AIMH
Artificial Intelligence for Media and Humanities

Università di Pisa

HPC lab

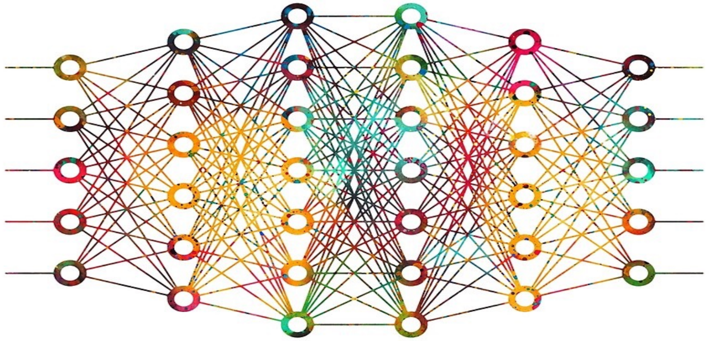
ISTI
Consiglio Nazionale delle Ricerche
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Towards Knowledge Distillation in Decentralized Learning for Edge AI: Challenges and Future Directions

Presenter: Molo Mbasa Joaquim

Introduction

1. AI emergence



Complex deep learning model

- Provide high accuracy
- Need large amount of data

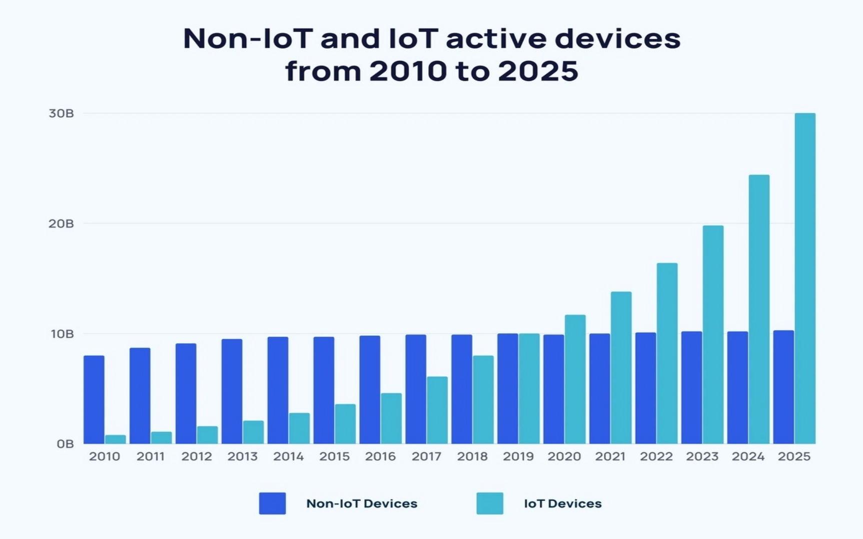
Cloud Computing platform

- computational resources (GPUs)
- Storage resources

Introduction

1. AI emergence

Statistics



International Data Corporation

- *Amount of data will exceed 90 ZB*

Consequence:

