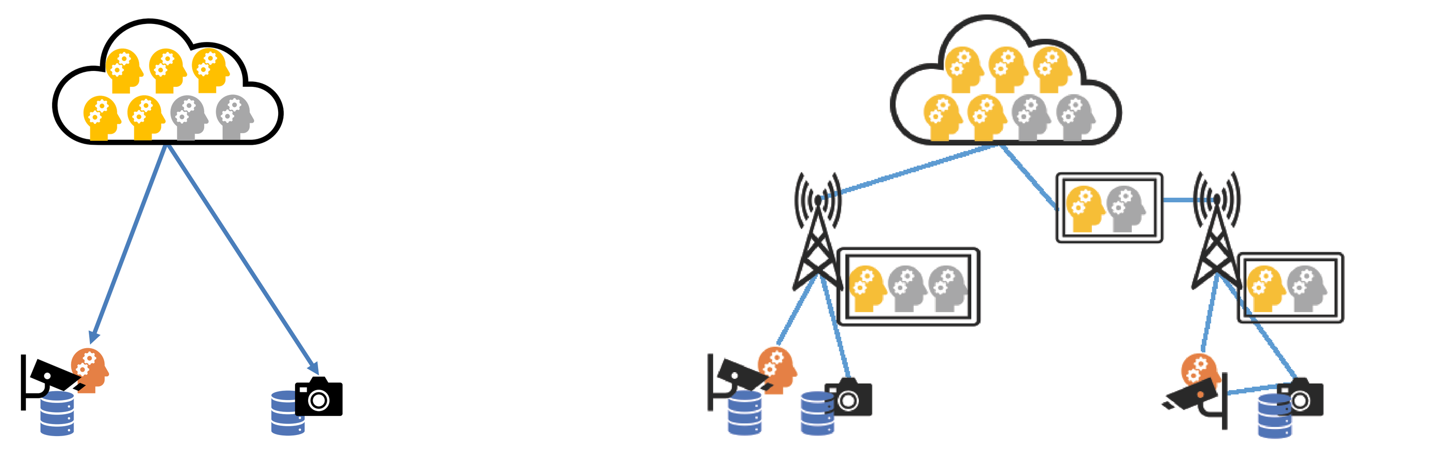
XXXVII Cycle

Efficient Distributed DNNs in the Mobile-edge-cloud Continuum

Giuseppe Di Giacomo

Supervisor: Prof. Carla Fabiana Chiasserini

Research context and motivation

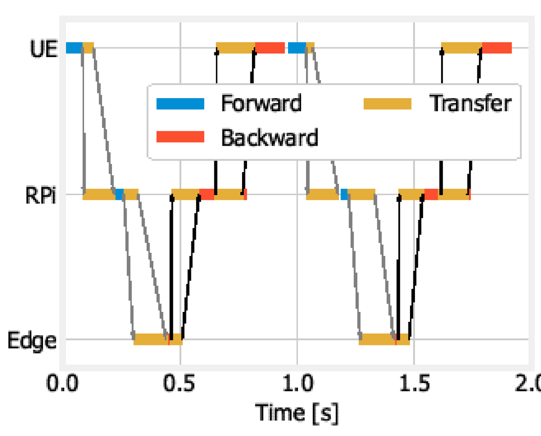


- Current scenario: Machine Learning (ML) applications at **cloud-based servers**
- Future scenario: edge intelligence paradigm, based on **Mobile-edge-cloud Continuum**

- **Distributed learning** envisions leveraging data and resources of multiple nodes in order to perform a common learning task. It is an excellent fit for edge intelligence scenarios:
  - capability of edge and mobile devices is growing faster than cloud ones
  - need to keep as much as possible the information local
  - recent works on ML techniques tailored for lower-powered devices has extended the types of devices that can be leveraged for distributed ML

Addressed research questions/problems

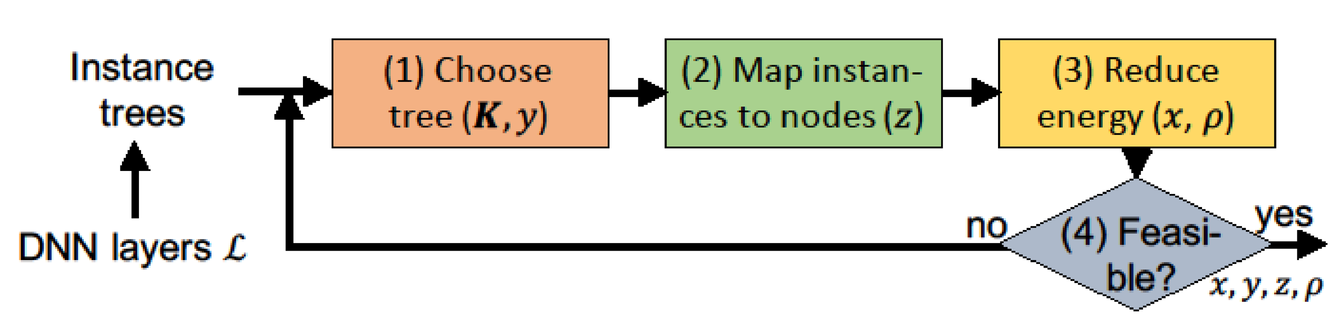
- Distributed learning is influenced by three main factors:
  1. the sequence of layers composing the Deep Neural Network (DNN)
  2. the available resources
  3. the quantity and diversity of the input data
- Decisions regarding such factors impact one another: it is important that they are made jointly

