Ericsson white paper
GFMC-18:000260
June 2018

ERICSSON

Artificial intelligence and machine learning in next-generation systems

5G, the next generation of mobile communication, will play a similar role in the evolution of digitalizing industries as cloud technologies have for the web industry. The ability to automate and leverage on data from distributed systems with real-time capabilities will be critical. Based on insights about future 5G systems and developments in manufacturing and ITS automation, this white paper reflects on the technical challenges the R&D community needs to address in order for ICT providers and other industry players to be able to fully capitalize on the potential of artificial intelligence (AI) and machine learning.

Introduction

The combination of billions of connected devices, petaflops of computing resources and advanced communication capabilities that enable real-time interactions is leading to the creation of systems on a scale and complexity level that is beyond the ability of humans to fully comprehend and control. Management and operation of these systems require an extremely high degree of intelligent automation.

An upside of the growing scale and multitude of interactions that leads to increasing system complexity is the fact that those same characteristics make it possible to gather vast amounts of data from different parts of the system. The gathered data can be fed into models and methods to derive deep insights that can be used to optimize and tailor the behavior of the system.

At Ericsson we are convinced that machine intelligence, which encompasses the fields of artificial intelligence (AI) and machine learning, offers the best opportunity to achieve the high levels of automation necessary to manage the complexity and optimize system performance. However, to fully capitalize on the potential of machine intelligence, we must first overcome the six technical challenges outlined in this white paper.

Machine intelligence for industries

While machine intelligence is certain to play a key role in the creation of next-generation systems in a wide variety of industry sectors in the near future, it is particularly relevant in rapidly developing industries such as ICT, manufacturing and transportation.

The ICT industry and 5G

Across the globe, mobile operators are getting ready to deploy the 5th generation of 3GPP mobile wireless networks (5G). Compared to the mobile infrastructure that is currently in place, 5G will bring higher throughput, lower latency, more efficient signaling, support for more spectrum bands, more programmability and other additional advanced techniques to maximize usage and optimize costs [1]. The number of connected devices will greatly increase as a result of this improved performance: sensors will benefit from more affordable bandwidth to the internet; heavy users of uplink traffic like video cameras will be able to share more data; fast-moving devices (drones, cars) will have more reliable connectivity and so on. These new devices will be the catalysts of a new wave of innovation for all involved industries.

The proliferation of these new devices will also increase the difficulty of manually managing a telecom network. Traditionally, a mobile network is designed and managed by telecom experts who rely heavily on their extensive knowledge of the network topology, the subscribers' mobility and usage patterns and the radio propagation models to design and configure the policies that continuously orchestrate the network. With 5G, these topologies will grow more complex with small cells and new antennas, usage patterns will become less predictable for humans alone, the radio propagation models will become harder to compute with new radio spectrum bands and denser topologies. As a result, machine intelligence will play a key role in assisting operators in laying out and operating 5G networks. More and more policies will be machine-learned, leveraging on constant radio measurements from the field and best-in-class simulators.

A mobile network is a highly distributed and decentralized system. As shown in **Figure 1**, machine intelligence capabilities will be added to several layers in the 5G architecture to enable data processing for various purposes, both locally (close to where data is created) and centrally (where data can be consolidated). Local learning and decision making will be done in distributed sites while data and knowledge across sites will be blended for a comprehensive global understanding of networks, services and functions.

