Title : quantization and low complexity implementation of recent neural network architectures

The advancements in deep learning architectures such as Convolutional Neural Networks (CNNs), Transformers, and emerging models like Mamba have led to significant improvements across various domains. However, the computational and memory costs of these models present challenges for deployment on embedded systems. Quantization and Pruning techniques aim to reduce model size and computational complexity while maintaining accuracy. In addition to traditional fixed-point, integer quantization and mixed-precision training methods [GHO22], in this thesis we will explore efficient power-of-two (PoT) quantization methods where weights and activations are restricted to PoT or sum of PoT values to enable highly efficient hardware implementations through simple bit-shift operations. We will also evaluate the performance, accuracy and computational cost of the proposed solutions using Field-Programmable Gate Arrays (FPGA) proof of concept. The use of FPGA for quantized neural network inference represents a promising approach to combine hardware flexibility, energy efficiency and high performance. Unlike graphical processors (GPUs) or dedicated accelerators (TPUs, NPUs, Groq LPUs), FPGAs offer fine-grained programmability of hardware resources, allowing to optimize each step of the neural computation according to the specific requirements of the quantized model.

The joint optimization of hardware and quantization algorithms will make it possible to fully exploit the potential of FPGAs to make artificial intelligence more efficient and accessible, particularly in the fields of edge computing and IoT.

Background and literature review

Two classes of quantization algorithms can be considered: Post-Training Quantization (PTQ) and Quantization-Aware-Training (QAT)][MEN23]. QAT relies on backpropagation to update the quantized weights reduces the quantization noise errors during training and achieves better solutions than PTQ. PTQ quantizes trained NNs but generally achieves limited performance when considering low-bit quantization [NAG20]. Different quantization strategies have been proposed including uniform and power-of-two quantization [ZHOU21] [WAN22].

The integration of FPGAs in embedded systems is already being exploited, notably with solutions such as Xilinx Alveo or Intel Stratix FPGAs, which offer specialized DSP blocks for accelerating neural networks. These devices are increasingly used for real-time applications, such as computer vision or natural language processing on resource-constrained devices. One interesting approaches is to implement parallel and pipelined computational units tailored for convolution and matrix multiplication operations, which are the fundamental components of convolutional neural networks (CNNs). In the past, several studies have shown that hardware implementation of quantized networks on FPGAs can achieve comparable or even superior performance to GPUs, while reducing energy consumption [ALB25]. The solution proposed in [UMU17], called FINN, which is a binarized neural network (BNN) architecture based on FPGA, demonstrates that quantized models can take advantage of the low latency of configurable logic circuits to accelerate inference.

Recent works have started to explore these methods showing promising results in maintaining accuracy while drastically improving hardware efficiency. The approach based on Successive Vector Approximation for PTQ proposed in [COE25] will be also a starting point of this thesis.

Scientific objectives

In this thesis, we will first focus on recent CNN architectures. For those pre-trained architectures, we will propose PTQ considering both classical approach (using available data) and Data Free Quantization (DFQ) [NAG19].

We will conduct research works in the following directions :

1. **Analyze the impact of the different quantization techniques** on CNNs, Transformers and more recent NN architectures in terms of accuracy, efficiency, and computational cost. Standard data set from computer vision, telecommunications and real-world applications will be considered for evaluation. Models will be implemented using frameworks such as PyTorch.
2. **Develop novel quantization techniques, including power-of-two methods,** tailored to modern NN architectures that preserve model performance while optimizing inference speed and hardware complexity.
3. **Investigate architectural solutions.** We will implement the proposed solutions using FPGA proof of concept and will evaluate the performance, accuracy, computational cost …

References

[GHO22] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). « A survey of quantization methods for efficient neural network inference", In Low-Power Computer Vision (pp. 291-326). Chapman and Hall/CRC

[MEN23] Menghani G., "Efficient deep learning: A survey on making deep learning models smaller, faster, and better," ACM Computing Surveys, vol. 55, no. 12, pp. 1–37, 2023.

[NAG20] Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T, "Up or down? adaptive rounding for post-training quantization", In Intern. Conference on Machine Learning, pp. 7197–7206. PMLR, 2020.

[ZHOU21] Zhou, S., & Lin, H. (2021), "Power-of-Two Quantization for Convolutional Neural Networks: A Hardware-Efficient Approach", IEEE Transactions on Circuits and Systems for Video Technology.

[WAN22] Wang, T., Li, X., & Zhang, Y. (2022), "Towards Ultra Low-Power Deep Neural Network Inference via Power-of-Two Quantization", Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2022.

[NAG19] Nagel, M., Baalen, M. v., Blankevoort, T., and Welling, M, "Data-free quantization through weight equalization and bias correction", In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1325–1334, 2019.

[COE25] L. S. Coelho, P. S. R. Diniz, D. Le Ruyet & L. Lovisolo (2025), "Multiplierless MLP Using Successive Vector Approximation in Post-Training Quantization", accept to ISCAS 2025

[ALB25] J. Albrecht et al., "Comparative Analysis of FPGA and GPU Performance for Machine Learning-Based Track Reconstruction at LHCb," *arXiv preprint arXiv:2502.02304*, 2024..

[VEN21] Venkatesh, G., Das, R., & Hegde, R. (2021). *Efficient FPGA-Based Implementation of Quantized Neural Networks for Edge AI.* IEEE Transactions on Circuits and Systems.