Edge machine learning and signal processing for wireless networked systems

Francesco Binucci

Introduction

Our research focuses on **Edge Machine Learning** (EML) within 5G/6G networks.

- New generation mobile networks support diverse applications, each with specific Key Performance Indicators (KPIs).
- Many of them are based on Artificial Intelligence and Machine Learning.
- EML allows the deployment of ML/AI services with **low latency** and **low energetic consumption** enabling the User Equipments to offload the data towards the network edge.

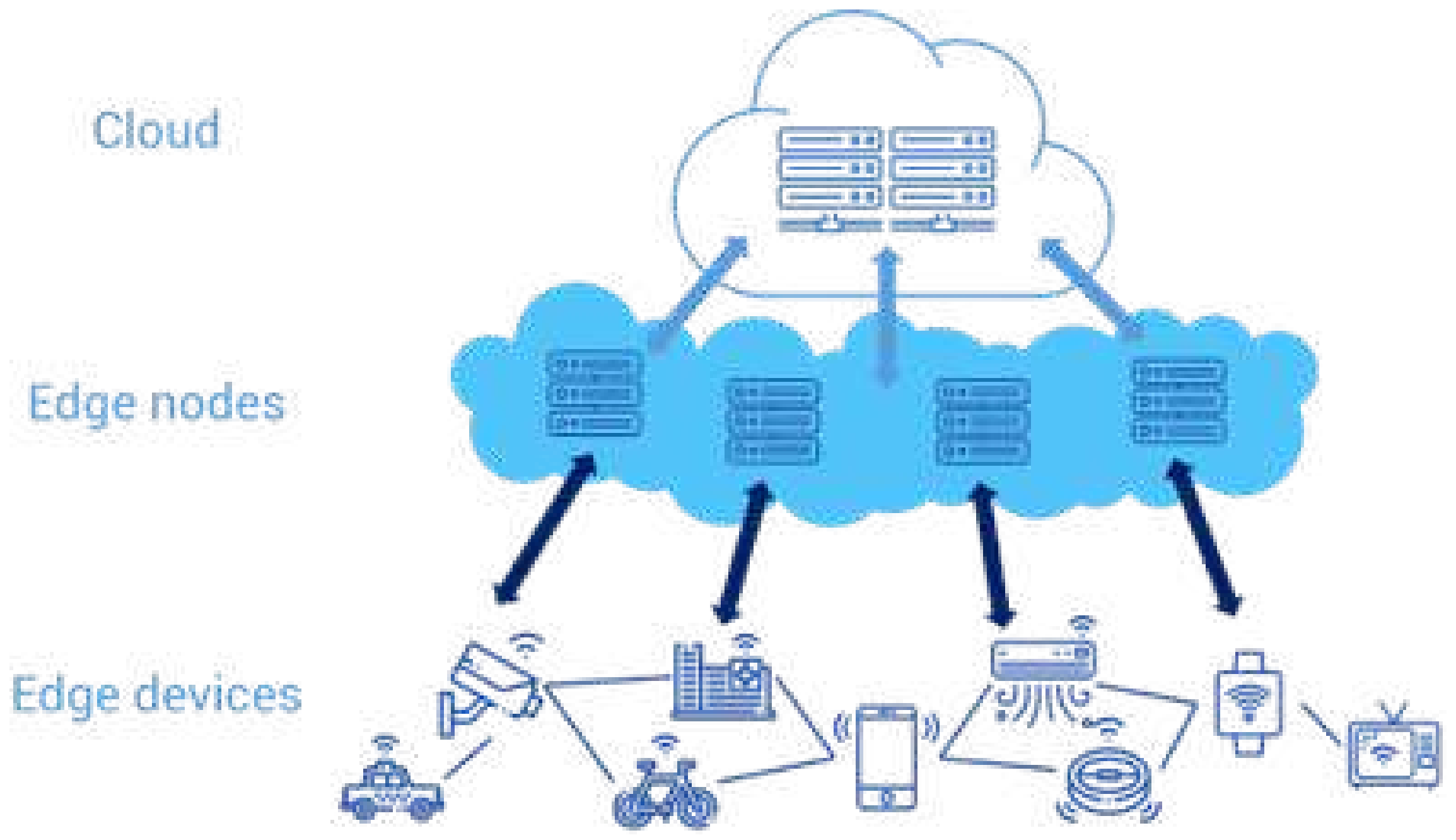

Figure 1. Schematic representation of a new generation mobile network, with edge and cloud networks. Credits: https://www.alibabacloud.com/fr/knowledge/what-is-edge-computing.

