M2 internship subject:
Optimal quantization of ReLU neural networks

Supervisors: Elisa Riccietti (MCF ENSL), Rémi Gribonval (DR Inria),
Pascal Carrivain (Ingénieur de Recherche, Inria) – Ockham team
Theo Mary (CR CNRS, Sorbonne)

September 2024

Context: inducing parsimony in machine learning The last years have witnessed an impressive success of artificial intelligence (AI) that has lead to the development of powerful tools such as convolutional neural networks, transformers, and generative models, which ensured great achievements in several domains, such as speech recognition, recommender systems, or object detection. Their success is explained by mainly two factors: the use of large models and large datasets. However, the training and deployment of such models leads to enormous energy consumption and carbon emission. The cost of annotating, maintaining and storing large datasets is also important. All of this undermines a democratic access to deep learning, hinders its use in contexts in which resources are limited, and has a negative impact on the environment. It is thus of crucial importance to make deep learning more parsimonious, by reducing its data-hungry nature and its dependence on large deep neural networks.

Among the most used techniques to achieve this goal are quantization and sparsification. Many approaches have been proposed in the literature, but the majority of them rely on heuristics and the question of optimality has not been addressed. In this project we will mainly focus on quantization.

Project description The aim of the Ockham team is to *develop principled approaches to neural networks quantization by exploiting scaling invariances of the problem.* In particular, in this internship we will make a first step towards this ambitious objective by considering simple neural networks with a single neuron.

Problems involving ReLU neural networks show important scaling invariances. Specifically, the output of these networks is invariant with respect to specific rescaling of the parameters. Because the ReLU function is positively homogeneous, multiplying by a scalar λ both the weight of the incoming edge of a neuron and its bias, and by $1/\lambda$ the weight of the outcoming edge, the realization of the network does not change.

A similar property is leveraged in [2, 3] to find optimal quantizations of rank-one matrices. This principled approach has been exploited to propose an efficient heuristic quantization schemes for butterfly matrices, which can be seen as a special (linear) case of neural network.

An exciting challenge is thus to extend this principled approach to a more general nonlinear setting, involving ReLU neural networks. This will involve in priority:

- to propose a heuristic quantization strategy for the nonlinear case, starting from the simple case of a single neuron,
- establish optimality measures adapted to the nonlinear context,
- investigate if it is possible to prove some optimality result in this context,
- studying possible extensions to more general architectures [1]: single layer networks with more neurons, more layers...

- implementing, testing, documenting and illustrating with examples the proposed heuristics, and integrating it in `pyfaust` [4] .

References

[1] A. Gonon, N. Brisebarre, E. Riccietti, and R. Gribonval. Path-metrics, pruning and generalization. Preprint, Under review, 2024.

[2] R. Gribonval, T. Mary, and E. Riccietti. Optimal quantization of rank-one matrices in floating-point arithmetic—with applications to butterfly factorizations. Preprint, Under review, June 2023.

[3] R. Gribonval, T. Mary, and E. Riccietti. Scaling is all you need: quantization of butterfly matrix products via optimal rank-one quantization. In *29ème Colloque sur le traitement du signal et des images (GRETSI)*, number 2023-1193 in Actes du GRETSI 2023, pages 497–500, Grenoble, France, August 2023. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

[4] Ockham team. `pyfaust` python package.
`https://faust.inria.fr`.