Pervasive Artificial Intelligence in Next Generation Wireless: The Hexa-X Project Perspective

Miltiadis C. Filippou[1], Vasiliki Lamprousi[2], Jafar Mohammadi[3], Mattia Merluzzi[4], Elif Ustundag Soykan[5], Tamas Borsos[6], Nandana Rajatheva[7], Nuwanthika Rajapaksha[7], Luc Le Magoarou[8], Pietro Piscione[9], Andras Benczur[10], Quentin Lampin[11], Guillaume Larue[11], Dani Korpi[12], Pietro Ducange[13], Alessandro Renda[13], Hamed Farhadi[14], Johan Haraldson[14], Leonardo Gomes Baltar[1], Emilio Calvanese Strinati[4], Panagiotis Demestichas[2] and Emrah Tomur[5]

[1]*Intel Next Generation & Standards, 85579 Neubiberg, Germany*
[2]*WINGS ICT Solutions, Greece*
[3]*Nokia Bell Labs, Stuttgart, Germany*
[4]*CEA-Leti, Université Grenoble Alpes, F-38000 Grenoble, France*
[5]*Ericsson Research, Turkey*
[6]*Ericsson Research, Hungary*
[7]*6G Flagship, University of Oulu, Finland*
[8]*b<>com, 35510 Cesson-Sévigné, France*
[9]*Nextworks, Pisa, Italy*
[10]*Insitute for Computer Science and Control (SZTAKI), Hungary*
[11]*Orange, France*
[12]*Nokia Bell Labs, Espoo, Finland*
[13]*Department of Information Engineering, University of Pisa*
[14]*Ericsson Research, Stockholm, Sweden*

Abstract

The European 6G flagship project Hexa-X has the objective to conduct exploratory research on the next generation of mobile networks with the intention to connect human, physical and digital worlds with a fabric of technology enablers. Within this scope, one of the main research challenges is the ambition for beyond 5G (B5G)/6G systems to support, enhance and enable real-time trustworthy control by transforming Artificial Intelligence (AI) / Machine Learning (ML) technologies into a vital and trusted tool for large-scale deployment of interconnected intelligence available to the wider society. Hence, the study and development of concepts and solutions enabling AI-driven communication and computation co-design for a B5G /6G communication system is required. This paper focuses on describing the possibilities that emerge with the application of AI & ML mechanisms (with emphasis on ML) to 6G networks, identifying the resulting challenges and proposing some potential solution approaches.

Keywords

connecting intelligence, 6G networks, sustainability, trustworthiness, ML for air interface design, edge AI, explainable AI

AI6G'22: First International Workshop on Artificial Intelligence in beyond 5G and 6G Wireless Networks, July 21, 2022, Padua, Italy.
miltiadis.filippou@intel.com (M. C. Filippou)

1. Introduction

Mobile communications systems have aggressively developed during the last 50 years, starting from analog cellular technology, then shifting to digital cellular technology in the 1990s and further evolving after 2000 towards third- (3G) and fourth-generation (4G) mobile technology devising packet switching technology and offering higher data throughput and voice quality. Until 4G, most applications were personal subscriber-oriented. In 2019, the global deployment of the fifth-generation (5G) technology for cellular networks started; this time, the focus was on applications serving vertical industries, besides further enhancing personal consumer experience.

Recently, research on B5G/ 6G technology has kicked-off worldwide influenced by the advent of technology trends, e.g., network softwarization and virtualization for flexible utilization of (virtual) network resources [1], Multi-access Edge Computing (MEC), a key technology enabling a rich computing environment at the edge of an access network [2], and the steady growth of network and application data generation with corresponding analytics opportunities. Apart from technological trends, societal needs have also triggered the need for 6G research; such needs refer to the digital inclusion of the remote and vulnerable, service trustworthiness and sustainability, expressed in terms of energy consumption, carbon footprint and cost. The European Flagship 6G project Hexa-X [3] has identified six fundamental challenges: i) connecting intelligence involving wide use of Machine Learning (ML) models across the network, ii) "network of networks", i.e., from indoor to wide area networks iii) sustainability, iv) global service coverage, v) extreme experience and vi) trustworthiness. In this paper, we address the challenges i) and vi) by concentrating on two design directions: a) "ML for air interface design" and b) "Communications for more efficient AI/ML" [4].

