ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXV Cycle

SEM-O-RAN: Semantic and Flexible O-RAN Slicing for NextG Edge Mobile System

Corrado Puligheddu

Supervisor: Prof. Carla Fabiana Chiasserini

Research context and motivation

- Vehicle to Everything communications (V2X) are enabling drone-based delivery and autonomous vehicles, a market that will reach a global revenue of $50 billions by 2024.
- To perform their mission-critical operation, autonomous vehicles and drones will require the continuous execution of complex computer vision (CV)-based tasks, often based on Deep Learning (DL) models. An example is the multi-object classification of blockages, intersections, and people from high resolution images or 3D LIDAR data.
- However, mobile devices are often battery-powered, therefore, with a low power budget, they don't always have the compute resources needed to locally execute CV tasks.
- By offloading the tasks to the Edge, where there is not such power limit and resource are more freely available, the mobile device can speed up the task execution while saving energy. Nonetheless, continuously sending multimedia the Edge may eventually saturate the RAN capacity. Thus, careful slicing is required to accommodate DL tasks at the Edge.

Addressed research questions/problems

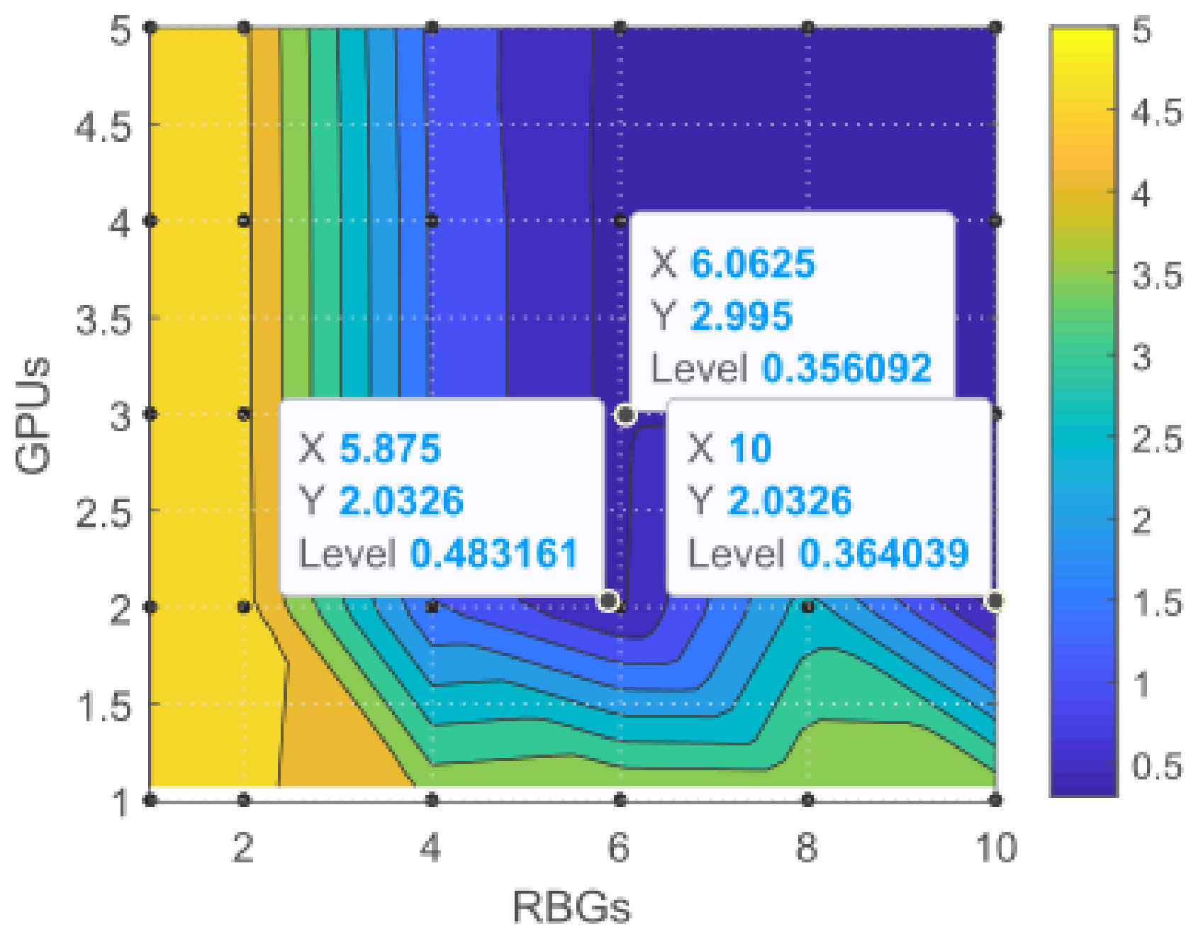
- In the Cityscapes dataset, a large scale database of videos recorded in 50 different cities, each image has a 100 KB size on average. If we assume that a real-time self-navigation system requires DL inference on frames collected from 4 cameras each 10 ms, the traffic load would be 32 Gb/s if 100 vehicles are connected to the RAN at the same time. Clearly, the traffic load would largely exceed the current fastest RAN capacity (~2 Gb/s).
- **Given limited edge (radio and compute) resources, how to maximize the number of task admitted to be offloaded to the edge?**

