UNIVERSITÀ DEGLI STUDI DI MILANO

25-07-2024

DEPLOYMENT OF LLMs AT THE EDGE OF THE 6G NETWORK

Cloud Computing Technologies

PRESENTED BY

Alessandro Biagiotti

Introduction

- Key aspects of 6G networks
- Three problems
- Three solutions

4 key aspects

1 AI and machine learning

2 New spectrum technologies

3 Security and trust

4 New architectural models

[Back to list](#)

AI and machine learning

AI for the network vs AI for the user and LLM lifecycle

pre-training

The model undergoes a general training procedure

fine-tuning

The model is re-trained for alignement pourposes

inference

the trained model is put to use in unseen situations

Back to list

New spectrum technologies

Predicted spectrum usage of 5G vs 6G

5G	6G
High band (24 - 71 GHz)	Sub-THz band (> 92 GHz)
Medium band (2.6 - 4.9 GHz)	Local hotspot band (24 - 71 GHz)
Low band (600 - 2600 MHz)	Urban subsection band (7 - 20 GHz)
	Urban band (2.4 - 4.9 GHz)
	Wide area coverage band (600 - 2600 MHz)
	Extremely wide area coverage band (470 - 690 MHz)

Security and trust

Eavesdropping

When in the field of extremely high frequencies there's the risk of eavesdropping.

Model poisoning

Third parties might tamper with the training or fine-tuning datasets.

Personal privacy

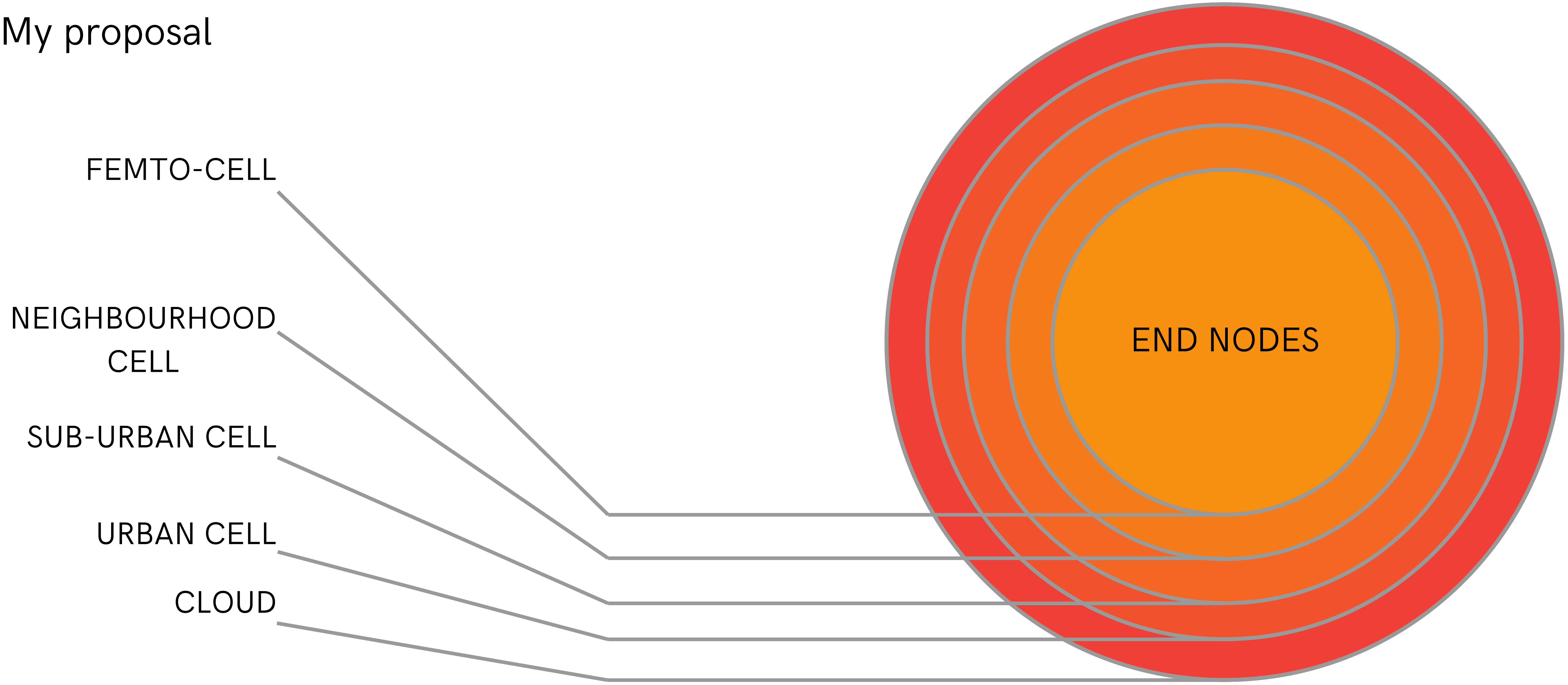
AI will be handling sensitive users' data (personal health, personal finance, private life information etc...).