The rest of the paper is organized as follows: Section 2 attempts to reply to the question of why ML is needed in 6G networks. Section 3 explores ML-driven wireless transceiver design approaches, as well as ML-driven radio resource management approaches for high communication performance and resilience to radio condition changes. Section 4 elaborates on networks for high-performing, sustainable and trustworthy AI. A detailed description of technical enablers elaborated in Sections 3 and 4 appear in [4], unless otherwise referenced. Section 5 presents main challenges of applying AI/ML in 6G and, finally, the paper concludes with Section 6.

2. Connecting Intelligence - why do we need ML in 6G?

Algorithm deficiency occurs when a (network) design task is well defined, nonetheless, the complexity of obtaining a well-performing algorithm is prohibitively high. In wireless communications, problems of allocating radio resources, e.g., power, time, frequencies, antennas fall under this category. On the other hand, model deficiency characterizes network (channel) and device (hardware) conditions that cannot be summarized by a universally applicable formal model. To tackle these deficiencies, ML is foreseen to be integral part of 6G, which can be attributed to several technology enablers. Computing capacity, is becoming more and more ubiquitous, calling for smart allocation of various AI workloads depending on computing cost, device constraints, or how data intensive these workloads are. As AI/ML gained focus in several industries, the usage of dedicated hardware solutions has increased to support and accelerate both training and inference tasks. The characteristics of structures that are hard or cannot be represented by an explicit mathematical ruleset, can often be modeled by ML models by learning from large datasets available. The value of data is already well recognized, and numerous data sources are made accessible internally within systems, as well as externally for business and open use. Intelligence will be distributed across agents in self-governed sub-systems of network functions (i.e., interconnected functional building blocks within a network infrastructure) and applications which will then interact with other sub-systems. Trusted and efficient communication support among these entities is an essential enabler for scaling out AI capabilities. The recent expansion in the volume of accessible compute and data resources resulted in several major breakthroughs in the areas of AI algorithms and architectures, which brings an explosion of AI/ML services and applications.

Some representative new 6G use cases relevant to ML-driven air interface design are the ones of "merged reality game/work", "interacting and cooperative mobile robots" and "flexible manufacturing". These use cases share some "conventional" communication Key Performance Indicators (KPIs), such as bit rate, latency, communication outage probability, some new ML related KPIs, i.e., convergence speed, flexibility, data quality, and complexity gain and some environmental characteristics that need to be considered, e.g., coverage, connection density and positioning accuracy. Relevant "soft goals" (also termed after as Key Value Indicators - KVIs) are generalizability, deployment flexibility, distributed learning with frugal AI, trustworthiness and resistance to adversarial attacks. Accordingly, some 6G use cases relevant to in-network learning are the three mentioned above, plus the ones of "digital twins for manufacturing", "immersive smart city", "AI partners", and some new enabling services, i.e., AI-as-a-Service, Compute-as-a-Service (CaaS) and AI-assisted Vehicle-to-Everything. The KVIs related to in-network learning are explainability, fairness, data economy and model complexity. Finally, the KPIs related to in-network learning are latency, AI agent density, interpretability level, network energy efficiency and inferencing accuracy. For a more detailed view, please refer to [3, 4].

3. ML for better wireless communications

ML-based techniques for physical layer design, i.e., at bit-level transmission between wireless devices, show great promise for enhancing the spectral and energy efficiency of future 6G communication systems. They can introduce more flexibility, increase robustness against hardware impairments, facilitate joint optimization of transmitter and receiver, as well as reduce required signaling overhead. In the following, we investigate a few applications of ML for wireless communications.