Goal-Oriented communications [1] represents a novel communication paradigm that act as a key-enabler for Edge Machine Learning tasks.

Specifically, GOCs have the following appealing features:

- They focus on transmitting only essential information for specific tasks.
- They allows to save as much transmission resoures as possible while guaranteeing a prescribed level of learning performance.

Goal-Oriented Communications can be formalized throgh the Information Bottleneck (IB) principle [2], a theoretical framework based on rate/distortion theory arguments which aims to find a compact representation $\mathbf{Z}$ of a signal $\mathbf{X}$ which is as much *informative* as possible with respect to the outcome of a specific inference task $\mathbf{Y}$.

Main Scientific Contributions

Our main scientific contribution are related to the development of resource allocation strategies to support edge-assisted Goal-Oriented communications focusing on the best trade-off among **latency**, **energy** and **learning performance**.

- Optimal resource allocation for digital goal-oriented compression framework based on auto-encoders and JPEG encoding. [3, 4]
- Extension to **multi-carrier** transmission and digital transmission affected by a non null bit-error-rate [5, 6].
- Development of the **Opportunistic Information Bottleneck**, a Goal-Oriented compression scheme based on the optimal solution of the Gaussian Information Bottleneck [7].
- Optimal resource allocation for Goal-Oriented DNN splitting [8].

Scientific Results

The works [3, 4, 5, 6] investigate optimal resource allocations strategies for edge-assisted GOCs and they highlighted the benefit of GOCs in reaching the best trade-offs between **energy**, **latency** and **learning** performance considering both single/multi carrier transmission schemes and also taking into account noisy communications.