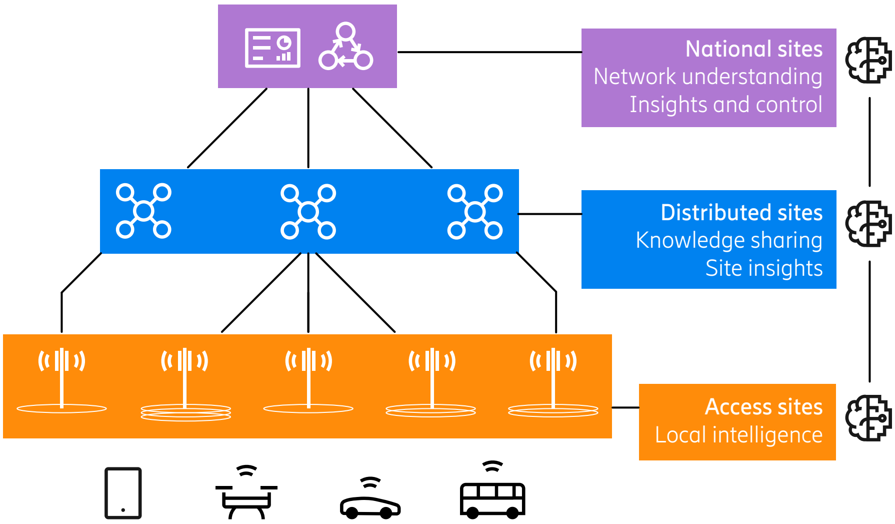

Figure 1: Local and global learning and decision making in large distributed networks

Each local site is a rich source of data about the state of the different components, the time series of events and associated contextual information. This can be used to build models for local behavior, whereas reasoning is required for the knowledge gathered across sites in order to infer system-wide insights. Ideally, the knowledge and insight derived at one site can be used at the other sites for better prediction.

More intelligence built in to the 5G system will allow for a shift from managing networks to managing services. Intelligent functions can be customized for each of these services allowing them to operate more resiliently and securely, taking the mobile network to a new level of innovation for the benefit of industry and society [2].

Intelligent manufacturing

Triggered by digitalization, the manufacturing industry has started its transformation toward a new era known as Industry 4.0. Data from connected production units and items in production opens up for a level of flexibility and efficiency improvements that were not possible before. Initially, anomaly detection and root-cause analysis will remove bottlenecks and increase the production yield. Various prediction algorithms will contribute to increase the level of automation. The flexibility in operations and level of produced quality will make it possible to reduce claims and returned products.

Smart manufacturing will further evolve with real-time capabilities due to the introduction of cloud and edge technologies and the increased level of connectivity that 5G can provide on the factory floor. For example, data that used to be contained in an isolated programmable logic controller (PLC) unit within each production unit (robot), can be shared with other robots through their connected PLC units. Eventually one can imagine the PLC being containerized as a microservice in a cloud infrastructure, enabling production optimization that takes the complete production line into account. Further, the manufacturing executing system that controls the overall production process can be further advanced by increased interaction with data down to PLC level.

While machine intelligence promises disruption opportunities in the manufacturing industry, it also poses many challenges: data is often heterogeneous and distributed, and there are hard requirements on real-time autonomous decision making. Further, the safety-criticality of many use cases in the industry, most of which arise during man-machine collaboration, sets stringent bounds on the requirements and adds to the complexity.

Intelligent Transportation Systems

Advances in Intelligent Transportation Systems (ITS) are leading to the introduction of more and more vehicles with autonomous driving capabilities. However, intelligent automation in ITS is not limited to autonomous vehicles alone. There are efforts underway to increase the efficiency of traffic systems at a strategic level such as the design of roads and artifacts (signal lights, traffic islands, bus stops, car parks and so on), the control of traffic signals and the configuration of directions based on mobility pattern predictions. All these applications require the processing of vast amounts of data to extract the required knowledge and make global decisions. Tighter control loops at the tactical level include operating and coordinating traffic lights for maximal throughput, and handling traffic congestion due to unforeseen events such as accidents. In many cases, even though the events seem to require only local intervention, without a global perspective, a local action can lead to gridlock in a larger zone.

ITS use cases can be categorized into layers, each having different requirements on the communication QoS. This imposes different constraints on the underlying models that have to be updated in real-time and deliver outputs with required latencies. Unlike in the manufacturing industry, the most prominent characteristic of ITS is the mobility of vehicles and people. This forces the applications and hence the machine intelligence capabilities to be mobile and the underlying communication architecture to be perpetually reconfigured to deliver the required QoS.