3.1. From hardware impairment mitigation to low complexity decoder: transceiver design by ML

Joint learning of transmitter and receiver improves spectral efficiency in various scenarios [5]. The key aspect is finding the proper balance between prior knowledge/restricted tasks and the freedom to learn several tasks jointly. A particularly interesting solution involves learning only the digital signal modulation scheme (constellation) on the transmitter side, while jointly learning the full receiver, thereby discarding the need to use pilot signals [5]. Neural Networks (NNs), due to their universal function approximation properties, demonstrate resilience against hardware impairments, be learning how to compensate for the the hardware imperfections. For instance, the transmitter can be trained to produce a waveform that results in 5–10 dB less out-of-band emissions caused by a nonlinear power amplifier [5]. At the receiver side, it has also been shown that a context-aware receiver can be used to compensate the impact of hardware impairment including oscillator phase noise at highband transmissions [6].

On the channel and environment awareness side, ML has demonstrated impeccable performance in a few areas. As for channel estimation, NNs inspired by the optimal Minimum Mean Square Error (MMSE) solution [7] demonstrated an excellent trade-off between complexity and accuracy. The benefit of using NNs for the task of matrix inversion, which is the most computationally expensive part of the MMSE estimator, as shown in [7] is twofold: smaller computational complexity and lower sample complexity, in comparison to the MMSE estimator. In [8], an NN designed via the deep unfolding method is initialized with a dictionary of steering vectors, which is then trained *online* in an *unsupervised* way, yielding great performance improvements. Mitigating dynamic blockage effects which occur due to movements, is crucial when using the higher frequencies in 6G. Sensing can enable situational awareness for the wireless system and the LiDAR appears as a strong candidate for this [9]. In the context of massive multi-antenna transmission with, ML has recently been applied to predict appropriate precoders based on the location of the targeted user. An NN comprising random Fourier features in order to allow high spatial frequencies to be learned has been proposed and trained on data containing the locations of users and their channels [10].

As for forward error correction, ML applications have already demonstrated significant improvements in mechanisms such as Turbo codes, Low-Density Parity-Check (LDPC) codes or polar codes. The code-agnostic, low-complexity weighted belief propagation inspired NN decoder demonstrated in [11] and developed in Hexa-X is a promising candidate to improve the decoding performance. This approach opens new perspectives for the design of codes suitable for low-complexity and fully explainable decoders, particularly suited to the needs of constrained 6G devices.

Applying ML-based solutions to enhance the communications at the physical layer allows for lower sample complexity for estimation tasks, higher accuracy for detection tasks, and better coping with non-linearity stemming from hardware imperfections. It becomes apparent that the future of air interface design involves, if not entirely, partially NN-based solutions.

3.2. AI-driven radio resource allocation & management

Radio resource management (RRM) often involves computationally complex sub-optimal solutions. RRM becomes even more challenging when dealing with 6G envisioned requirements. In the following, we present some applications of ML for RRM that can achieve higher performance. In [12], unsupervised learning is used for uplink power control in a cell-free massive antenna system to achieve max-min user fairness. It has a reduced computational complexity, while keeping tight bounds to the conventional solution. As wireless networks are most representable using graphs, Graph NNs (GNNs) can be used for RRM tasks in supervised or unsupervised manner. Furthermore, Access Point (AP) selection in cell-free network is formulated as a link prediction on a graph, where, given the measurements of known set of APs from a wireless device, a GNN predicts the candidate APs which can serve the device. Here, the GNN leverages the knowledge of static correlated fading, due to static environmental features to predict the APs. Certain RRM tasks rely on the knowledge of the *relative* location of devices and base stations e.g., cell search and hand-over. Additionally, a computationally efficient version of channel charting has recently been proposed, relying on a distance measure in the channel domain that is insensitive to small-scale fading and has been shown to perform well on other tasks, such as channel mapping and positioning . Incorporating timestamp information can greatly improve channel charting, as it allows using contrastive learning methods, such as triplet loss.

Given a large number of heterogeneous sensors connected to Federated Learning (FL) nodes at the network edge, load balancing is necessary to remedy potential hot spots and data diversity to ensure quality balance for the federated learners. By dynamic load rebalancing, we reconnect sensors to nodes in the radio network if load is uneven or some nodes receive insufficient variety of data for serving local models. Another challenge in RRM, related to the inferencing capability and energy efficiency of *edge AI*, is prioritizing learning data transmissions per a data importance criterion. Challenges lie in the joint presence of *channel uncertainty.* e.g., interference, noise, and *data uncertainty*, which is measured by entropy. State-of-the-art works, such as [13] mostly concentrate on time as a resource to be optimized per a data significance criterion, however, more parameters may be considered leading to *frugal* over-the-air learning.

