Machine Learning-Powered Management Architectures for Edge Services in 5G Networks

Corrado Puligheddu

Next-generation mobile networks are designed to allow vertical industries to offer a broad range of virtualized services to their users. However, increasingly evolved mobile services force more and more demanding performance requirements, which are particularly hard to guarantee, without expensive overprovisioning, in case of time-varying service and traffic demands. This work addresses the issue by taking two complementary approaches.

First, we propose a centralized automation solution for the provisioning of edge services and management of edge resources. Using service scaling (i.e., allocating more resources to a service, only when needed, to manage more traffic) the proposed solution can ensure the satisfaction of performance requirements of an automotive vertical service. We then improve our solution by integrating the concept of ML-as-a-Service (MLaaS) through a MLaaS Platform able to train and serve ML models to the elements of a 5G system, thus giving them the possibility of making smarter decisions. We demonstrate the new capabilities of our proposal by developing two ML-driven algorithms for network slice subnet sharing and run-time service scaling. Results show that service performance can be always satisfied while saving 30% on OPEX.

The second approach investigates distributed RAN orchestration and edge resource management using reinforcement learning. In this case, the decision-making logic is colocated with the services and applications it controls, allowing local fine-grained and low-latency actions. We propose two reinforcement learning frameworks for edge resource management: CAREM and VERA. CAREM operates in heterogeneous vRANs; it is able to select the best available radio link and the transmission parameters, enabling efficient radio resource allocation in time-varying scenarios. CAREM exhibits excellent performance when compared both to the closest existing scheme based on neural networks, and to a contextual bandit approach. Instead, VERA addresses the concurrent execution of two kinds of services at the edge, namely user applications and network functions. We show that often the computing resources required by these services are entangled since the data processed by the

former has to be transferred by the latter and vice versa. Acknowledging this complex dynamic, we propose a scalable reinforcement learning-based framework to orchestrate resources at the edge. Considering as services an LTE vRAN and a video transcoder, we demonstrate that VERA is able to meet services KPIs over 96% of the observation periods.