Novel contributions

- SEM-O-RAN is the first O-RAN slicing framework for edge-assisted mobile applications. With respect to the state of the art, it proposes two core innovations:
  - The task is defined in terms of required *end-to-end latency* and *accuracy-per-class performance,* thus allowing **flexibility** in the way edge resources of different types are allocated. Flexibility allows for the consideration of multiple edge allocations leading to the same task-related performance, ultimately leading to better resource utilization.
  - SEM-O-RAN considers the **semantics** of the DL task to further reduce the network overhead by compressing the images while guaranteeing the minimum accuracy-per-class performance. In fact, different DL classifiers can tolerate different levels of image compressions due to the semantic nature of the target classes.
- We formulate the Semantic Flexible Edge Slicing Problem (SF-ESP), demonstrate that is NP-hard and propose a greedy heuristics to solve it efficiently.
- We shows that SEM-O-RAN allocates up to 169% more tasks than the state of the art.

Flexible resource allocation

- Tasks resource allocation is not fixed, thus multiple combinations of resource allocations lead to the same latency performance. Therefore, the resource allocation can be chosen as to maximize the number of admitted tasks.
- Let us assume that resource types are Resource Blocks Groups (RBGs) and GPUs, and that 25 RBGs and 4 GPUs are available at the edge. Two tasks, requiring a latency of 0.4s, have to be allocated. Both (6, 3) and (10, 2) meet the latency requirements, however the available resources are not sufficient to allocate the minimum resources (6, 3) to both tasks. On the contrary, by allocating (10, 2) both tasks can be admitted for offloading since resources are enough.



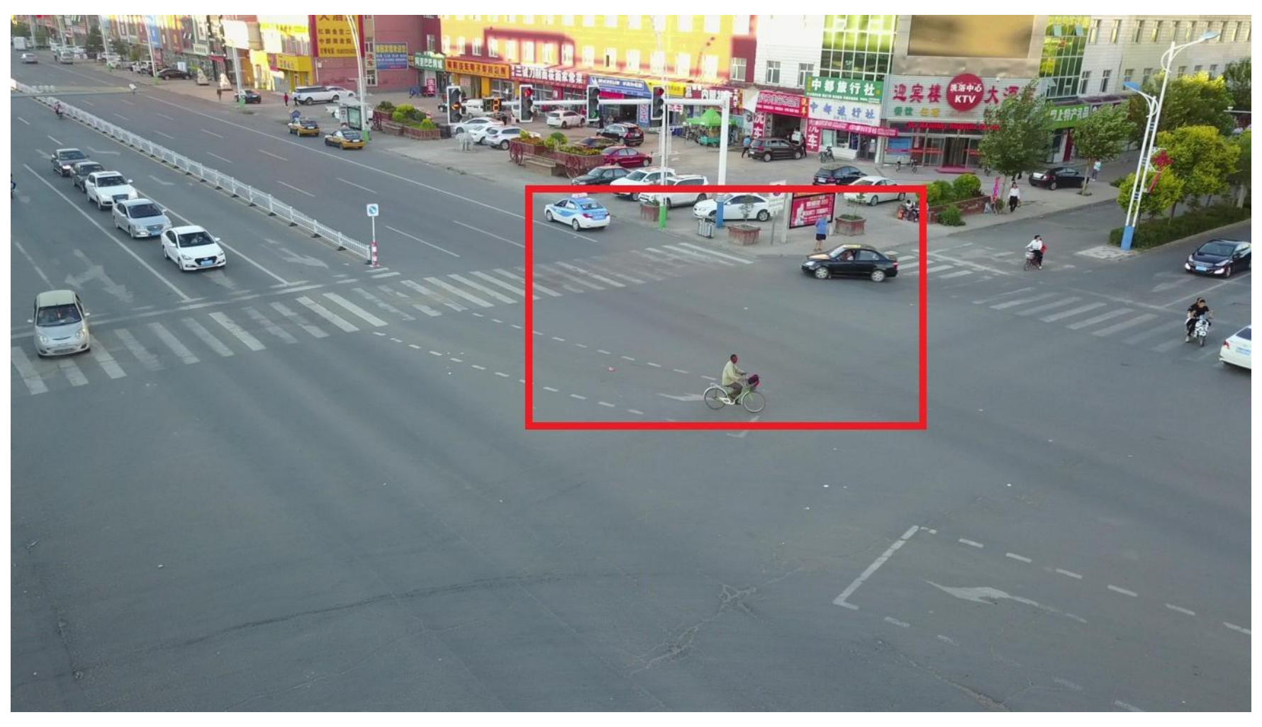
Latency for different resource allocations