It is well understood that machine intelligence is applicable to almost all operations across many industries and can achieve efficiency through intelligent and adaptive automation. The applications presented above illustrate several facets of machine intelligence in large industries.

Addressing the challenges

There are six main challenges that must be addressed to ensure the acceptance of machine intelligence as a viable approach for the intelligent automation of complex systems.
At Ericsson, we believe that solution approaches will require strong domain knowledge and a deep understanding of the underlying connectivity and communication aspects.

1. Real-time intelligence

More and more networked devices and systems are able to sense their surroundings, rapidly generating data and insights about the world and their own operations. Autonomous vehicles, manufacturing assembly lines, smartphones, IoT devices and sensors, and 5G networks themselves are all examples of such systems. The need for real-time intelligent decision making on live data is rapidly increasing and has been identified by major industry and academic players that are collaborating to co-create the required software components in initiatives such as the Berkeley RISELab [3].

Real-time requirements entail that predictions, model updates and inferences from knowledge bases are based on live-streaming data. This has implications for system architecture – how to distribute the models and knowledge bases over the cloud, edge and devices; whether model training should be offline or online; how to represent and prepare data for fast consumption by algorithms, and more. The systems must support flexible, programmable data pipelines for the volume, velocity and variety of real-time data and algorithms capable of real-time decision making.

In the past many capable frameworks have been developed to deal with data at rest, making sense of large heterogeneous data lakes. We believe there is an urgent need for similar advances for data in rapid motion, where latency requirements will approach the limits of modern communication networks. There will soon be a need to guarantee real-time intelligence with extremely small latencies; one or two orders of magnitude faster than we typically experience with a centralized cloud. Such advancement will be a cornerstone in the next wave of intelligent networked systems.

2. Distributed and decentralized intelligence

In a large distributed system, decisions are made at different places and different granular levels. Some decisions are based on local data and governed by tight control loops with low latency. Other decisions are more strategic, affect the system globally and are made based on data collected from many different sources. Decisions made at a higher global level may also need real-time response in critical cases such as power-grid failures, cascading node failures, and so on. An intelligent system that automates such large and complex systems must necessarily reflect the distributed nature and support the management topology.

Data generated at the edge, in a device or network edge node, will at times need to be processed in place. Data may not economically be transferred to a centralized cloud; there may be laws governing where data can reside and there could be privacy or security implications of data transfer. The scale of decisions in these cases is restricted to a small domain, so the algorithms and computing power necessary are usually fast and light. However, local models could be based on incomplete and biased statistics which may lead to loss of performance. There is a need to leverage the scale of distribution, make appropriate abstractions of local models and transfer the insight to other local models.

Learning about global data patterns from multiple networked devices or nodes without access to the actual data is also possible. A recent approach is so-called federated learning: learn local models based on local data patterns, send the local models to centralized cloud, average them and send back the average model to all devices. More research is needed along this direction, in particular to cater for different kind of models and model combinations and stronger privacy guarantees.

Global decision making, on the other hand, relies on knowledge of the global state and global data patterns – that is, patterns of events occurring across local nodes. The global state is built by combining the knowledge from multiple components and using it to make decisions. However, we must ensure that all the pieces of information were correct at certain points in time not so long ago. Such correctness can be ensured by gathering a distributed snapshot over the system components and performing reasoning on the snapshot instead of the raw data directly. For such a snapshot to be consistent, the variety of data coming from different sources must have equivalent semantics. Moreover, the snapshot must remain current at the time when it is used for reasoning.

A common distributed and decentralized paradigm is required to make the best use of local and global data and models, and determine how to distribute learning and reasoning across nodes to fulfill extreme latency requirements. Such paradigms themselves may be built using machine intelligence to incorporate features of self-management, self-optimization and self-evolution.