The optimization problems describing RRM are computationally intensive. The research work in Hexa-X demonstrates that techniques from ML can reduce optimization complexity by leveraging from structure in data.

4. Networks for high-performing, sustainable & trustworthy AI

Enabling edge AI in 6G networks calls for a holistic view of several KPIs/KVIs entailing energy efficiency, latency, trustworthiness, etc. In this section, we present the main technological enablers, which we envision to enable in-network AI in 6G, with a deeper look into such targets.

4.1. Joint communication/computation co-design for efficient Edge AI/ML

The convergence of communication and computing for edge AI will play a key role in 6G, as a means of enabling both lean network orchestration and an efficient platform for AI services. First of all, enabling in-network AI functionalities will need to support discovery and selection of the network node(s), such as a user device or an edge cloud server, capable of processing a workload (e.g., an ML model or any application-related processing load) with high performance, security and energy efficiency. The CaaS concept aims to offer such capabilities, to enable services, such as the offloading of computations from end devices to the discovered nodes (e.g. nearby edge cloud servers). To this end, to reduce the communication burden due to frequent exchange of data and/or local models, semantic and goal-oriented communications have been identified as two (possibly inter-playing) technical solutions, as their focus is on the actual interpretation of the receiver, also according to the effectiveness achieved in performing a task, rather than on error free transmissions.

To further improve efficiency, emerging AI architectures need to be explored, due to the radically different requirements on the communication infrastructure. Biologically inspired NNs, like Spiking NNs [14] are often mentioned as the next AI generation, where neurons communicate with spikes within a single device or over wireless channels. The sparsity of the events enables highly energy-efficient operations. These events can be timing sensitive, so the communication medium must keep this relation with high fidelity, while highly localized functions in neuromorphic systems may benefit from network features for local communications.

Changing perspective towards AI for automated radio network control, a promising solution is the concept of Explainable AI (XAI), whose goal is to investigate tools and techniques aimed at opening the so-called opaque (or black-box) models (e.g., Deep NNs - DNNs) or at devising intrinsically interpretable and accurate models (e.g., rule based systems). This concept can be used to explain and distinguish the effect of network configuration and user device load to network performance KPIs. The main technical difficulty is that network configuration (e.g., policy of allocating radio resources) affects the user device load, while the load itself has a more direct effect on the network KPIs. Therefore, XAI will primarily find the importance of the load and explain the predicted KPI based on load, mostly ignoring the explanation of how network configuration and control affects the KPI. The vision is to combine game [15] and information theoretic measures.

From a network resource orchestration standpoint, centralized AI solutions [16] suffer from orchestration reaction time, a crucial issue for 6G use cases with strict latency requirements. AI-based distributed orchestration at the network edge is a promising solution for automatic and efficient up/down scaling of User Plane Network Function instances at the edge. Distributed Network Data Analytics Functions (NWDAFs) [17] and other data sources at the same edge location would feed the AI algorithm for its (re)training, thus reducing the reaction time and the amount of sensitive data sent to the cloud, however with a limit on inference accuracy, due to the scarcity of edge resources. Indeed, the additional computational cost of training the models in each participating node is a challenge to be addressed. Some of the techniques aimed at minimum communication costs and efficient knowledge sharing are FL, blockchain and approaches that combine both. To ensure resource efficiency, privacy, learning quality, and resilience, when deploying AI/ML methods in large scale wireless networks, different resource management and knowledge sharing methods should be evaluated.

4.2. Privacy, security & trust-enhancing enablers for in-network AI/ML

Besides efficiency related KPIs/KVIs, trustworthiness is a requirement to be addressed throughout the ML pipeline, to ensure data protection and privacy, and avoid attacks affecting models robustness. In particular, distributed training agents may behave malevolent and pose model poisoning threat in collaborative learning settings. Certain attacks can be applied to the inference phase [18], such as membership inference, model extraction, or model inversion. To mitigate these attacks, the first step is to implement fundamental security and privacy solutions that constitute a first layer of defense, e.g., implementing proper identity management and access control methodologies for unauthorized data access attacks. Implementing basics will resolve well-known issues, but a second layer, with an ML/AI specific mitigation approach, is needed to handle residual threats. Advanced privacy enhancing technologies can be used to prevent those that cannot be solved via basic methodologies.