- *Need for computational resources*
- *Need for storage*
- *Need for bandwidth*

Introduction

2. Emergence of Edge Computing

Massive amount of real-world data

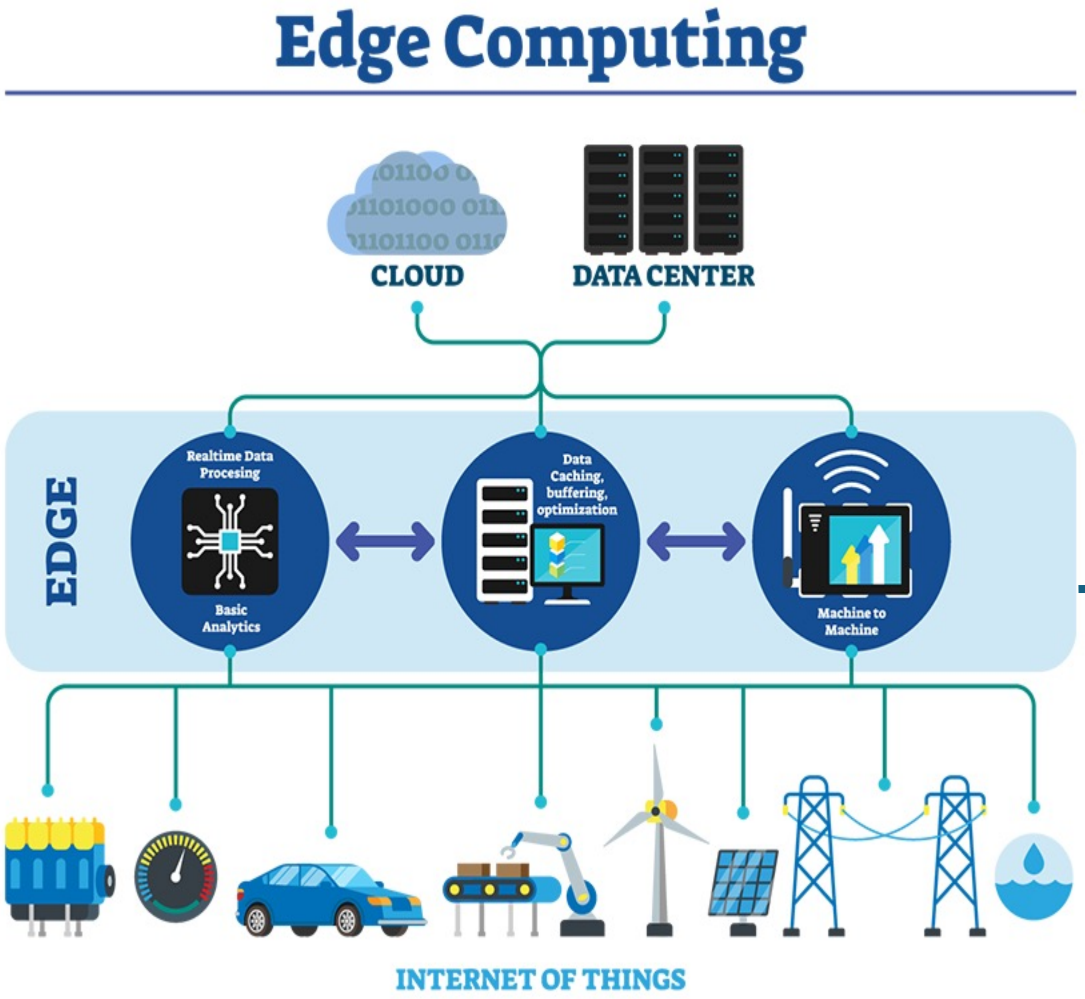
CHALLENGE ACCEPTED

Edge Computing

- Real-time processing of the data at the edge of the network

Introduction

❖ Edge Computing emergence



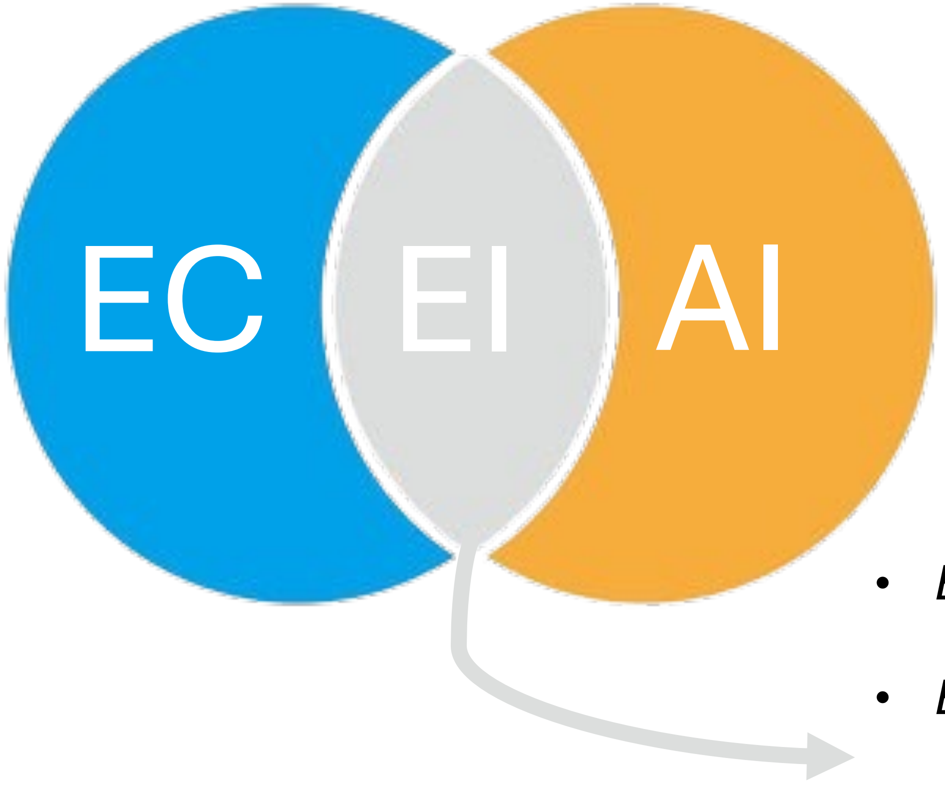
Edge Computing as distributed network architecture that processes data close to its source to reduce bandwidth usage and application latency

Edge Server : Process part of the computation

Edge devices ⟹ Resource constrained

Introduction

❖ Edge Intelligence emergence



- *EI* involves the integration of *EC* and *AI* utilizing majority of the resources at the edge of the network to offer intelligent insights independently of centralized resources.

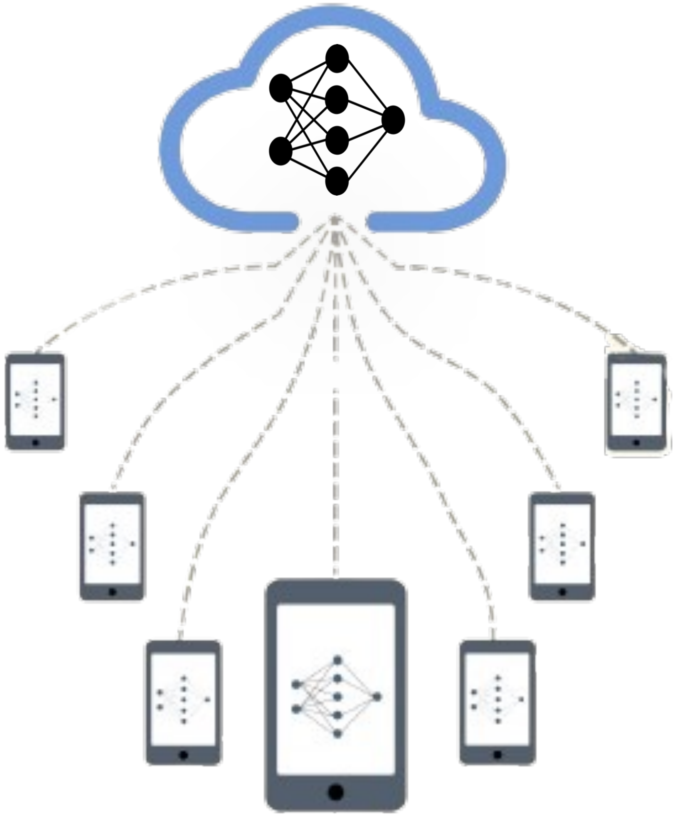
Key Components

- *Edge caching:* how to manage redundancy in EI application
- *Edge training: how to train EI application*
- *Edge inference:* how to infer intelligent application at the edge
- *Edge offloading:* how to sufficiently provide computing power to EI applications

Background of EI training scheme

❖ Collaborative Edge Training

Edge Training involves collaborative learning that put together multiple devices to train a shared model:



Federated Learning

Clients collaborate to train a model on a central server, while maintaining the decentralization of client data

↓

Key properties:

1. Non-IID
2. Unbalanced
3. Massively distributed
4. Limited communiaction

Federated Learning

Federated Averaging (FedAvg)

Server side: weights initialization:

$$w = \sum_{k=1} \frac{|D_k|}{\sum_k |D_k|} \cdot w_k$$

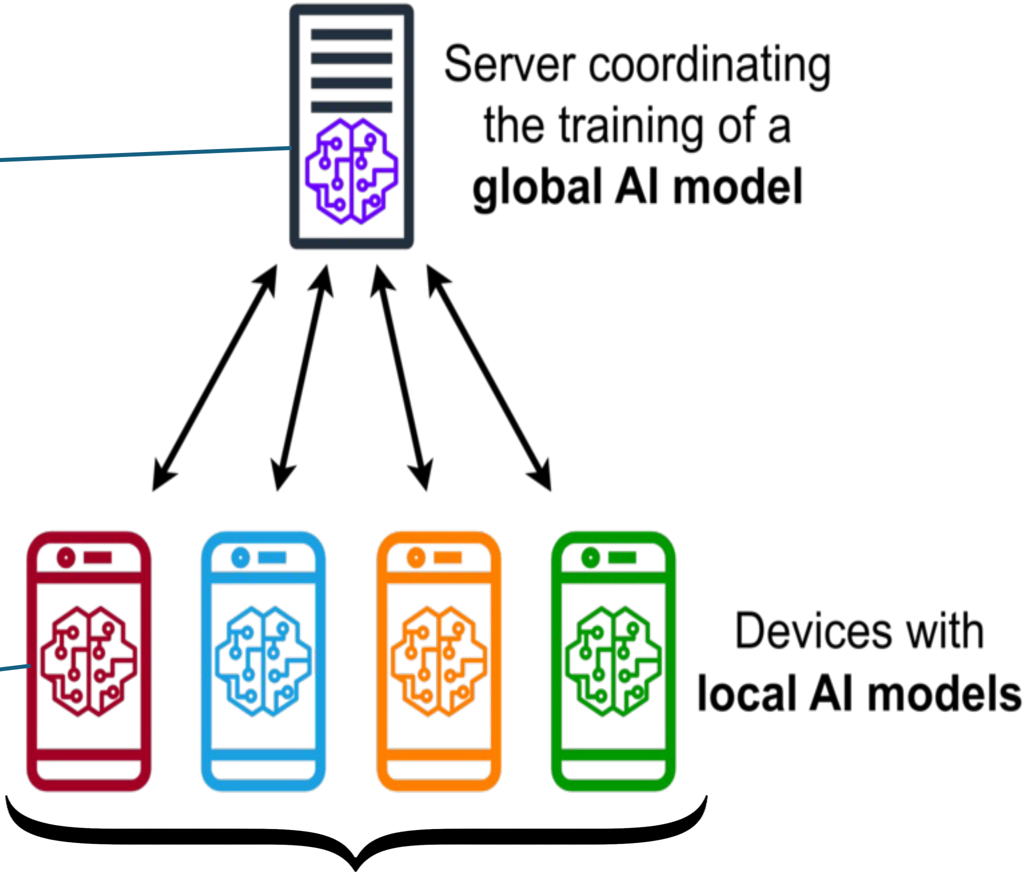
Communicate w

Client k:

$$w_k = w$$

$$w_k = w_k - \eta \frac{\partial \mathcal{L}_k(w_k)}{\partial w_k}$$

Communication w_k


Server coordinating the training of a **global AI model**
Devices with **local AI models**

Heterogenous device → differences in features, dimensions of collected data (multiple modalities)

Federated Learning

Extensions of Local Update Methods

- Drift due to local updates, and client sampling in cross-device FL
  - SCAFFOLD (Karimireddy et al., 2020)
  - FedProx (T. Li et al.,2020)

- Accelerate training through momentum

  - MIME (Karimireddy et al., 2021)

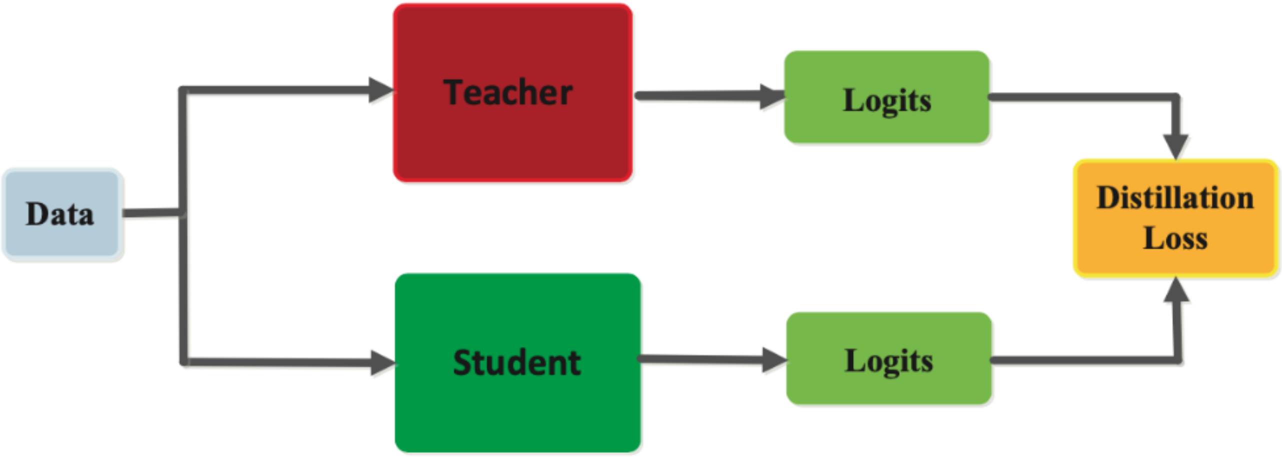
- Correct global model for clients performing different numbers of local updates
  - FedNova (wang et al., 2020)
  - FedLing (Mitra et al., 2021)
  - FedSuffle (Horvath et al., 2022)

CHALLENGE

- *System Heterogeneity :* the aggregation algorithm still struggle to solve model disparity

Knowledge Distillation

- Knowledge distillation → Teacher-Student Approach
- Method to extract knowledge from logits
- Used originally for model compression
- Allows the performance of the student close to that of the teacher
- In practice, it is used as compression method for resource constrained devices



Formulation

$$\mathcal{L}_{KD}(w) = \frac{1}{T}(p_i^S - p_i^T) = \frac{1}{T}\left(\frac{e^{z_i^S}/T}{\sum e^{z_i^S}/T} - \frac{e^{z_i^T}/T}{\sum e^{z_i^T}/T}\right)$$

$z_i^T, p_i^T \; and \; z_i^S, p_i^S$ are the logits the probabilities of the teacher and the students respectively.

Knowledge Distillation

Assuming that $z_i \ll T$:

$$\mathcal{L}_{KD}(w) = \frac{1}{T}\left(\frac{1+z_i^S/T}{n+\sum z_i^S/T} - \frac{1+z_i^{\mathcal{T}}/T}{n+\sum z_i^{\mathcal{T}}/T}\right)$$