3. Beyond games and simulations

Reinforcement learning is one of the key techniques used in the development of the AlphaGo Zero system [4], which famously achieved superhuman performance in the game Go in 2017. A reinforcement learning agent learns by taking actions in an environment with the goal of maximizing some notion of cumulative reward. It both exploits its current knowledge and explores the environment to gain new knowledge. It tries different actions in various system states in order to learn from the outcomes. For industrial systems, including telecom networks, however, it is not always possible to explore all actions when the system is operational because an undesirable action/state combination could lead to degraded performance. More research is required to figure out how to fully utilize the strengths of reinforcement learning and other state-of-the-art algorithms, and tailor them to industrial settings and constraints.

One potential avenue is to develop algorithms that can explore the environment in a controlled manner – for example, by identifying or learning state regions or temporal slices where all or some actions are permitted and confining exploration to them. Another direction is to start with a realistic simulator where the agent can explore without limits and then have a principled way of transferring this knowledge to an operational system. The operational system can potentially also give feedback to the simulator so that it in turn can learn to simulate even more realistically. As an example, a newly deployed radio base station cannot always be allowed to freely explore its environment, but could instead rely on its neighbors or simulations for a sufficiently accurate base model.

4. Machine learning meets machine reasoning

Machine learning and machine reasoning can both be used to build intelligent logic but they have different approaches. Machine learning is typically used for learning a complex function from vast amounts of data – for example, learning to classify images using supervised learning or learning to master the game of Go by reinforcement learning. Machine reasoning, on the other hand, implements abstract thinking as a computational system. Such a system contains a knowledge base storing declarative and procedural knowledge and a reasoning engine employing logical techniques such as deduction and induction to generate conclusions.

Declarative knowledge describes the facts of the domain, concepts and their relationships. One can reason using the declarative knowledge to understand complex relationships between the concepts of a certain domain. Procedural knowledge is the knowledge of "how to do/perform a task". It can be used to describe, for example, the dynamic behavior of the components in the domain – that is, the triggering events and actions that the components can take.

One of the main challenges in intelligent systems is effective integration of learning and reasoning. Statistical learning and symbolic reasoning have been developed largely by distinct research communities. We see that hybrid approaches will be useful in next-generation intelligent systems where robust learning of complex models is combined with symbolic logic that provides knowledge representation, reasoning and explanation facilities. This is particularly useful where knowledge can be codified as given knowledge either to guide, monitor or predict system behavior. The knowledge could be, for example, universal physical laws or the best-known methods in a domain.

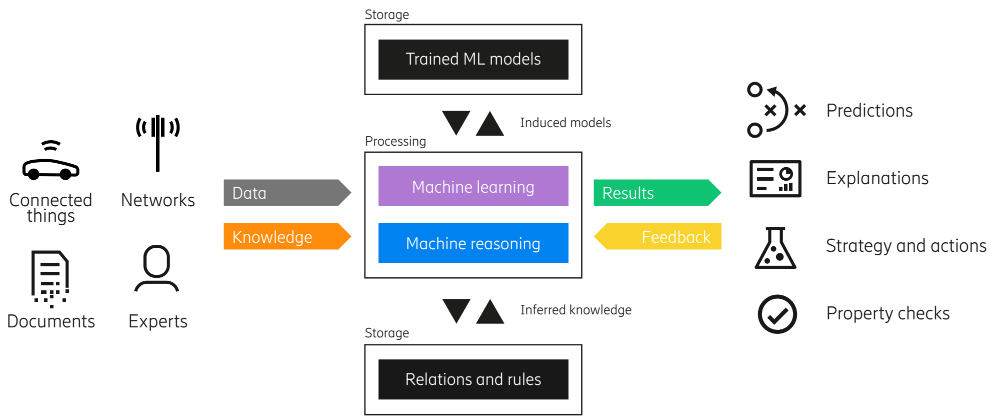

Figure 2: Interplay of machine learning and reasoning procedures in intelligent systems

There are several ways machine learning can be related to machine reasoning. For example, as shown in **Figure 2**, by using knowledge representation and reasoning, high-level concepts may be extracted from complex neural networks for the purpose of knowledge comprehension, validation and maintenance. Another area may include employing machine learning techniques to reinforce reasoning by learning from successfully explored strategies and guiding the state exploration mechanisms by adaptive heuristics. For example, in large manufacturing units, the system can predict a failure and suggest the best configurations to choose from so that the machines can be led to safe states in a cost-optimal way.

