing role of intelligence, autonomy, context-awareness and collaboration in the operation of edge applications, which rely on the user devices and edge infrastructure components. The edge-native AI paradigm relies on the key characteristics of edge computing, i.e. reactivity, scalability, distribution and sharing of resources, and optimization of QoE. Here, our edge-focused view combines both *AI for Edge computing* and *Edge Computing for AI*, leveraging the shared responsibilities of big data analytics and system control to the edge.

To summarize, edge-native AI provides edge applications with unprecedented access to secure, personalized, context-aware, localized and distributed application-specific AI methods, targeting specified topics across the aspects in Fig.1 from multiple perspectives, i.e. platform providers, 3rd party application providers, network operators, and end users. EdgeAI solutions will enable novel applications and radical innovations in the fields of urban computing, smart buildings and cities, personalized applications and context-aware mobile technologies, and so forth. In the future 6G era, edge-native artificial intelligence will be a crucial part of everyday computation and smart technologies.

Acknowledgements: This research is supported by the Academy of Finland 6Genesis Flagship (grant 318927) and by the Future Makers program of Jane and Aatos Erkko Foundation and Technology Industries of Finland Centennial Foundation.

REFERENCES

[1] C. Li, Y. Xue, J. Wang, W. Zhang, and T. Li, "Edge-oriented computing paradigms: A survey on architecture design and system management," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 39, 2018.

[2] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in internet of things: a survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 1–27, 2018.

[3] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless Network Intelligence at the Edge," *arXiv preprint*, 2018.

[4] Q. Mao, F. Hu, and Q. Hao, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2595–2621, 2018.

[5] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.