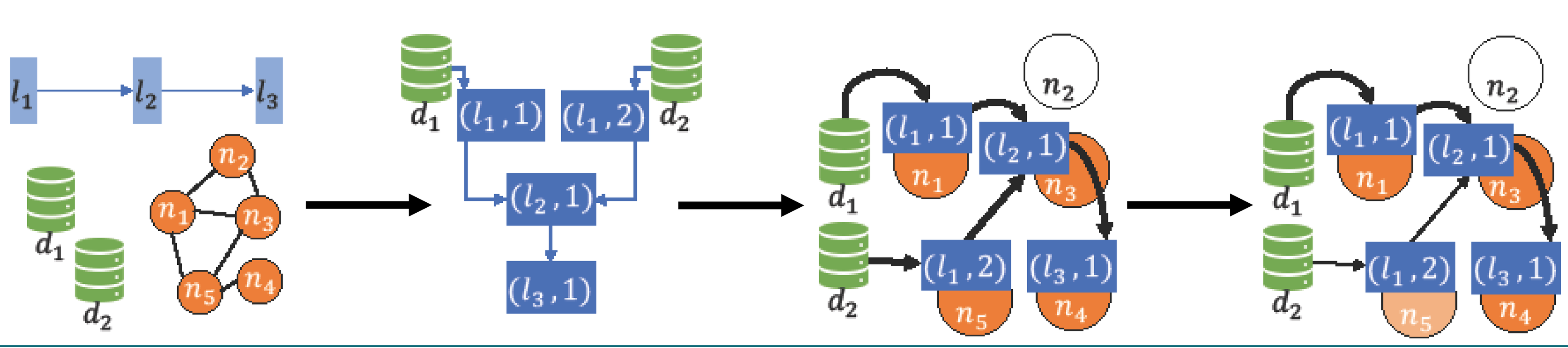


- Limits of state-of-the-art works on distributed ML:
  - emphasis on learning effectiveness over **efficiency**
  - Split Learning (SL) is a paradigm based on partitioning the DNN among the learning nodes. However, SL works focus only on placement decisions
  - in Federated Learning (FL) all nodes train the same DNN

Novel contributions

- **RightTrain**: framework able to make **joint and high-quality decisions** in order to adapt the DNN structure, the network resources and the size of the datasets to one another, while **minimizing the energy consumption** required to reach a **target learning quality** ε^{max}





Submitted and published works

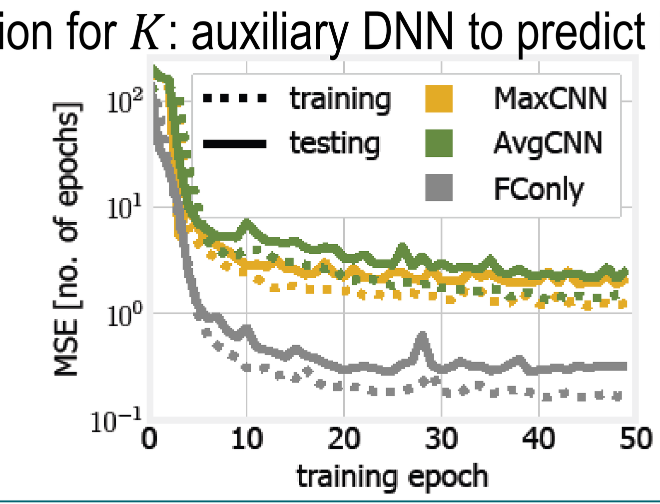
- Dang, V.N. *et al.*, "*Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation*", Medical Image Analysis, vol. 75, no. 102263, 2022
- Malandrino, F., Chiasserini, C. F., Di Giacomo, G., "*Energy-efficient Training of Distributed DNNs in the Mobile-edge-cloud Continuum*", WONS 2022, 2022
- Di Giacomo, G., Haerri, J., Chiasserini, C. F.,"*Edge-assisted Gossiping Learning: Leveraging V2V Communications between Connected Vehicles*", ITSC, 2022