Intelligent systems must expose more deliberate behavior in order to fulfill given objectives, robustness to be able to solve a problem in several different ways, and flexibility in decision making by utilizing various pieces of both prepopulated and learned knowledge.

5. Human-machine collaboration

As machine intelligence capabilities proliferate into more and more systems and augment human decision-making processes, it is imperative that we continue to deepen our understanding of how human intelligence and machine intelligence complement each other. Humans play a critical role in defining requirements for the systems and enhancing the quality of the models through supervision, reinforcement and in debugging the algorithms. At the same time, there is an opportunity to increase the efficiency of human-driven operations tremendously through augmenting human intelligence.

In recent years there have been major advances in natural language technologies and computer vision, leading to interfaces where humans can interact more meaningfully with machines using speech, images and video. The deeper intelligence of such systems lies in augmenting human intelligence with the most relevant information for the current task, and the best suggestions or decisions.

There is an urgent need for solution frameworks that can support the symbiotic interaction of humans and machines. These frameworks must capture a variety of functional, performance and compliance-related requirements to be satisfied by the intelligent systems (the collaborative systems of humans and machines). They should also have tools and methods to ensure that despite their self-optimizing and self-evolving nature, the intelligent systems always satisfy the requirements.

6. Safety and trust

One of the risks inherent in the human-machine interaction is the potential for human users to misuse machine intelligence systems. Misuse can have far-reaching consequences in terms of compromising safety and profitability.

Central to addressing this problem is the concept of calibrated trust, which means that the human should trust a machine to perform its tasks in accordance with the capability and "willingness" of the machine to perform them. For instance, a human who over-trusts an automated system tends to misuse it. An example of this is a human driver who uses an autonomous vehicle with level 3 automation as if it had level 5, which would reduce the overall level of safety. Another example is when a human user chooses not to use an automated system out of fear or lack of trust, which would negatively impact the effectiveness of the process. In order to reach true collaborative scenarios between humans and machines, calibrated trust is essential.

The more autonomy we give to systems the more important it is to be able to provide safety and trust guarantees, especially in safety-critical use cases such as ITS or manufacturing processes that include human workers. New verification methods are needed that take into account the adaptive, evolving nature of machine intelligence, as well as monitoring solutions that not only detect and resolve but also predict and prevent. There is a growing body of research on the topic of effectively communicating the rationale behind the system's insights and decisions to the human operators, which is welcomed.

Conclusion

Based on our insights about future 5G systems and ongoing developments in manufacturing and ITS automation, this white paper has highlighted six technical challenges that we believe must be addressed in order to be able to fully capitalize on the potential of artificial intelligence technologies. The technical challenges involve: (1) real-time intelligence, (2) distributed and decentralized intelligence, (3) moving beyond games and simulations to industrial scale, (4) combining machine learning with reasoning, (5) human-machine collaboration, and (6) safety and trust. We believe these challenges are best addressed through approaches based on strong domain knowledge, including a deep understanding of the underlying connectivity/communication aspects. Ericsson is working to address all of them and we urge the larger research and development community to join us in finding solutions.

Glossary

AI	Artificial Intelligence
ITS	Intelligent Transport Systems
ML	Machine Learning
PLC	Programmable Logic Controller

References

[1] 5G systems – enabling the transformation of industry and society, Ericsson white paper, January 2017, available at: https://www.ericsson.com/assets/local/publications/white-papers/wp-5g-systems.pdf

[2] The 5G business potential – Industry digitalization and the untapped opportunities for operators, Ericsson report, October 2017, available at: https://www.ericsson.com/assets/local/narratives/networks/documents/report-bnew-17001714-uen-rev-a.pdf

[3] RISELab, available at: https://rise.cs.berkeley.edu

[4] Silver, David, et al. "Mastering the game of go without human knowledge," Nature 550.7676 (2017): 354.