Ongoing research is pushing to have a Reinforcement Learning model as a security system that can deploy countermeasures based on real time threat analysis

Back to list

New Architectural models

My proposal



There is multiple of these cells and for each cell there is a constellation of powerful edge nodes that can handle more demanding computations as well as the network infrastructure management.

Back to list

Some numbers

Two numbers for scale

350

GigaBytes

VRAM Required to load in memory the GPT-3 175B model with 16b floats

665

Years

The amount of time required to complete a single training run for GPT-3 175B on an RTX 8000

Three problems

Size and compute

LLMs are too big to be stored on end or low-order edge nodes.

Latency

We cannot do inference in cloud anymore, we are latency-bound.

Privacy

Training and fine-tuning have to be operated in an extremely secure environment.

Three solutions

Federated Learning

Each device fine-tunes its own version of the model.

Split Learning

The model is split horizontally and then a certain number of devices handle the resulting slices.

LoRA

Low-rank matrix injection for model compression.

Federated Learning (FL)

Each device trains a **clone** of the model stored in a head node with **local data**.

The result of the training is later encrypted and sent upstream to the head node that **cumulates the effect** of the training in some way (e.g. averaging).

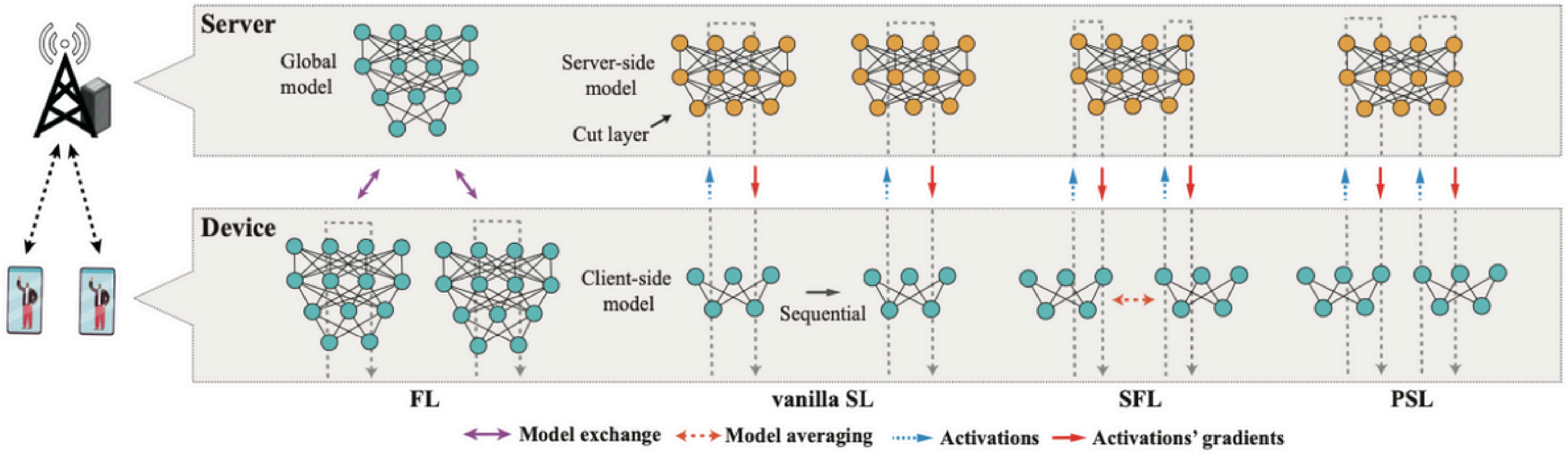
While this approach maintains a **privacy** focus it cannot be used on end devices to train LLMs because end devices lack **computational power**.