Adopted methodologies

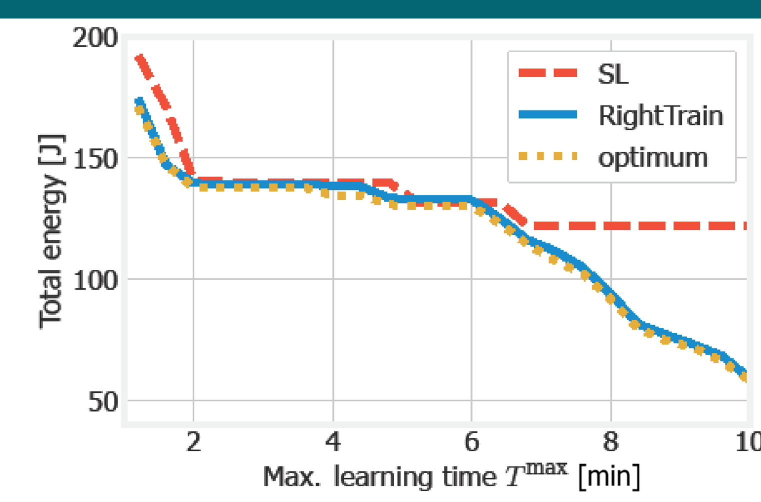
- Description of system model and problem formulation

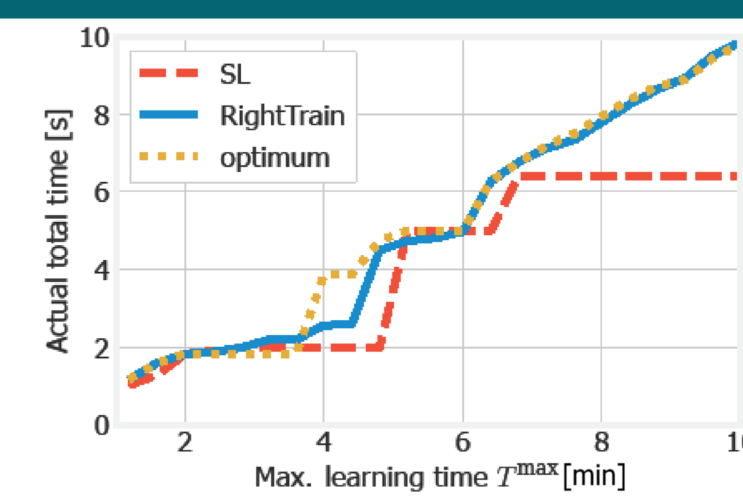
$$\min_{x,y,z,\rho} K(y,\varepsilon^{max})E(x,y,z,\rho)$$

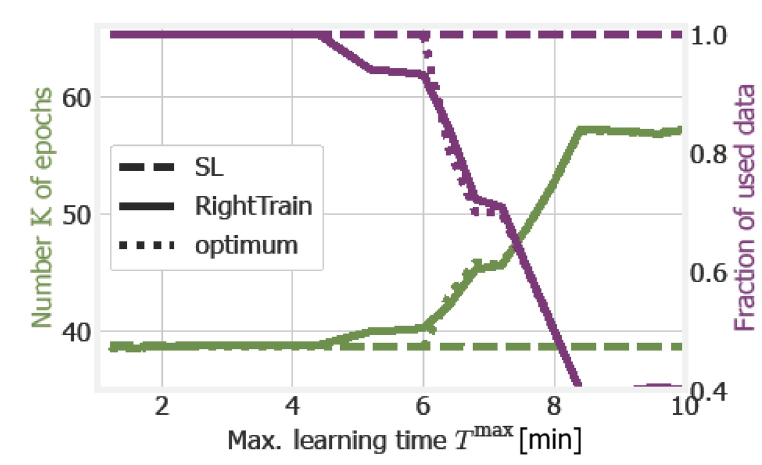
  - $x(l,i,m,j)$: data flowing from layer instance (l,i) to instance (m,j)
  - $y(l,i,m,j) \in \{0,1\}$: expresses whether layer instance (l,i) is connected to layer instance (m,j)
  - $z(l,i,n) \in \{0,1\}$: expresses whether instance (l,i) is deployed at node n
  - $\rho(l,i)$: computational resources to be assigned to instance (l,i)
- No closed-form expression for K: auxiliary DNN to predict its value