Submitted and published works

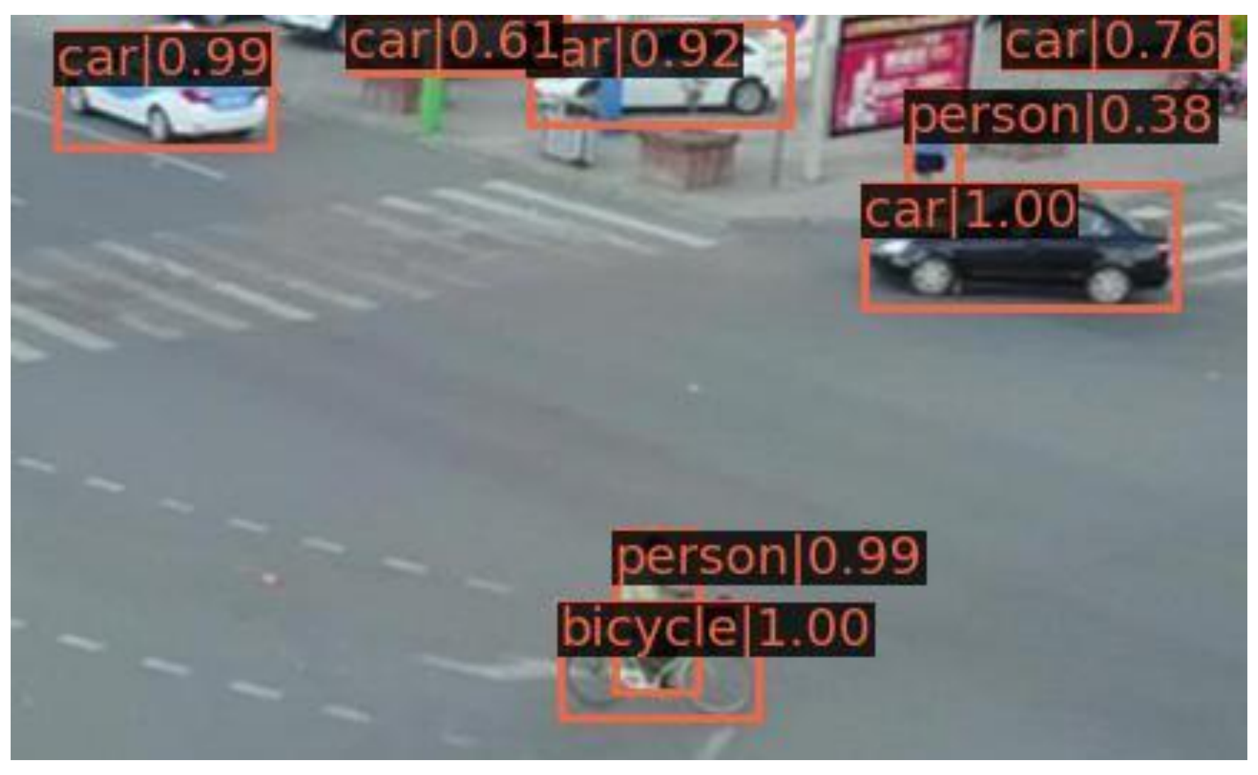
- S. Tripathi, C. Puligheddu, S. Pramanik, A. Garcia-Saavedra and C. F. Chiasserini, "VERA: Resource Orchestration for Virtualized Services at the Edge“, ICC 2022 - IEEE International Conference on Communications, 2022, pp. 1641-1646
- C. Casetti et al., "ML-driven Provisioning and Management of Vertical Services in Automated Cellular Networks“, IEEE Transactions on Network and Service Management
- C.Puligheddu, J. Ashdown, C. F. Chiasserini, and F. Restuccia, “*SEM-O-RAN: Semantic and Flexible O-RAN Slicing for NextG Edge-Assisted Mobile Systems*”, INFOCOM 2023 - IEEE International Conference on Computer Communications

