Novel AI/ML models to achieve zero-touch service management (ZSM) in 6G

1. Context

Zero-touch service management (ZSM) is a novel paradigm for managing new generations of networks. ZSM aims to rely on AI/ML algorithms to reduce human intervention at the minimum up to null to allow networks and systems to be autonomous and able to run self-configuration, self-healing, and self-correction. ZSM uses the principle of a closed-control loop, which usually consists of three functions or elements: Monitoring System (MS), Analytical Engine (AE), and Decision Engine (DE). These three elements are already known in the Infrastructure as a Service (IaaS) management process of virtual resources, principally relying on a centralized element (i.e., orchestrator) that runs the three entities. However, with the emerging 6G networks that 6G systems are expected to serve a massive number of extremely heterogeneous Network Slices that cross multiple technological domains (i.e., RAN, Edge, Cloud, and Core), posing significant challenges to classical centralized management and orchestration approaches in terms of scalability and sustainability, the three elements are highly distributed among the actors managing and orchestrating network slice components and resources. MS is in charge of monitoring Key Performances Indicators (KPI) and various relevant events from the different components running a network service (NFV Infrastructure - NFVI, SDN Controller, RAN, etc.) as well as from the VNFs and PNFs composing a network slice. Note that MS may use a DB to store the collected monitoring data for future analysis. It can be used for instance to understand the long-term evolution of a system. MS periodically transmits monitoring information to AE processes the data, and provides the required analysis output to the DE. The latter, using a pre-deployed policy or automatic decision mechanisms, decides on the LCM action to apply; in the case of a VNF a LCM action is to scale up or down the resources (e.g., CPU) or migrate the VNF to another NFVI. These actions are enforced by the DE using interfaces with the components managing the network service (i.e., orchestrator). AE and DE are highly driven by AI/ML techniques aiming at learning appropriate LCM decisions to consider, according to the state of the infrastructure, the state of network slice components (obtained from AE), and depending on the technological domain where each slice component is deployed (e.g., RAN).

In this thesis, we will focus on the AI/ML algorithms that will run inside the key elements of the closed-control loop, i.e., AE and DE. To this aim, we will concentrate on novel AI/ML solutions that are applicable to the challenges that arise in 6G and new-generation networks. As stated earlier, 6G needs a distributed and scalable management plane that should rely natively on AI/ML, which requires novel AI/ML for training and transferring models between the different domains composing the infrastructure that runs 6G services. In this case, we will move toward a distributed closed control loop that is composed of not only one AE, but different AEs, each for each domain. Similarly, the DE needs to be distributed and jointly cooperate for decision-making policies. Hence, this calls for new distributed and cooperative AI/ML algorithms.

2. Objectives and approach

As introduced earlier, this thesis project will shed light on novel algorithms for distributed learning and decision-making, particularly adapted to the distributed and highly scalable 6G system. In more detail, the objectives of the thesis are related to the topic of distributed machine learning and decision-making by addressing the following topics:

Distributed, federated, and split learning

Distributed and federated AI/ML72 is an emerging approach that aims at leveraging distributing data, computing, and memory resources, for model training while exploiting the (often unused) capabilities of edge devices without violating or preserving data privacy. Popular examples include federated learning and split learning73,74 using a parameter server or in meshed topology, which focus on training of a global ML model or parts thereof at different computing nodes in the network. In this thesis, we will go beyond the existing models by exploring the concept of Composable AI, which h will efficiently realize a new concept of ML as a Service (MLaaS) that suits distributed network systems well, such as 6G. In such a new concept, an ML model trained within one or more network domains leveraging such distributed approaches as federated learning or split learning and using local data and resources, can be made available to others. Similarly, a domain requiring an ML model can receive an already-trained model and further refine it using its own data, thus greatly saving computational power and memory. Composable AI, therefore, enables ensemble as well as transfer learning to let ML models be collaboratively built, and it leverages meta-learning techniques to always select the best available ML model for the execution of a specific task in each domain and context.