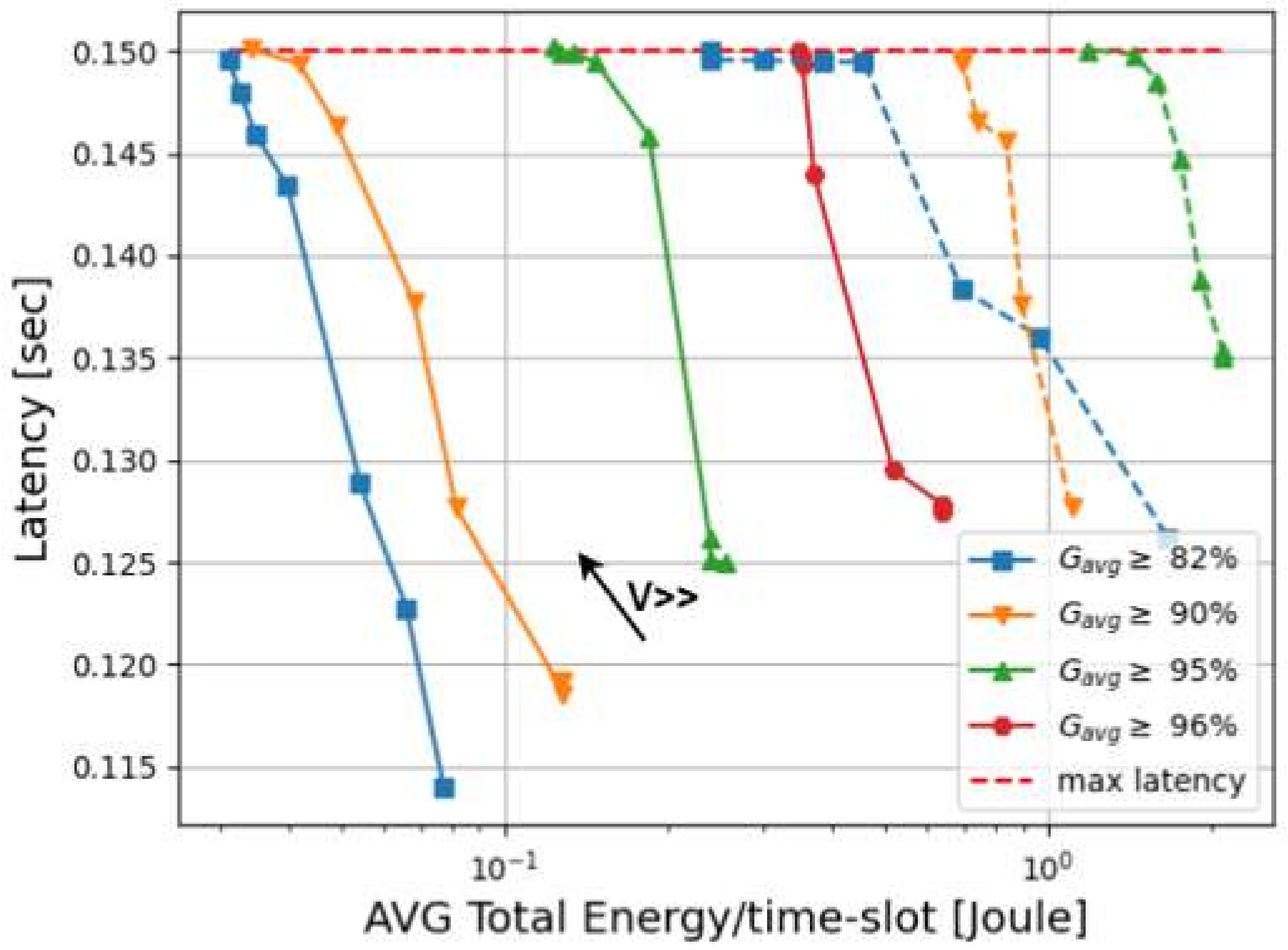

Figure 2. Energy/Accuracy trade-off comparisons considering an Auto-Encoder based GOC scheme (solid) and a compression scheme based on downsampling with anti-aliasing pre-filtering (dashed).

Our **Opportunistic Information Bottleneck** framework, presented in [7], exploits the closed-form solution of the Gaussian Information Bottleneck [9] to solve an *opportunistic* regression sub-task between a Gaussian Transformation $h()$ of the input data and the output of the first linear layer $\mathbf{L_0}$ of a Deep-Neural network used to solve a general inference task (e.g., image classification). This allows to deploy a **theoretical principled** and **low-complexity** GOC scheme with a better compression/accuracy trade-offs with respect to the competitors.