and under the zero-mean hypothesis, i.e. $\sum_j z_i = 0$,

$$\mathcal{L}(w) = \frac{1}{NT^2}\left(z_i^S - z_i^{\mathcal{T}}\right)$$

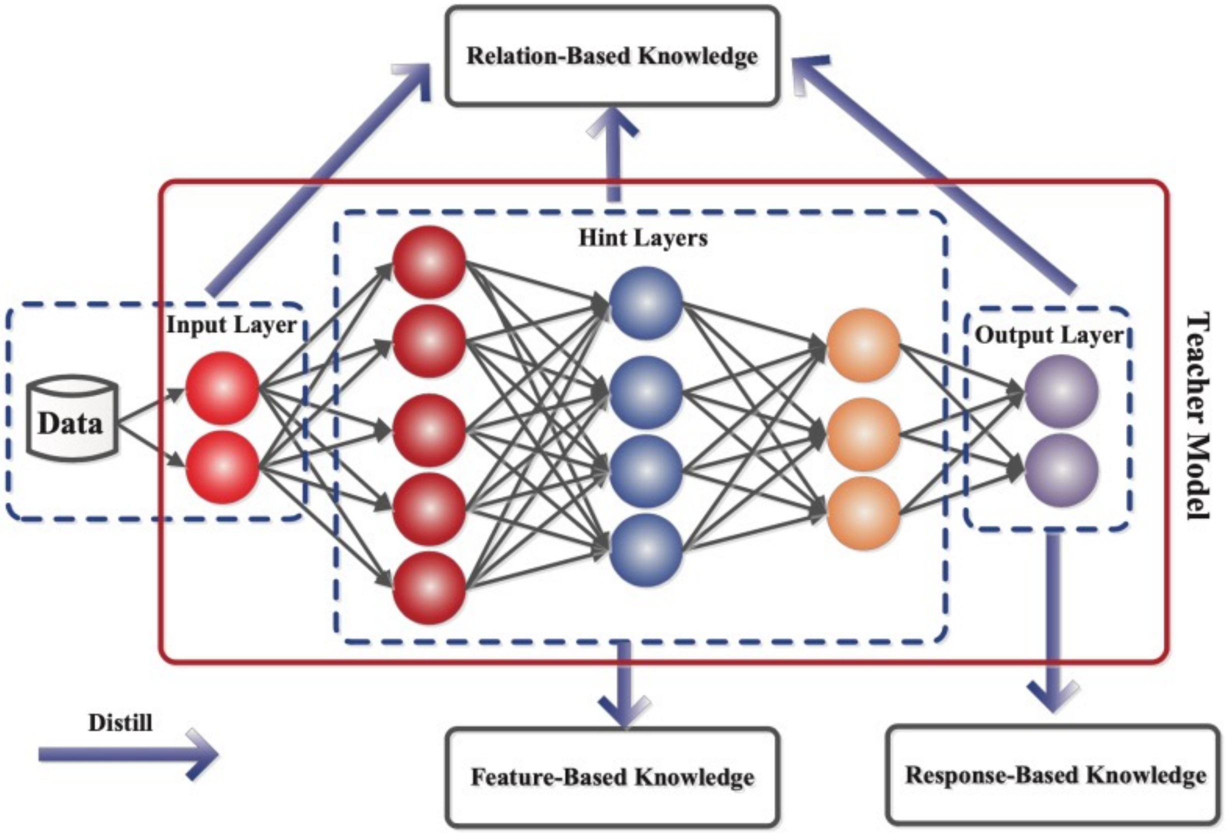
i.e. the loss is equivalent to matching the logits of the two models, as done in model compression

The overall loss is computed as the linear combination of the student loss and distillation loss:

$$\mathcal{L} = \alpha\mathcal{L}_{stu} + (1-\alpha)\mathcal{L}_{KD}$$

Types of Knowledge Distillation

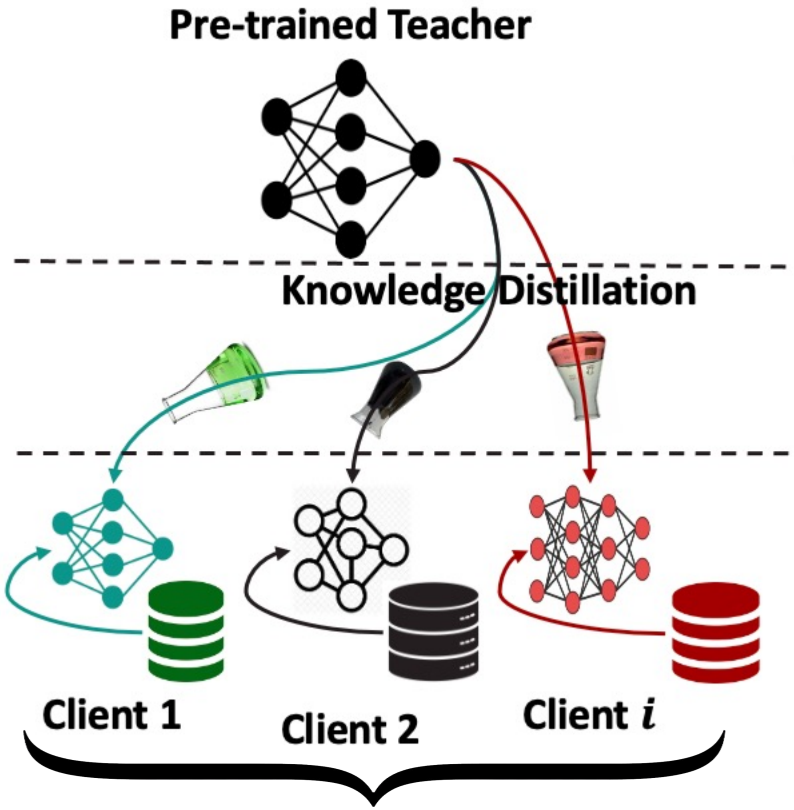
Typical Knowledge distillation uses logits (inputs to the final softmax)
Instead the probability produced by the softmax.



- Response-based knowledge,
- Feature-based knowledge,
- Relation-based knowledge