Distributed and multi-agent Reinforcement Learning

6G will involve different domains limiting the centralized approach of existing decision making solution. In this context, distributed decision making is needed, where distributed Reinforcement Learning approaches such as MARL7 have emerged to improve scalability. Although in distributed RL approaches training and control are performed in a decentralized way, targeting global optimality is essential for minimizing system costs. Since the agents' decisions are usually based on partial observations of their local environment, achieving a global optimal solution may be hard. One way to do so is to use a common state-action value (Q) function for all elements. Further, three main existing challenges may inhibit MARL from achieving global optimality. First, when multiple RL agents are acting in the same environment, the environment becomes non-stationary from the perspective of any individual agent11. Second, policy-based approaches present high variance when multiple agents learn collaboratively. In this thesis, we will address all the above-mentioned (three) challenges faced by today's MARL approaches and essentially strive to balance contradictory requirements, in many situations weighing the benefits of centralization and decentralization, and making appropriate trade-offs to smartly use the advantages of each type of infrastructure. More specifically, regarding the co-existence of multiple learning agents leading to non-stationary environment, project the envisioned algorithms will focus on devising novel modelling approaches that allow for approximations that bear stationary characteristics, e.g., such as multiple exponential stages that can approximate general arrival distributions as Markovian ones.

Attention learning and Transformer architecture for time series

Most of the problems the ML model needed to predict the performance of new generation systems, such as 6G, rely on time-dependent data. For instance, the traffic to be handled by a base station is generally time-dependent; detecting the failure of a system, like a base station, depends on variables that evolve with time. RNN and LSTM were mainly used to predict the performances of a new generation network that depend on data in form of time series. However, RNN and LSTM have limitations when it comes to capturing long-range dependencies in time series data, as they suffer from the vanishing gradient problem. Transformer, on the other hand, is a new ML architecture, initially designed for natural language processing tasks, which has shown a high ability when it comes to time series data analysis. It offers several advantages that make it well-suited for handling time series data, although it wasn't initially designed for such data. Here's how the Transformer architecture can be helpful in time series analysis. In this thesis, we will explore the usage of Transformer architecture, particularly the self-attention mechanism which allows to capture dependencies and relationships between data points at different time steps to address networking problem in 6G that needs efficient mechanisms to handle time series, such as failure detection, traffic models, etc.

3. Organization

The PhD project is divided into three phases:

- The first phase, from M0 to M6, will be dedicated to related work on 6G, Zero touch Service Management (ZSM), Autonous Networks, Distributed ML models: federated learning, split learning, MARL, ML models for time series: RNN, LSTM, Transformer.
- The second phase, from M6 to M30, will consist of devising algorithms and mechanisms to address the challenges cited earlier.
- The last phase, from M30 to M36, will be dedicated to the PhD document and the preparation of the final defense.

For the dissemination activities, we aim to publish and demonstrate the devised works in peer-reviewed conferences, such as IEEE ICC, Globecom, ICML and Infocom. During the final year, one or two publications will be submitted to peer-reviewed journals. Moreover, we aim to implement some mechanisms on top of 6G orchestror to run demonstration in international conferences.

Contact:

Prof. Adlen Ksentini (adlen.ksentini@eurecom.fr)
Prof. Bouziane Brik (bouziane.brik@u-bourgogne.fr)

References:

[1] H. Chergui, A. Ksentini, L. Blanco and C. Verikoukis, "Toward Zero-Touch Management and Orchestration of Massive Deployment Slices in 6G Network", in IEEE Wireless Communications Magazine, 2022

[2] J. Konecny, et.al., "Federated optimization: Distributed optimization beyond the datacenter," arXiv preprint arXiv:1511.03575, 2015

[3] P. Vepakomma, at.al., "Split learning for health: Distributed deep learning without sharing raw patient data", 2018

[4] D. Rosendo, et al. "Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review", Jou of Parallel and Distributed Computing, 2022.

[5] R. Lowe, et al., "MultiAgent Actor-Critic for Mixed Cooperative-Competitive Environments", NIPS'17, 2017

[6] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J.Yan, L. Sun, "Transformers in Time Series: A Survey", in the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)