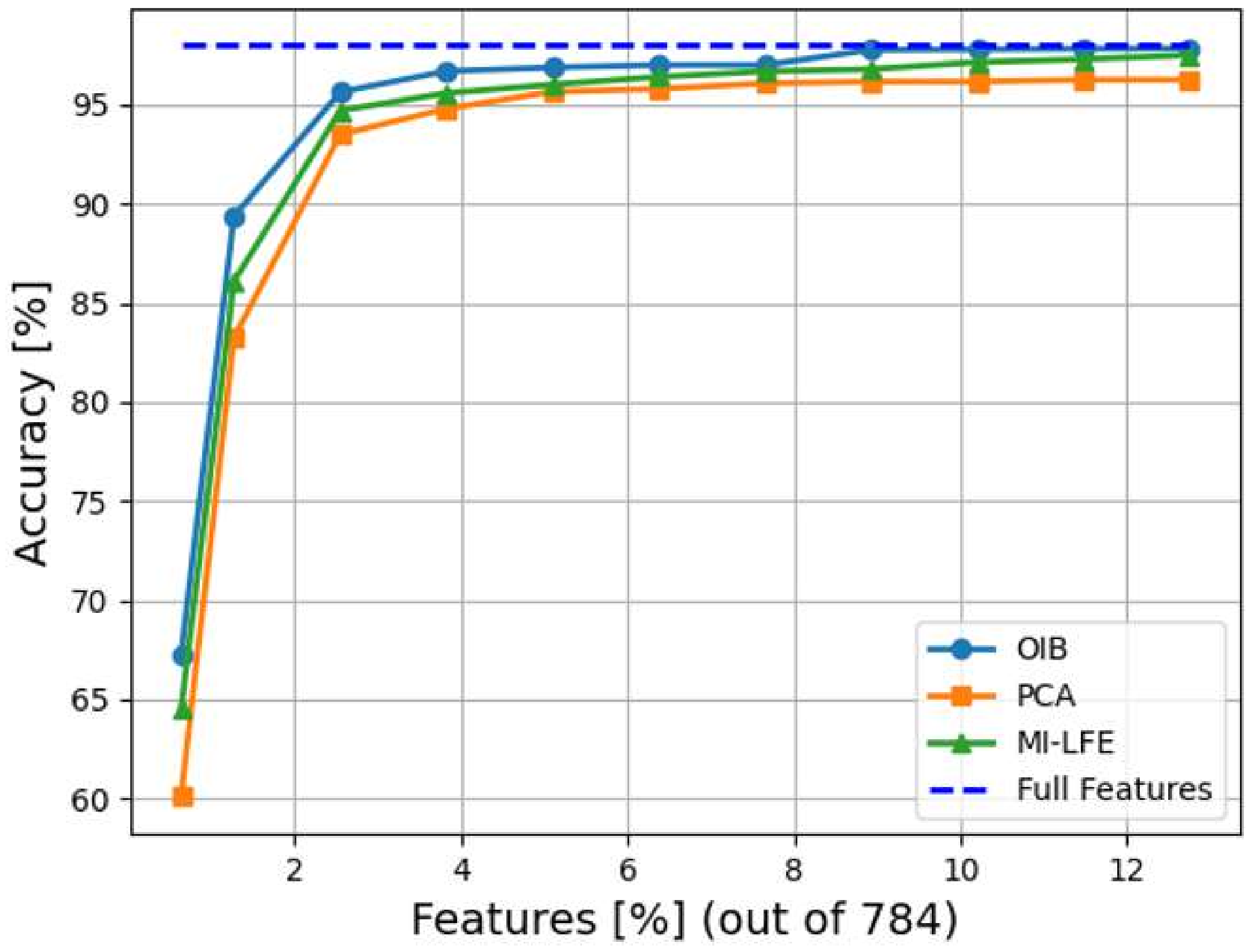

Figure 3. Comparison of the proposed opportunistic Information Bottleneck with Principal Component Analysis and a feature extraction algorithm based on Mutual Information Maximization.

Scientific Results (contd.)

In [8] we developed an optimal resource allocation strategy for goal-oriented **deep neural network splitting**. The main contributions are

- Analysis of the accuracy degradation in different SNR regimes as a function of the splitting-point.
- Dynamic optimization of the computational and transmission resources as well as the splitting point selection taking into account the accuracy degradation due to the noise effect.