Further reading

1. Ericsson white paper, Big data analytics, 2015, available at: https://www.ericsson.com/assets/local/publications/white-papers/wp-big-data.pdf

2. Ericsson Research Blog, available at: https://www.ericsson.com/research-blog/tag/machine-intelligence/

3 Tackling IoT complexity with machine intelligence, Ericsson Technology Review, 2017, available at: https://www.ericsson.com/en/ericsson-technology-review/archive/2017/tackling-iot-complexity-with-machine-intelligence

Contributors

The contributors to Ericsson's opinion on this topic are Elena Fersman, Julien Forgeat, Rickard Cöster, Swarup Kumar Mohalik and Viktor Berggren.

Elena Fersman
Elena Fersman is a research director in machine intelligence and automation at Ericsson Research. She is also a docent and an adjunct professor in cyber-physical systems specialized in automation at KTH Royal Institute of Technology in Stockholm. She holds a Ph.D. in computer science from Uppsala University, Sweden, and a master's degree in management from St. Petersburg Polytechnic University in Russia. She did her postdoctoral work at École Normale Supérieure Paris-Saclay in France. At Ericsson, she has held various positions ranging from product management to research leadership. Her current research interests include modeling and analysis of autonomous knowledge-intensive cyber-physical systems. Elena has co-authored over 20 scientific publications and over 50 patent families.

Julien Forgeat
Julien Forgeat is a senior specialist in machine intelligence infrastructure at Ericsson Research. He joined Ericsson in 2010 after spending several years working on network analysis and optimization. He holds an M.Eng. in computer science from the National Institute of Applied Sciences in Lyon, France. At Ericsson, Julien has worked on mobile learning, Internet of Things and big data analytics before specializing in machine intelligence infrastructure. His current research focuses on the software components required to run AI and machine learning workloads on distributed infrastructures.

Rickard Cöster
Rickard Cöster is an expert in data science and machine learning at Ericsson Research. He has 20 years of experience in the field, and his current focus is on machine learning for self-learning networks, and novel applications in industrial IoT. Rickard works closely with academia and research institutes, and supervises a Ph.D. student in the Wallenberg AI, Autonomous Systems and Software Program (WASP). He obtained a Ph.D. in computer science in 2005 and joined Ericsson in 2010. Rickard is the co-author of more than 15 publications and 40 filed patent applications.

Swarup Kumar Mohalik
Swarup Kumar Mohalik is a principal engineer at Ericsson Research. He has been with Ericsson since 2015, working on various applications of AI planning technology in IoT and 5G. Swarup has a Ph.D. in computer science from the Institute of Mathematical Sciences in Chennai, India, and did a postdoctoral fellowship at LABRI at the University of Bordeaux in France. His other areas of expertise include theory of automata, logic and formal methods. Prior to joining Ericsson, Swarup worked at Intel, General Motors and Hewlett-Packard, focusing on formal specification and verification of real-time embedded software, and model-based testing.

Viktor Berggren
Viktor Berggren is a principal researcher in machine intelligence and automation at Ericsson Research. He holds an M.Sc. in industrial and management engineering from Linköping Institute of Technology, Sweden. His current focus is on exploring the combined technologies of artificial intelligence and machine learning. Since 2015 he has been engaged in industry collaborations, aiming to shape 5G to fit the need of future industries. Viktor has more than 30 years of experience at Ericsson in leading positions within research, product management and systems management.

Acknowledgements

The contributors would like to acknowledge the following people for their valuable input to this white paper: Andreas Johnsson, Andrii Berezovskyi, Aneta Vulgarakis, Jörgen Gustafsson, Klaus Raizer, Konstantinos Vandikas, Leonid Mokrushin, M. Saravanan, Martin Isaksson, Mattias Lidström, Meral Shirazipour, Rafia Inam and Ramamurthy Badrinath.