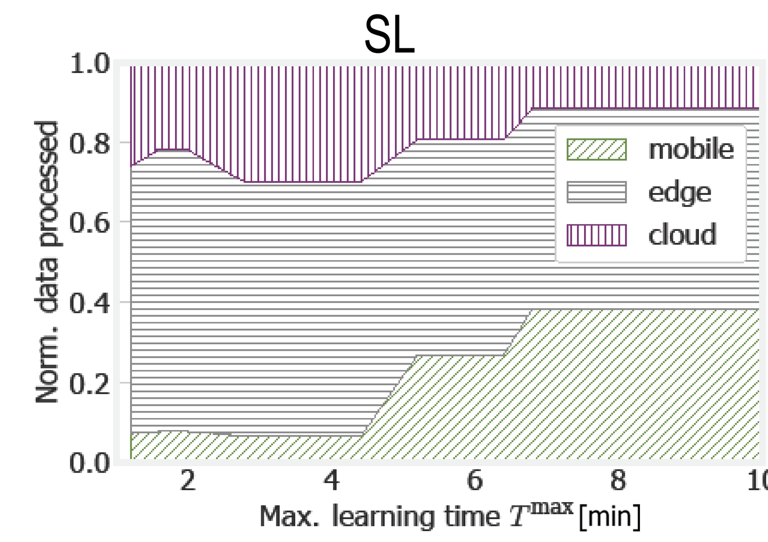


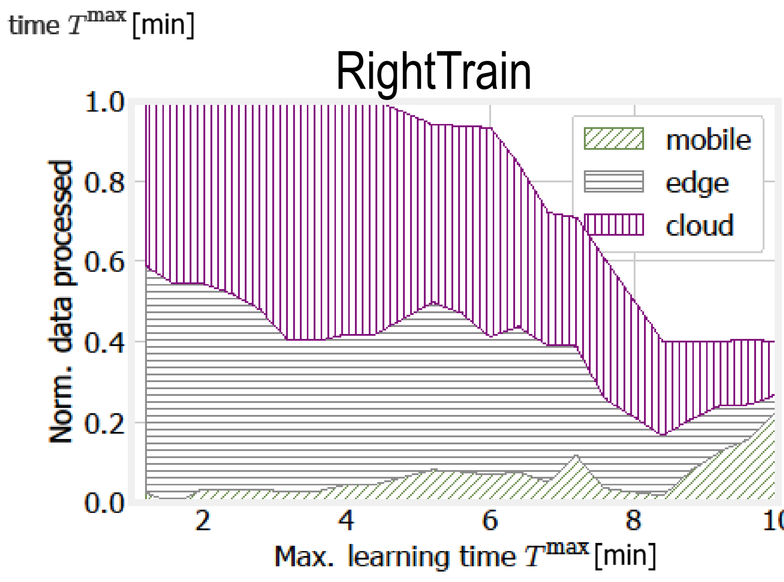
Performance evaluation



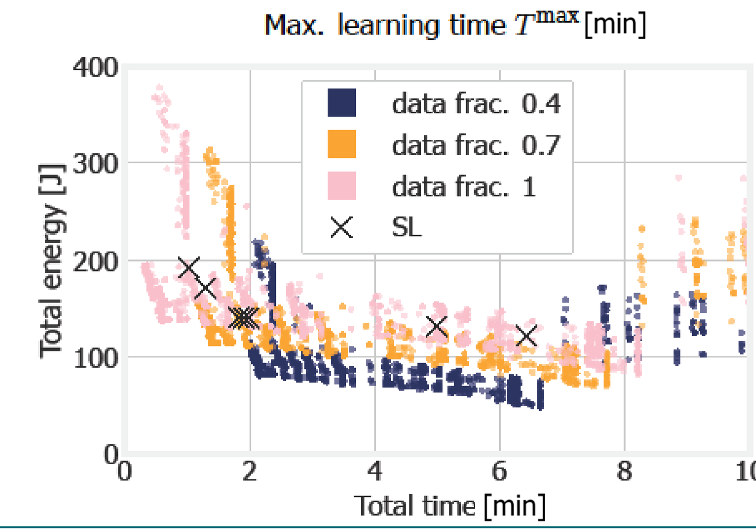








- Effect of different fractions of used data



Future work

- Use more complex DNNs
- Generalize our results using other datasets
- Integrate RightTrain with techniques like pruning and knowledge distillation

List of attended classes

- 01TRARV – Big data processing and programming (1/3/2022, 4 credits)
- 01DTPRV – Connected Vehicles (didattica di eccellenza) (23/6/2022, 4 credits)
- 01QTEIU – Data mining concepts and algorithms (3/2/2022, 4 credits)
- 01RGBRV – Optimization methods for engineering problems (7/6/2022, 6 credits)
- 02SFURV – Programmazione scientifica in matlab (21/04/2022, 6 credits)

PhD program in

Electrical, Electronics and Communications Engineering