Energy efficient deployment of applications in the edge-network-cloud continuum

Contract: PhD grant - Fixed-term contract - 3 years
Financed by the *PEPR CareCLOUD*

Starting date: Can start early 2024
Location: Center Inria of Université Côte d'Azur
Address: 2004 Route des Lucioles - BP 93 06902 Sophia Antipolis cedex France

Supervision: Frédéric Giroire (DR CNRS, Coati team),
Contact: frederic.giroire@inria.fr

Research program

Today, with the deployment of a new generation of networks, such as 5G, the future 6G and the Internet of Things, the traditional paradigms for deploying many cloud and network services have changed profoundly. Indeed, the development of hardware capabilities of devices on the edge and in the network on the one hand, and the new network constraints (e.g. latency and delay) of new applications (such as the automated car, telemedicine) on the other, creates a new scenario in which services are deployed along the entire edge-network-cloud continuum [1]. An application can therefore be launched either on the edge (in a cell phone, for example), in the network (on an antenna, for example) or in the cloud.

At the same time, we are witnessing a massive development of artificial intelligence on the edges of the network [2], with, for example, the automatic completion and correction of messages, or the analysis of photos or videos. A new field of research is focusing on the deployment of (deep) learning models on network edge equipment, such as gateways, microcontrollers, antennas or cell phones [3]. These devices present strong constraints in terms of disk space, computing power and memory, which limit the latency and accuracy of these models [4].

It is therefore necessary to find new methods for launching complex neural networks in equipment with high constraints [5]. One solution is to design lightweight models that can be used in specific devices, such as MobileNet [6] for telephones. Another solution is to use model compression techniques. The principle is to reduce the size of an efficient neural network so that it uses less memory and storage, while minimizing its accuracy [7]. Recent work [8,9] has proposed adjustable deep neural networks that can be configured during the inference phase without the need for retraining.

The aim of this thesis is to study how to effectively use these new model compression techniques when deploying applications in the edge-network-cloud continuum, with the aim of reducing cloud and network energy consumption. Indeed, the energy efficiency of an adjustable model is very good at the beginning of its execution, then decreases very sharply.

Summary of the project into which the research is carried out

The thesis will be carried out as part of the PEPR CLOUD project CARECloud (Comprendre, Améliorer, Réduire les impacts Environnementaux du Cloud computing). Cloud computing and its many variations offer users considerable computing and storage capacities. The maturity of virtualization techniques has enabled the emergence of complex virtualized infrastructures, capable of rapidly deploying and reconfiguring virtual and elastic resources, in increasingly distributed infrastructures. This transparent resource management gives users the illusion of access to flexible, unlimited and virtually immaterial resources. However, the power consumption of these clouds is very real and a cause for concern, as are their overall greenhouse gas (GHG) emissions and the consumption of critical raw materials used in their manufacture. At a time when climate change is becoming more visible and impressive every year, with serious consequences for people and the planet on a global scale, all sectors (transport, construction, agriculture, industry, etc.) must contribute to the effort to reduce GHG emissions. Clouds, despite their ability to optimize processes in other sectors (transport, energy, agriculture), are no exception to this observation: the increasing slope of their greenhouse gas emissions must be reversed, or their potential benefits in other sectors will be wiped out. This is why the CARECloud project aims to drastically reduce the environmental impact of cloud infrastructures.

Activities

1 Propose new algorithmic and optimization methods for scheduling machine learning applications in the cloud.
2 Investigate the trade-off between energy efficiency and accuracy of neural network compression techniques.
3 Reduce cloud energy consumption by using new methods to place, schedule and compress machine learning applications.

Skills to be acquired (2-3 lines)

1 Algorithms for scheduling applications in the cloud.
2 Optimization techniques for cloud and network resource placement.
3 Neural network compression techniques.

References

[1] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, "Edge comput- ing with artificial intelligence: A machine learning perspective," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource- constrained distributed machine learning," in *IEEE INFOCOM 2018- IEEE conference on computer communications*. IEEE, 2018, pp. 63– 71.

[3] G. Drainakis, P. Pantazopoulos, K. V. Katsaros, V. Sourlas, and A. Amdi- tis, "On the distribution of ml workloads to the network edge and beyond," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2021, pp. 1–6.

[4] W. Gao, Q. Hu, Z. Ye, P. Sun, X. Wang, Y. Luo, T. Zhang, and Y. Wen, "Deep learning workload scheduling in gpu datacenters: Taxonomy, challenges and vision," *arXiv preprint arXiv:2205.11913*, 2022.

[5] J. Lin, W.-M. Chen, J. Cohn, C. Gan, and S. Han, "Mcunet: Tiny deep learning on iot devices," in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[7] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126– 136, 2018.

[8] J. Yu and T. Huang, "Autoslim: Towards one-shot architecture search for channel numbers," 2019. [Online]. Available: https://arxiv.org/abs/1903.11728

[9] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-all: Train one network and specialize it for efficient deployment," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HylxE1HKwS