ETH zürich

Master's Thesis Proposal

Entropy-Aware Quantization for Compression.

Advisors: Mohamed-Hicham Leghettas, Emil Schätzle
Professor: Markus Püschel

March 26, 2026

The high number of parameters of modern machine learning models led researcher to find new approaches to reduce their storage and memory bandwidth cost. In this regard, two techniques that are often utilized consist in using *quantization* (i.e., using lower bit representations) [1], [2] or *pruning* (i.e., removing) weights and using sparse computations [3]. The goal of this project is to leverage a technique called *Entropy Coding* to unify both approaches: entropy coders (such as [4]) can be used to store weights with variable precision thereby interpolating between quantization and pruning. Such an approach could decrease the size (and hence increase inference speed) of large machine learning models even further.

Your contribution In this project, your goal is to carefully review existing quantization and pruning methods, reformulate their optimization criteria to turn them entropy-aware and throroughly evaluate the achieved compression/loss trade-offs. In particular, your tasks are:

a) Familiarize yourself with existing work in the field of quantization and pruning. You should evaluate the literature with respect to its eligibility for entropy-awareness.

b) Make the existing approaches entropy-aware by adapting their optimization strategy. It should be possible to control the target bit-rate or loss/perplexity in order to explore different trade-offs.

c) Optionally, design your own quantization scheme from what you learned in task b).

d) Evaluate the achieved compression/loss trade-offs of the quantization methods considered in tasks b) and c). Compare them to the related work you found in task a). Analyze which methods worked well and why.

Deliverables

Final report: A digital copy of the project report containing a detailed problem description, an overview of related work and existing approaches, a description of the implementation and design choices, and an evaluation of the results. The final report must be written in English.

Reproducible experimental setup: Implementations, configuration scripts and instructions to reproduce the results reported in the thesis must be submitted to the provided git repository.

Presentation: The results of the project must be presented during an Advanced Computing Laboratory seminar. The presentation is limited to 20 minutes (excluding questions) and should give an overview as well as the most important details of the work.

Contact If you are interested in pursuing this project, please contact emil.schaetzle@inf.ethz.ch or pueschel@ethz.ch.

References

[1] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, *Gptq: Accurate post-training quantization for generative pre-trained transformers*, 2023.

[2] J. Lin, J. Tang, H. Tang, *et al.*, *Awq: Activation-aware weight quantization for llm compression and acceleration*, 2024.

[3] E. Frantar and D. Alistarh, "Sparsegpt: Massive language models can be accurately pruned in one-shot," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23, Honolulu, Hawaii, USA: JMLR.org, 2023.

[4] E. Schätzle, T. Pegolotti, and M. Püschel, *Fast entropy decoding for sparse mvm on gpus*, 2026. arXiv: `2603.01915 [cs.PF]`. [Online]. Available: `https://arxiv.org/abs/2603.01915`.