In addition to this, advanced mechanisms searching for vulnerabilities are needed, especially in distributed and real-time operation contexts. In this case, data sanitization could rely on techniques, such as anomaly detection, removal of negative impact (elimination of training data with negative impact on the accuracy), training with micromodels (to reduce the risk of attacks), and usage of Generative Adversarial Networks (GANs). The effect of data poisoning (i.e., injection of malicious points in a dataset) in FL, can spread across the network, where the model is distributed. Leveraging on the GAN paradigm, realistic samples resembling adversarial data are generated and used in the training, obtaining FL models immunity to the attack. Furthermore, removing biased data from the original data set is a fundamental aspect to improve accuracy. In 5G, this issue may occur in NWDAF [17] which is composed of two logical components: the Analytics Logical Function (AnLF), performing inference, and the Model Training Logical Function (MTLF), for training the ML model, which can be then consumed on-demand by the AnLF. The data source of MTLF could be biased, leading to potential unfair decisions by the AnLF. To this end, a continuous verification and an unfairness mitigator help avoiding bias and unfair decisions, respectively, the former for detecting sources of unfairness from the AnLF, while the latter for mitigating biased decisions taken by the MTLF. In such a direction, AI systems must also meet explainability requirements, as already mentioned in Section 4.1. However, most of existing FL solutions, conceived for collaborative training of ML models in a privacy preserving fashion, disregard the explainability requirement and only comprise models optimized via Stochastic Gradient Descent, e.g., DNNs. The Federated XAI (FED-XAI) vision is about devising methods and approaches compliant with both FL and XAI paradigms and it specifically targets the collaborative learning of inherently explainable models.

5. Which are the challenges of applying AI/ML in 6G?

The Hexa-X 6G vision involves generation, extraction, collection and processing of data from distributed network entities for different functionalities to enhance network performance. These functionalities introduce challenges from different perspectives ranging from fairness in data processing and inference, processing sensitive data in compliance with the regulations, distributed trust and ownership to sustainability. The distributed nature of data resources and the diversity of the use cases creates a huge amount of data with different granularity ranging from telemetry data, application specific data to management data. More sensitive data will not only be transported by 6G, but also be processed within the 6G system. Therefore, not only following the privacy-by-design approach, but also developing advanced privacy enhancing mechanisms enabling distributed and collaborative computations will become inevitable.

Decentralized and disjoint AI-driven 6G network functions are expected to be deployed as cloud services. The ownership of the 6G components, AI capabilities, and the data being processed may belong to different parties. In addition, the underlying cloud infrastructure may or may not be operated by those parties handling 6G functionalities. Thus, the diversity of the ownership and the control in the architecture create complex trust relationships. AI life-cycle comprising training, development, and deployment of the model and operating and managing the AI service should be carefully designed taking these complexities into account. In such a distributed learning environment involving network domains of different levels of trust, the generalization capability of an ML model may be affected, due to introduced limitations to data ingestion from different geographical areas where different data use regulations may be in effect. From a standardization point of view, a challenge is to specify interoperable interfaces facilitating AI agent discovery and selection, while, from a technology standpoint, the challenge is enhancing ML model generalization capability by not violating personal and corporate data regulations. Sustainability is another fundamental aspect of the 6G vision [3]. Since data traffic grows at high pace, also considering the pervasive deployment of computing resources needed to enable AI, along with the unavoidable network densification, ever higher efficiency is required to limit network energy consumption, thus keeping a positive balance between direct and indirect effects of the ICT industry on the global carbon footprint. Sustainable AI in 6G poses unprecedented challenges in terms of hardware re-usability, lean architectural design and operations, as well as higher reliance on renewable energy sources.