Semantics-based compression

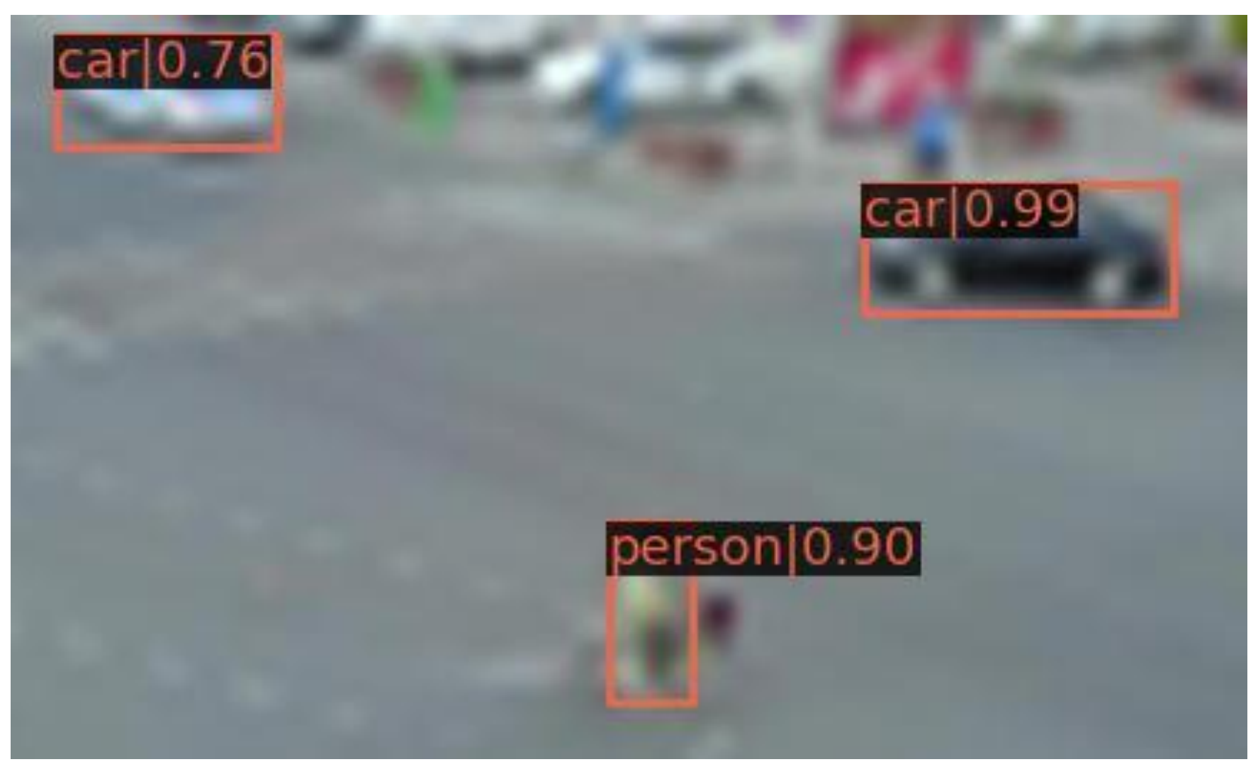
- In the task definition, the Virtual Network Operator (VNO) provides the target accuracy to be obtained on selected object classes.
- A frame captured by a smart city camera is compressed using a light and a heavy compression factors. In the first case, the DL classifier is able to detect the bicycle object, which is semantically harder to recognize compared to a person or a car. In the second case, only the two cars crossing the intersection and the person in the bicycle are detected.