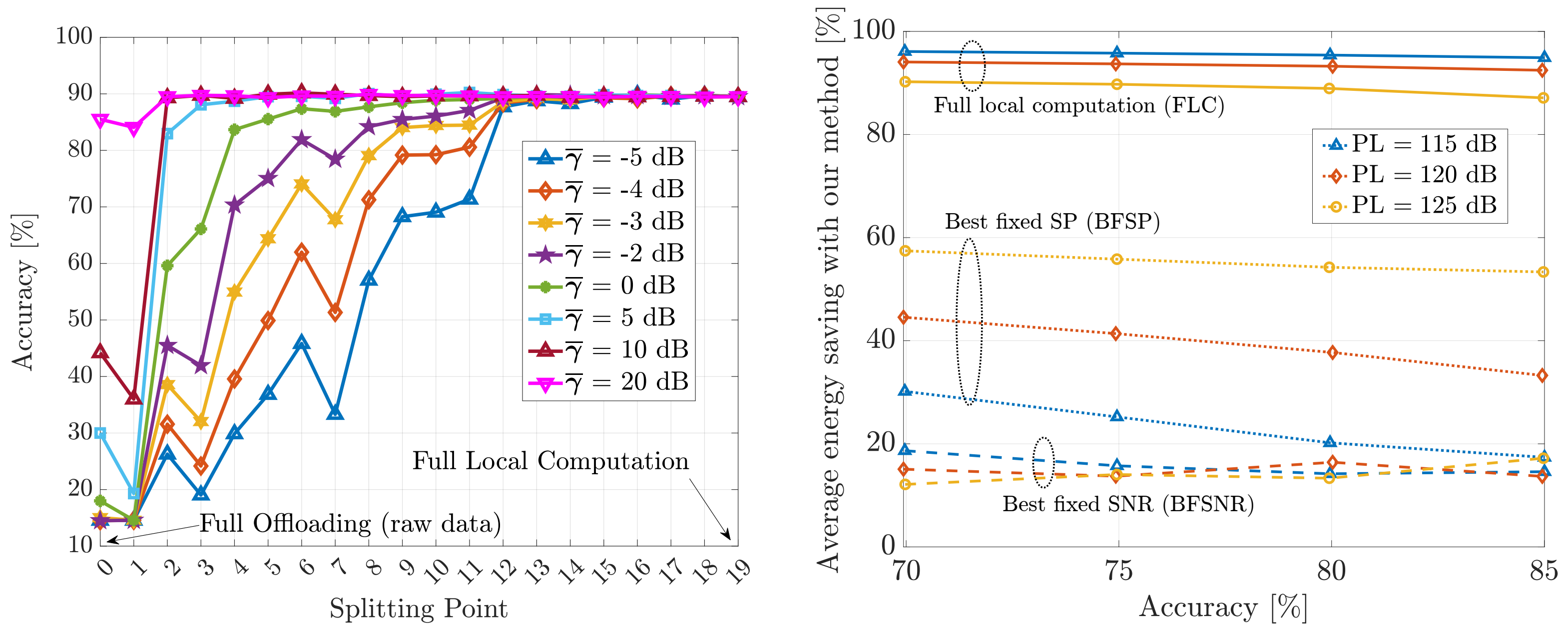

Figure 4. (a) shows the accuracy degradation as a function of the splitting point for different SNRs. (b) shows the benefits of the proposed dynamic SP selection strategy with respect to the competitors.

Ongoing Research and future work

- Development of resource allocation strategies to ensure **average** and **strict** reliability guarantees using Adaptive Conformal Prediction [10].
- Optimal Resource allocation for Edge-Assisted Opportunistic Information Bottleneck.
- Investigation of EML and GOCs in Federated Learning scenarios.

References

[1] E. C. Strinati and S. Barbarossa, “6g networks: Beyond shannon towards semantic and goal-oriented communications,” *Computer Networks*, vol. 190, p. 107930, 2021.

[2] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.

[3] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, “Adaptive resource optimization for edge inference with goal-oriented communications,” *EURASIP Journal on Advances in Signal Processing*, vol. 2022, no. 1, p. 123, 2022.

[4] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, “Multi-user goal-oriented communications with energy-efficient edge resource management,” *IEEE Transactions on Green Communications and Networking*, 2023.

[5] F. Binucci and P. Banelli, “Ber-aware dynamic resource management for edge-assisted goal-oriented communications,” in *ICASSP 2023 - 2023 IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, pp. 1–5, 2023.

[6] F. Binucci and P. Banelli, “Goal-oriented water-filling for dynamic management of edge-assisted ofdm communications,” in *ICC 2023 - IEEE International Conference on Communications*, pp. 5761–5766, 2023.

[7] F. Binucci, P. Banelli, P. Di Lorenzo, and S. Barbarossa, “Opportunistic information-bottleneck for goal-oriented feature extraction and communication,” *IEEE Open Jour. of the Commu. Soc.*, vol. 5, pp. 2418–2432, 2024.

[8] F. Binucci, M. Merluzzi, P. Banelli, E. C. Strinati, and P. Di Lorenzo, “Enabling edge artificial intelligence via goal-oriented deep neural network splitting,” *arXiv preprint arXiv:2312.03555*, 2023.

[9] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for gaussian variables,” *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[10] I. Gibbs and E. Candes, “Adaptive conformal inference under distribution shift,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 1660–1672, 2021.

francesco.binucci@studenti.unipg.it