6. Conclusions

In this paper, we have introduced the main motivations, envisioned opportunities and incurred challenges of widely using AI (with emphasis on ML) in future 6G mobile communications systems, per the vision of the EU 6G Flagship project Hexa-X. First, we have introduced several technical enablers for designing the 6G air interface using data-centric techniques, applicable to wireless transceivers for high performance signal transmission and reception, aiming to offer wireless device cost, energy and complexity reductions and deal with wireless transceiver hardware impairments. Enablers for efficient radio resource and interference management have also been elaborated to address large-scale transmission problems suffering algorithm deficiencies. Then, methodologies for joint communication and computation system co-design have been also described to address engineering trade-offs for different collaborative learning structures. Privacy, security and trust-enhancement enablers for in-network AI/ML have been also elaborated. Finally, challenges of applying AI/ML in 6G networks have been detailed, calling for innovation by both AI and wireless communications communities regarding both aspects of “learning to communicate” and “communicating to learn” in 6G cellular networks.

7. Acknowledgments

This work has been partly funded by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956).

References

[1] N. M. K. Chowdhury, R. Boutaba, A survey of network virtualization, Computer Networks 54 (2010) 862–876.

[2] F. Giust, X. Costa-Perez, A. Reznik, Multi-access edge computing: An overview of ETSI MEC ISG, IEEE 5G Tech Focus 1 (2017) 4.

[3] Hexa-x deliverable D1.2 - expanded 6G vision, use cases and societal values – including aspects of sustainability, security and spectrum, 2021. URL: https://hexa-x.eu/wp-content/uploads/2021/05/Hexa-X_D1.2.pdf.

[4] Hexa-x deliverable D4.1 - AI-driven communication & computation co-design: Gap analysis and blueprint, 2021. URL: https://hexa-x.eu/wp-content/uploads/2021/09/Hexa-X-D4.1_v1.0.pdf.

[5] D. Korpi, M. Honkala, J. M. J. Huttunen, F. A. Aoudia, J. Hoydis, Waveform learning for reduced out-of-band emissions under a nonlinear power amplifier, 2022. arXiv:2201.05524.

[6] H. Farhadi, M. Sundberg, Machine learning empowered context-aware receiver for high-band transmission, in: 2020 IEEE Globecom Workshops (GC Wkshps), 2020, pp. 1–6.

[7] Y. Chen, J. Mohammadi, S. Wesemann, T. Wild, Turbo-AI, Part II: Multi-dimensional iterative ML-based channel estimation for B5G, in: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 2021, pp. 1–5.

[8] L. Le Magoarou, S. Paquelet, Online unsupervised deep unfolding for MIMO channel estimation, 2020. arXiv:2004.14615.

[9] D. Marasinghe, N. Rajatheva, M. Latva-aho, LiDAR aided human blockage prediction for 6G, in: 2021 IEEE (GC Wkshps), 2021, pp. 1–6.

[10] L. Le Magoarou, T. Yassine, S. Paquelet, M. Crussière, Deep learning for location based beamforming with NLOS channels, in: IEEE (ICASSP), 2022.

[11] G. Larue, L.-A. Dufrene, Q. Lampin, P. Chollet, H. Ghauch, G. Rekaya, Blind neural belief propagation decoder for linear block codes, in: 2021 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit), 2021, pp. 106–111.

[12] N. Rajapaksha, K. B. Shashika Manosha, N. Rajatheva, M. Latva-aho, Deep learning-based power control for cell-free massive MIMO networks, in: IEEE ICC, 2021, pp. 1–7.

[13] D. Liu, G. Zhu, Q. Zeng, J. Zhang, K. Huang, Wireless data acquisition for edge learning: Data-importance aware retransmission, IEEE TWC 20 (2021) 406–420.

[14] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, A. Maida, Deep learning in spiking neural networks, Neural Networks 111 (2019) 47–63.

[15] C. Frye, C. Rowat, I. Feige, Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, NIPS 33 (2020) 1229–1239.

[16] X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina, Network slicing in 5G: Survey and challenges, IEEE Communications Magazine 55 (2017) 94–100.

[17] 3GPP, Technical Specification 23.288, Architecture enhancements for 5G System (5GS) to support network data analytics services (Release 17), 2021.

[18] L. Melis, C. Song, E. D. Cristofaro, V. Shmatikov, Inference attacks against collaborative learning (2018). arXiv:1805.04049.