Knowledge Distillation for decentralized learning

Most research used in the literature

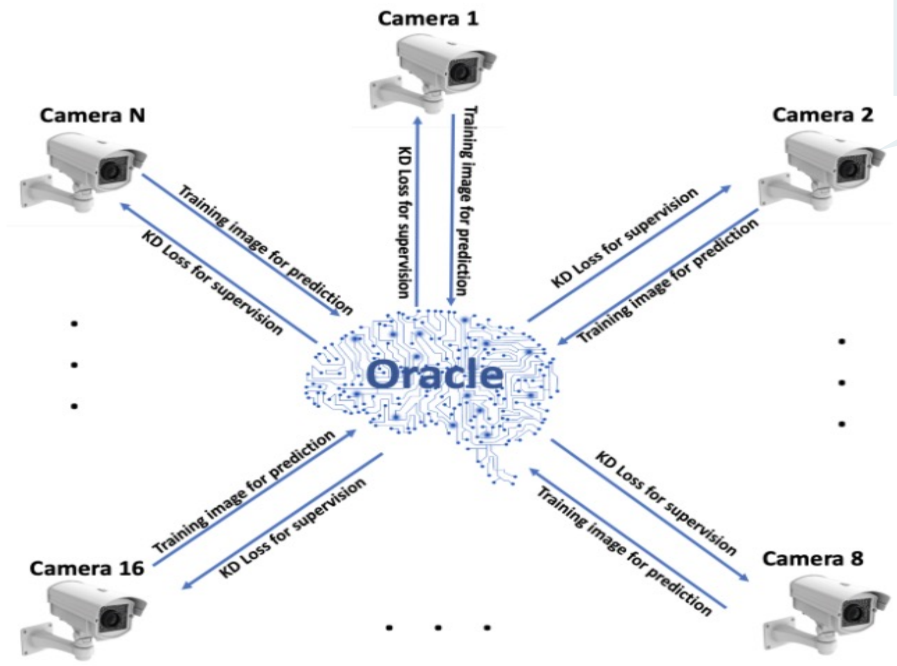


CHALLENGE

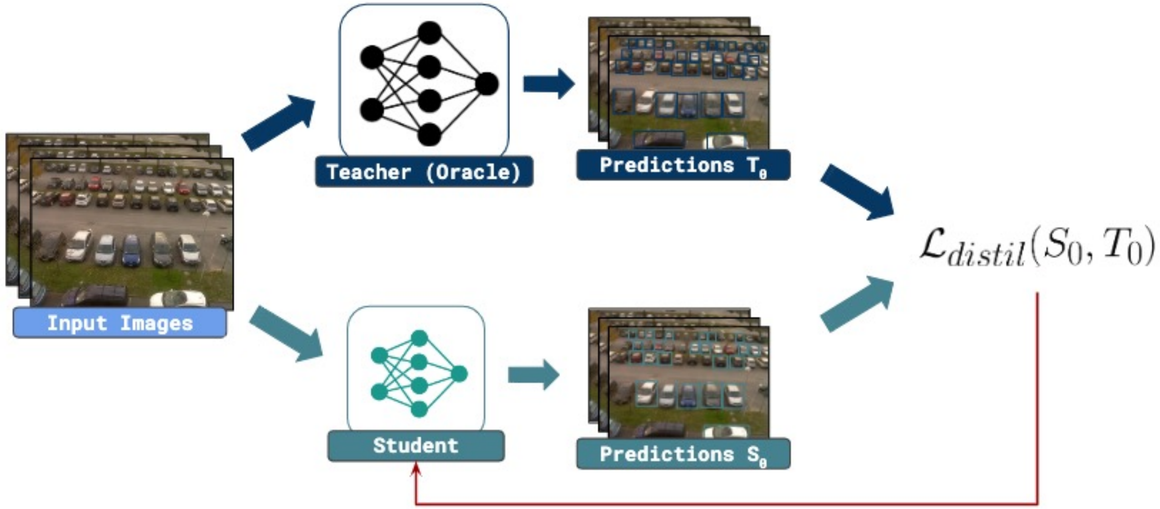
- Require transmission of data from the clients to the teacher
- The central server constitute a bottleneck when the number of devices increase

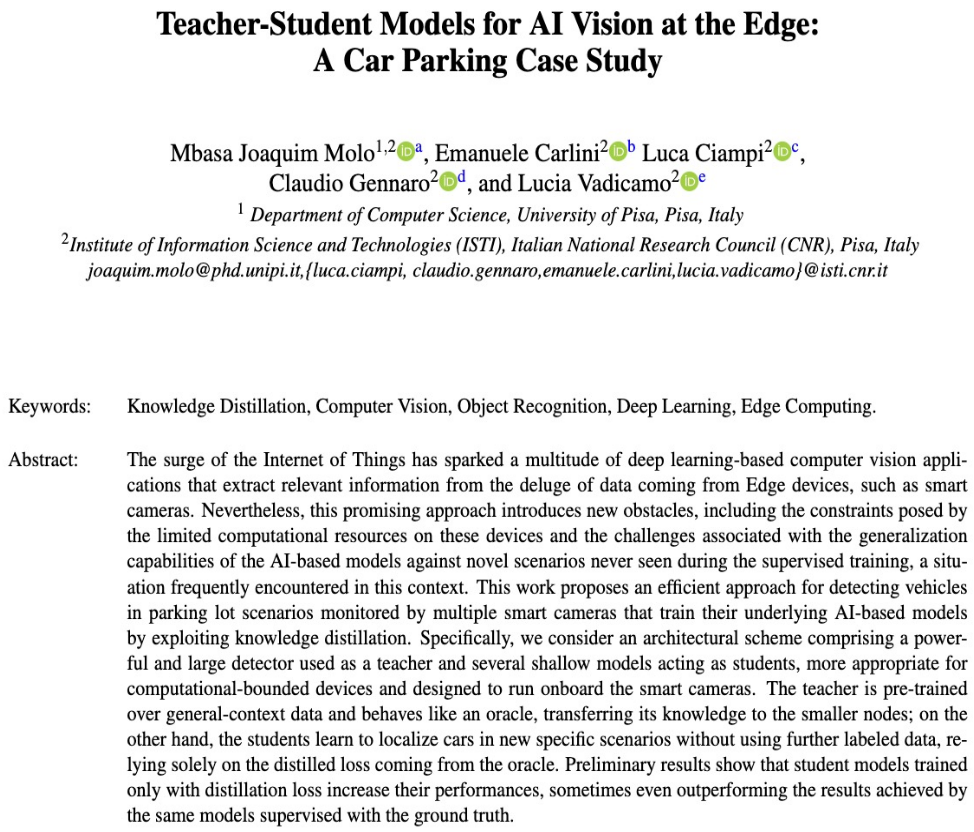
Heterogenous device → differences in features, dimensions of collected data (multiple modalities)

Use case



Lightweight model suitable for small cameras





Teacher-Student Models for AI Vision at the Edge: A Car Parking Case Study

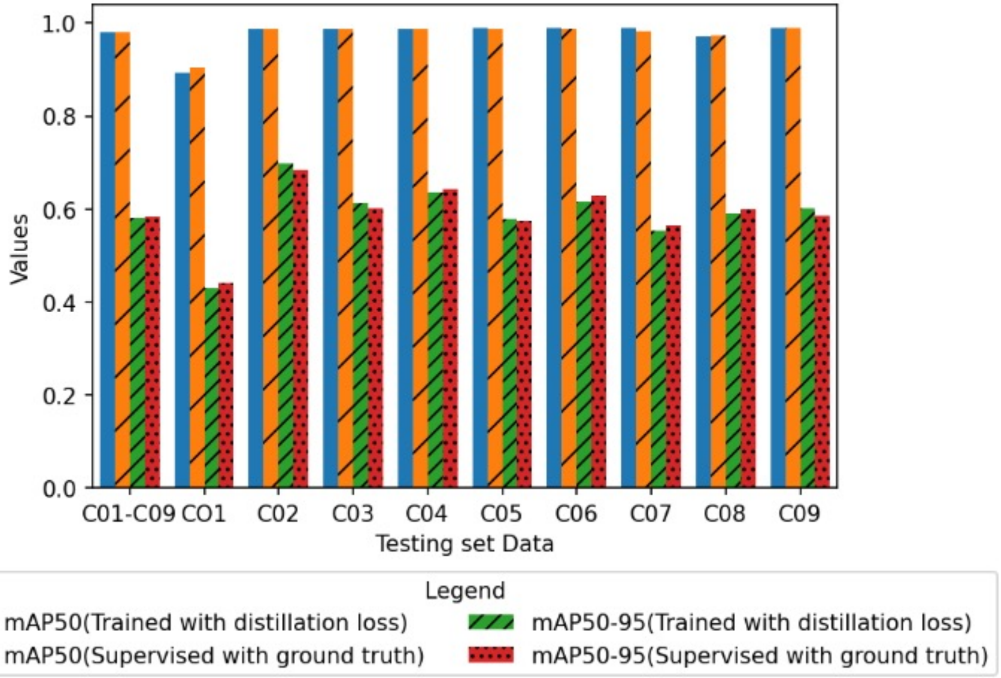
Mbasa Joaquim Molo[1,2][a], Emanuele Carlini[2][b] Luca Ciampi[2][c],
Claudio Gennaro[2][d], and Lucia Vadicamo[2][e]
[1] *Department of Computer Science, University of Pisa, Pisa, Italy*
[2] *Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), Pisa, Italy*
joaquim.molo@phd.unipi.it,{luca.ciampi, claudio.gennaro,emanuele.carlini,lucia.vadicamo}@isti.cnr.it