On the other hand I can see FL deployed **close to the upper-end** of the network stack to share the weight of pre-training bigger models.

Split Learning (SL)

The original model is **split** horizontally in one or more slices that are then divided between nodes based on their computational power.

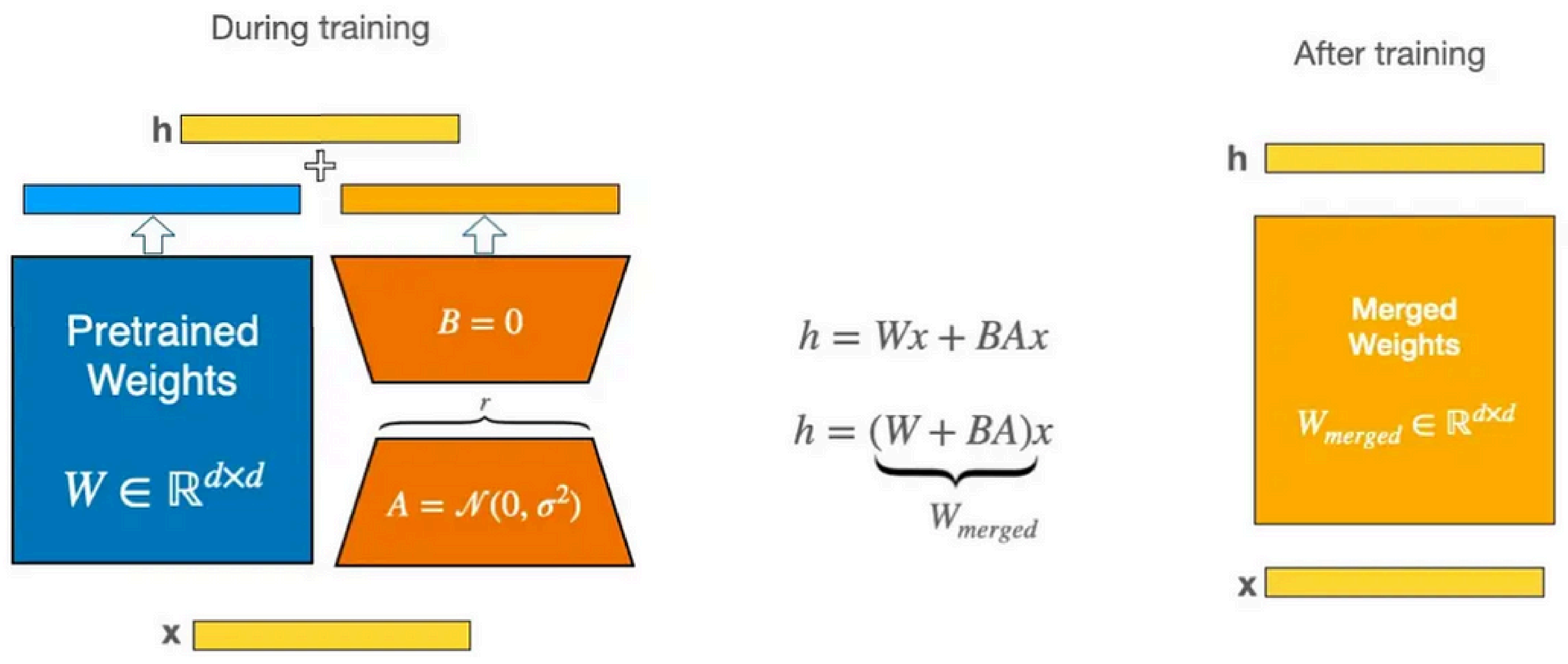
SL addresses the **size** and the **privacy** constraints but there are some latency complications when a user is in a mobilitiy phase.



Reinforcement learning algorithms are being studied to find the optimal splitting point(s) for the model depending on a series of environmental conditions as well as characteristics of the model itself.

Low Rank Adaptation

When performing fine-tuning, instead of using the full-size weight-update matrix we can store its **Low Rank representation** for the more expensive layers (like the fully connected layers).



The pre-trained wights are frozen, only the low rank representation undergoes the training procedure and then after training the two matrices are summed together.

Thank you for your attention