(a) Original frame without compression



(b) Crop of the lightly compressed frame

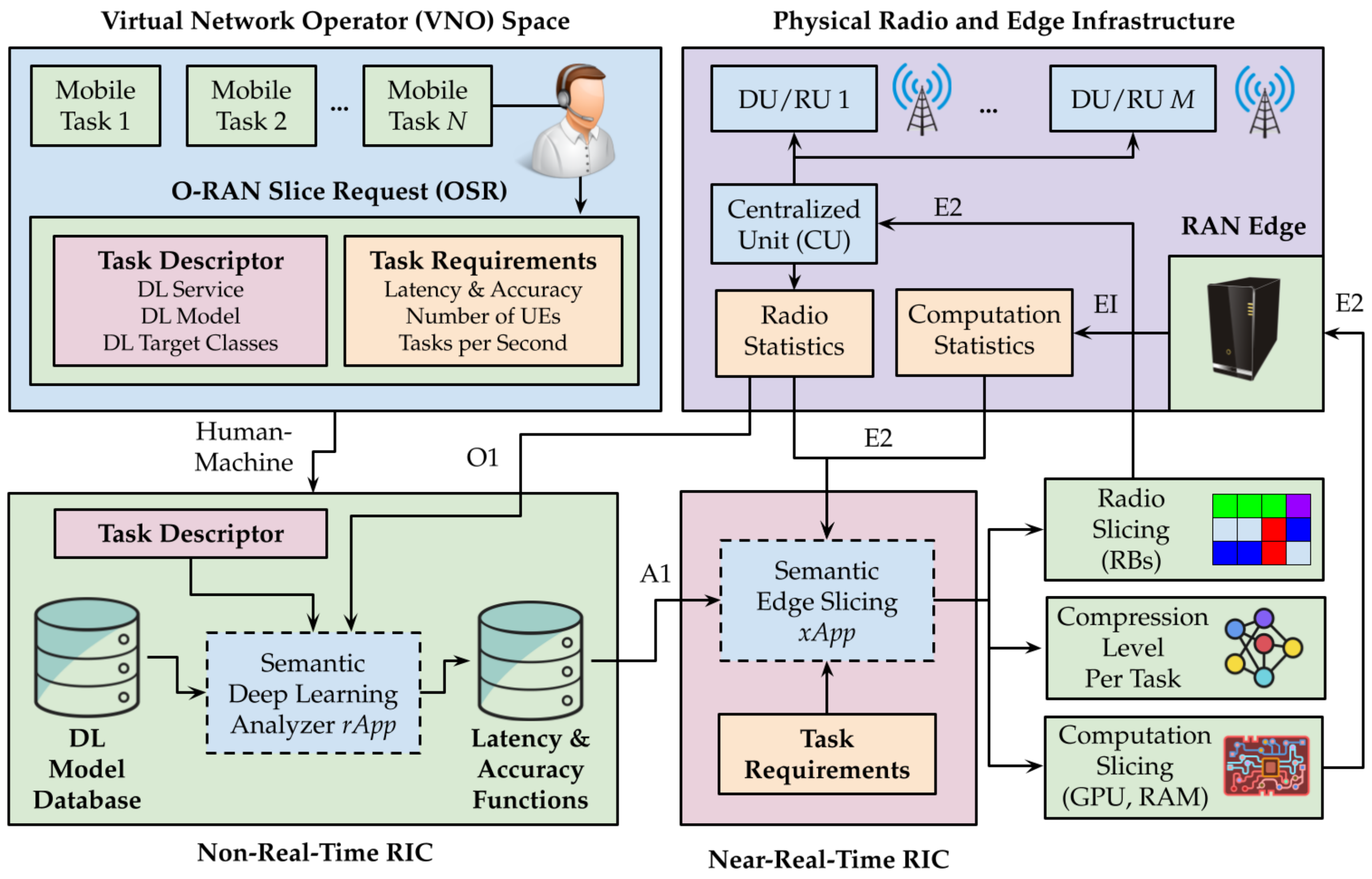


(c) Crop of the highly compressed frame

- The highest compression factor allows for the highest network bandwidth saving, thus allowing more tasks to be offloaded concurrently with the same resource consumption.

Methodology

- SEM-O-RAN exploits the O-RAN architecture to control slice resource allocation according to the VNO task requests and the network status. It is constituted by two modules:
  - Semantic Deep Learning Analyzer (SDLA) rApp: it receives the task requests by the VNO, it calculates the accuracy functions for the offered models using datasets that are representative of the task, and the latency functions for the mobile devices connected to the RAN. It runs in the O-RAN Non-RT controller.
  - The Semantic Edge Slicing xApp: it admits tasks for offloading by solving the SF-ESP in the O-RAN Near-RT controller using the latency and accuracy functions computed by the SDLA and the real time radio statistics from the RAN.



Future work

- Image compression is an intentional degradation of the quality of the task input data performed to save bandwidth. **What if the data quality is compromised by external factors** (e.g., bad weather)?
- Rain could degrade the image quality and prevent compression to mantain the minimum task accuracy.

List of attended classes

- 01DTPRV – Connected Vehicles (23/06/2022, 4 credits)
- Lipari School on Advanced Networking Systems (04/07/2022, 6 credits)

Politecnico di Torino
1859

PhD program in
Electrical, Electronics and Communications Engineering