Scheduling precision for quantized training

Master thesis
OMLS Group

Low-precision or quantized training can improve the computational efficiency of training deep neural networks (DNN). In addition, this automatically enables the quantization of the model at inference time, allowing them to be deployed on resource-constrained devices.
While the precision for the weights and activations is typically chosen a priori and remains fixed throughout training, recent work [Dremov et al., 2025][Wolfe et al., 2024] suggest that changing the precision during training is in fact beneficial.

Given a fixed memory and compute budget, the question arises how to allocate the budget optimally throughout training. For fixed precision formats, Pareto-optimal quantization has been studied for ResNets [Abdolrashidi et al., 2021] and MobileNets [De Putter et al., 2025], but less so in more modern model architectures, such as transformers or state-space models.

The goal of this project is to further explore the potential of scheduling or adaptively adjusting the precision over the course of training. This involves studying the trade-off between quantizing weights and activations and exploring critical learning periods [Achille et al., 2018] for quantization.

Prerequisites

- Experience in Linear Algebra, Calculus
- Experience with optimization algorithms
- Preferably a good grade for Continuous Optimization and Foundations of Deep Learning
- Experience in using Python libraries, such as Torch, particularly in implementing the model training.

Key research questions

- Does the pareto-optimal quantization curve differ between transformers and ResNets and convolutional networks?
- Can one improve upon the pareto-optimal quantization frontier by scheduling the precision of weights and activations?
- Are there periods in training, which are more critical to quantization than others?

Related works

Gong, Ruihao, et al. "A survey of low-bit large language models: Basics, systems, and algorithms." *arXiv preprint arXiv:2409.16694* (2024).

Dremov, Aleksandr, et al. "Compute-Optimal Quantization-Aware Training." *arXiv preprint arXiv:2509.22935* (2025).

Wolfe, Cameron R., and Anastasios Kyrillidis. "Better schedules for low precision training of deep neural networks." *Machine Learning* 113.6 (2024): 3569-3587.

Achille, Alessandro, Matteo Rovere, and Stefano Soatto. "Critical learning periods in deep networks." *International conference on learning representations*. 2018.

Abdolrashidi, AmirAli, et al. "Pareto-optimal quantized resnet is mostly 4-bit." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

De Putter, Floran, Sherif Eissa, and Henk Corporaal. "POQ: Is There a Pareto-Optimal Quantization Strategy for Deep Neural Networks?." *IEEE Access* (2025).

Contact

Jim Zhao	jim.zhao [at] unibas.ch	OMLS Group
Aurelien Lucchi	aurelien.lucchi [at] unibas.ch	OMLS Group