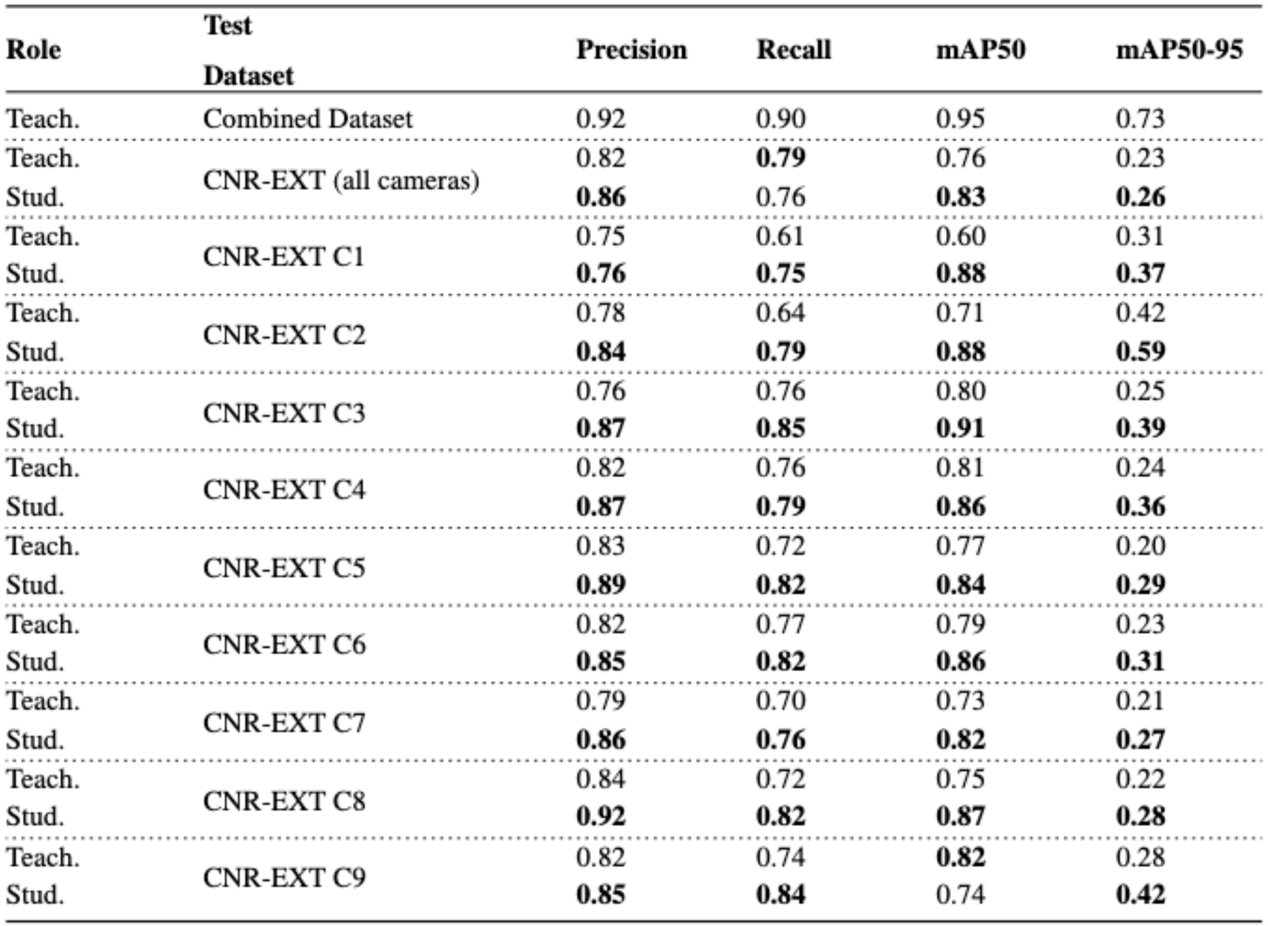
Keywords: Knowledge Distillation, Computer Vision, Object Recognition, Deep Learning, Edge Computing.

Abstract: The surge of the Internet of Things has sparked a multitude of deep learning-based computer vision applications that extract relevant information from the deluge of data coming from Edge devices, such as smart cameras. Nevertheless, this promising approach introduces new obstacles, including the constraints posed by the limited computational resources on these devices and the challenges associated with the generalization capabilities of the AI-based models against novel scenarios never seen during the supervised training, a situation frequently encountered in this context. This work proposes an efficient approach for detecting vehicles in parking lot scenarios monitored by multiple smart cameras that train their underlying AI-based models by exploiting knowledge distillation. Specifically, we consider an architectural scheme comprising a powerful and large detector used as a teacher and several shallow models acting as students, more appropriate for computational-bounded devices and designed to run onboard the smart cameras. The teacher is pre-trained over general-context data and behaves like an oracle, transferring its knowledge to the smaller nodes; on the other hand, the students learn to localize cars in new specific scenarios without using further labeled data, relying solely on the distilled loss coming from the oracle. Preliminary results show that student models trained only with distillation loss increase their performances, sometimes even outperforming the results achieved by the same models supervised with the ground truth.

Use case

Student performance is relatively the same wihen the learning is done with groubd truth or with the distillation loss





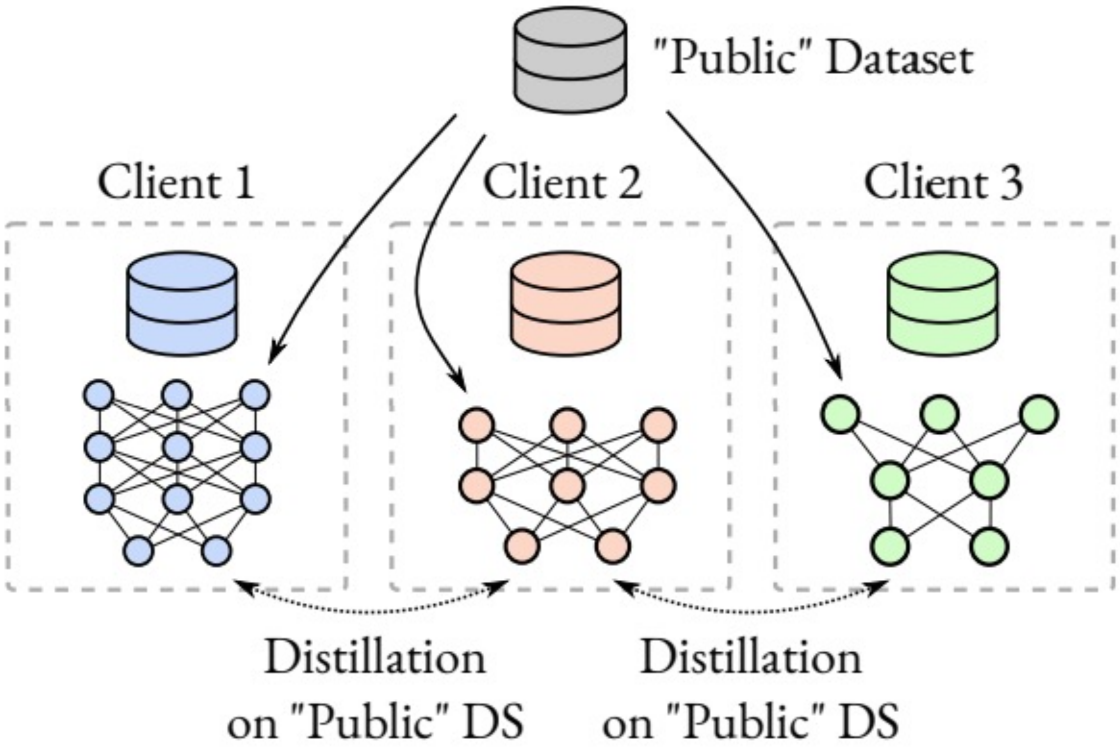
Role	Test Dataset	Precision	Recall	mAP50	mAP50-95
Teach.	Combined Dataset	0.92	0.90	0.95	0.73
Teach.	CNR-EXT (all cameras)	0.82	**0.79**	0.76	0.23
Stud.		**0.86**	0.76	**0.83**	**0.26**
Teach.	CNR-EXT C1	0.75	0.61	0.60	0.31
Stud.		**0.76**	**0.75**	**0.88**	**0.37**
Teach.	CNR-EXT C2	0.78	0.64	0.71	0.42
Stud.		**0.84**	**0.79**	**0.88**	**0.59**
Teach.	CNR-EXT C3	0.76	0.76	0.80	0.25
Stud.		**0.87**	**0.85**	**0.91**	**0.39**
Teach.	CNR-EXT C4	0.82	0.76	0.81	0.24
Stud.		**0.87**	**0.79**	**0.86**	**0.36**
Teach.	CNR-EXT C5	0.83	0.72	0.77	0.20
Stud.		**0.89**	**0.82**	**0.84**	**0.29**
Teach.	CNR-EXT C6	0.82	0.77	0.79	0.23
Stud.		**0.85**	**0.82**	**0.86**	**0.31**
Teach.	CNR-EXT C7	0.79	0.70	0.73	0.21
Stud.		**0.86**	**0.76**	**0.82**	**0.27**
Teach.	CNR-EXT C8	0.84	0.72	0.75	0.22
Stud.		**0.92**	**0.82**	**0.87**	**0.28**
Teach.	CNR-EXT C9	0.82	0.74	**0.82**	0.28
Stud.		**0.85**	**0.84**	0.74	**0.42**

Performance of the student with a teacher agnostic to the data distribution used for the during training

Knowledge Distillation for decentralized learning

Recently realised by Google AI

(Zhmoginov et al., 2023)



$$\mathcal{L}_k^t(w_k) = \mathcal{L}_k(w_k) + \sum_\beta \mathbb{E}_{x\sim\mathcal{D}_*}\mathcal{L}_{dist}^\beta\left(\psi_k^\beta(x), \phi_k^{\beta,t}\right)$$

$$\phi_k^{\beta,t}(x) = \left\{\phi_j^\beta(x) | j \in e_t(k)\right\}$$

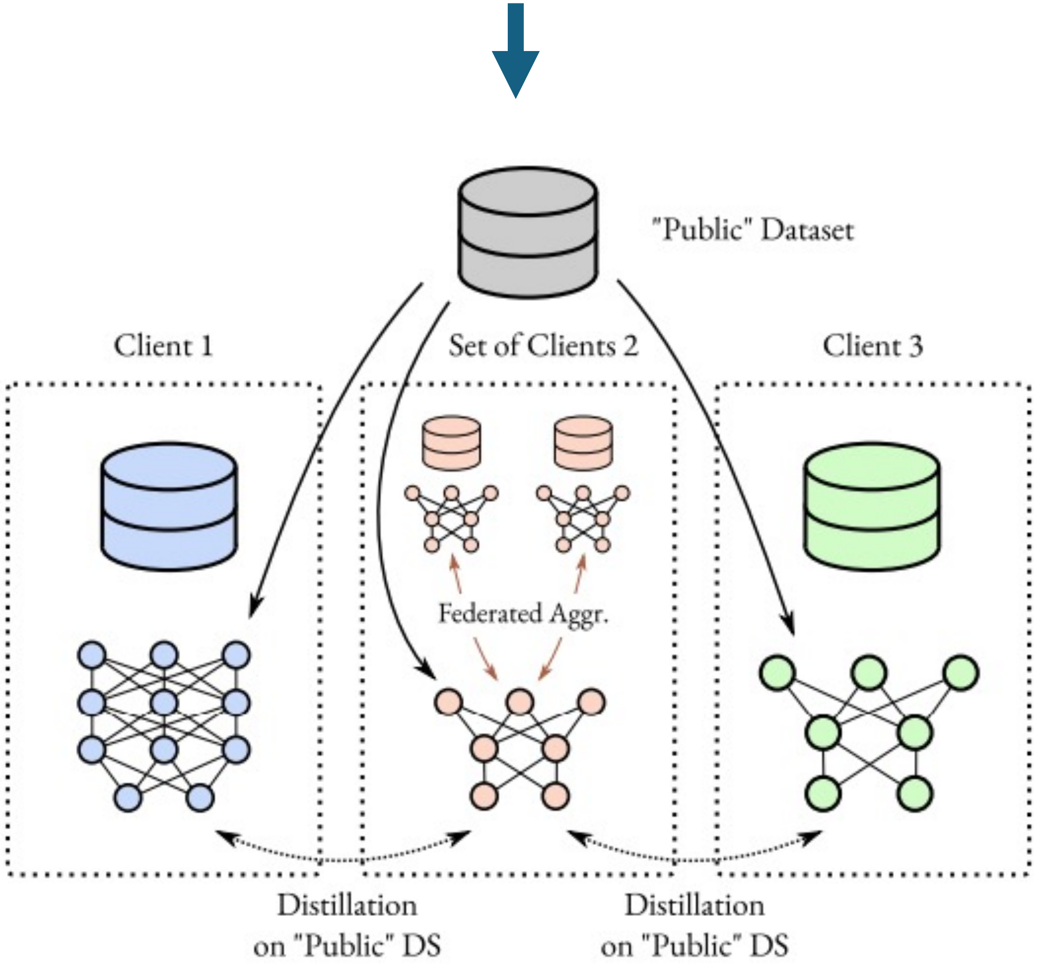
- Let $h_k(w_k)$ the main head of the model in client C_k
- $h_k^{aux}\left(w_k, \phi_j^\beta\right)$ the auxilary head

$$\mathcal{L}_{dist}^{aux}[h^{aux}, h] = -\nu_{aux}\sum_{j\in e_t(k)} h_j \log h_k^{aux}(x)$$

Knowledge Distillation for decentralized learning

Recently realised by Google AI

(Zhmoginov et al., 2023)



- Individual clients may have different architectures and different objective functions.
- Other clients may themselves be collections of models trained using federated learning

Open Challenge in Knowledge Distillation

1

Access to high-quality teacher

- Involves training a large pre-trained teacher on a large dataset, then use its predictions as soft targets to train the student.
- Obtaining high-quality teachers can be computationally expensive

2

Knowledge understanding

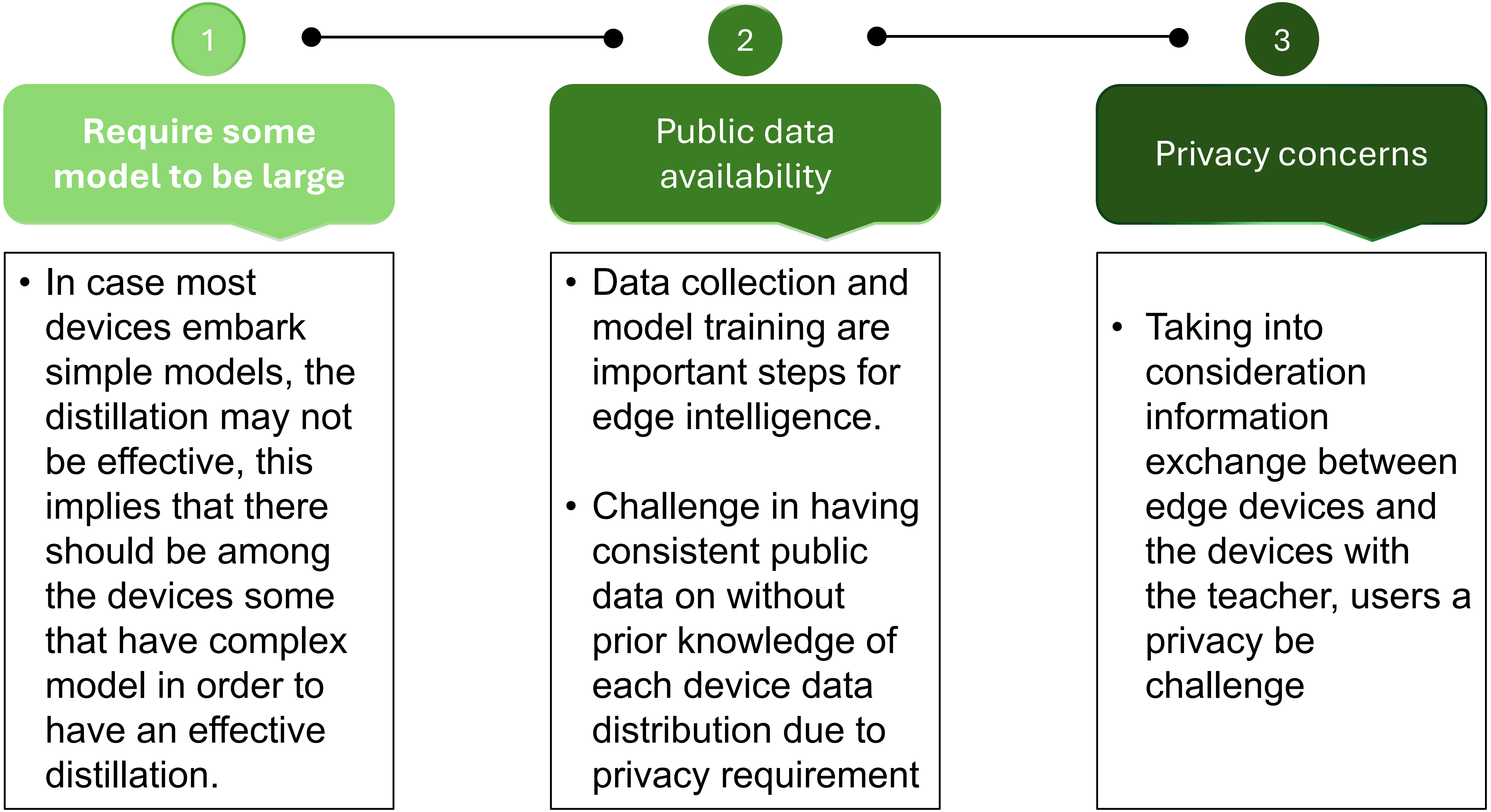
- Few empirical evaluations on the efficacy of KD have been performed
- Theoretical analysis of the quality of knowledge or the quality of the teacher-student architecture, is still difficult to attain

3

Hyper-parameter selection

- Selecting appropriate hyperparameters such as temperature scaling factor for soft target generation or balancing between mimicking complex decision boundaries and avoiding overfitting

Open Challenge in Decentralized Distillation



1

Require some model to be large

- In case most devices embark simple models, the distillation may not be effective, this implies that there should be among the devices some that have complex model in order to have an effective distillation.

2

Public data availability

- Data collection and model training are important steps for edge intelligence.
- Challenge in having consistent public data on without prior knowledge of each device data distribution due to privacy requirement

3

Privacy concerns

- Taking into consideration information exchange between edge devices and the devices with the teacher, users a privacy be challenge

Future directions

1

Decentralized KD expansion

Investigate KD for Edge AI beyond edge training by including edge offloading and edge caching

2

Model compression

Investigate how KD can be combined with other model compression approaches to have more effective lightweight models on portable platforms

3

Decentralized learning framework based on KD for multimodal data

Development of decentralized learning framework based on KD using multimodal data and trade-off investigation with FL