ANR CominLabs LeanAI
Final Scientific Report

Inria Centre at Rennes University – Nantes University
Silviu Filip, Olivier Sentieys – Anastasia Volkova

1 Context and objectives of the project

Deep learning has achieved remarkable success across a wide range of applications, but these advances are largely driven by high-precision arithmetic and large-scale computing infrastructures. In contrast, many application domains targeted by the LeanAI project—embedded systems, edge devices, and resource-constrained platforms—require strict control of computational cost, memory footprint, latency, and energy consumption. This discrepancy raises fundamental scientific and technological challenges related to numerical precision, error control, and hardware efficiency.

The central objective of the LeanAI project was to develop principled and practical approaches to *mixed-precision deep learning*, spanning theory, algorithms, software tools, and hardware prototyping. The project aimed to go beyond uniform low-precision approaches by addressing: (i) sound error and precision analysis of neural network computations, (ii) efficient arithmetic operators and hardware building blocks, (iii) fine-grained software frameworks for simulating custom numerical formats, (iv) FPGA-based acceleration and validation, and (v) learning-aware techniques such as quantization-aware training and structured pruning.

LeanAI adopted a co-design perspective, where numerical analysis, learning algorithms, and hardware constraints are considered jointly, with the dual goal of improving efficiency and preserving correctness guarantees when required (e.g., in safety-critical or embedded control applications).

2 Axis A: numerical foundations and modeling

2.1 Sound mixed-precision fixed-point quantization

A major scientific contribution of LeanAI is the development of a sound and fully automated method for mixed-precision fixed-point quantization of feed-forward neural networks. The approach formulates precision selection as a mixed-integer linear programming (MILP) problem, combining interval-based range analysis with explicit constraints on roundoff error propagation and overflow prevention.

By exploiting the structure of neural networks (layered dot-products, bias addition, piecewise-linear activations), the method scales significantly better than generic precision tuning tools applied to fully unrolled code. The resulting precision assignments are guaranteed to satisfy a user-specified global error bound, while minimizing a cost function related to implementation complexity. This work provides a rigorous bridge between real-valued models used in verification or analysis and deployable fixed-point implementations.

Related publications: [9]

2.2 Probabilistic analysis of limited-precision stochastic rounding

The project also advanced the theoretical understanding of stochastic rounding in low-precision computations. While stochastic rounding is often analyzed under idealized assumptions, LeanAI investigated the effects of *limited randomness*, which is a key constraint in practical hardware and software implementations.

A probabilistic error analysis was developed that explicitly models the number of random bits used by the rounding operator. The results show that insufficient randomness can introduce bias and alter error propagation properties. The analysis yields practical guidelines for selecting the amount of randomness required in summation and inner-product operations, aligning theoretical guarantees with implementable designs.

Related publications: [13]

2.3 MPtorch: fine-grained mixed-precision simulation

LeanAI delivered MPtorch, an open-source PyTorch-based framework enabling fine-grained simulation of custom and mixed-precision arithmetic during neural network training and inference. MPtorch allows users to specify numerical formats and rounding modes at the level of individual operations and execution phases, while relying on standard floating-point arithmetic internally and explicit rounding to emulate reduced precision.

MPtorch serves both as a research tool for exploring numerical effects and as a methodological bridge between algorithmic development and hardware-oriented design choices.

MPtorch: `https://github.com/mptorch/mptorch`

3 Axis B: hardware efficiency and learning-aware optimization

Survey publication: [2]

3.1 Hardware-aware quantization and arithmetic operators

The project investigated quantization strategies explicitly tailored to hardware constraints, including multiplierless and operator-specialized implementations. This work highlights that efficiency gains depend not only on bit-width reduction, but also on the compatibility between numerical formats and available arithmetic building blocks (e.g., shift-add structures, constant multipliers).

These results complement the sound quantization framework by ensuring that validated numerical choices can be mapped efficiently to concrete hardware architectures.

Related publications: [3, 5, 6, 8, 7, 11]

3.2 FPGA acceleration and structured pruning

LeanAI explored FPGA-based acceleration strategies combining quantization with structured sparsification. Pattern-aware pruning techniques were studied to induce regular sparsity patterns that are well-suited to hardware mapping. FPGA case studies demonstrate that structured pruning can significantly reduce resource usage while preserving performance and predictability.

Related publications: [12]

3.3 Quantization-aware and adaptive training

On the learning side, the project produced results on quantization-aware training, including adaptive bit-width strategies where precision choices are optimized during training. These approaches improve the robustness and efficiency of low-precision deployments and naturally integrate with the MPtorch framework and hardware-oriented objectives.

Related publications: [1, 4, 10]

4 Quantitative outcomes and impact

The project resulted in multiple peer-reviewed publications in international journals and conferences, as well as archived preprints. These publications cover sound mixed-precision quantization, stochastic rounding analysis, hardware-aware quantization, FPGA acceleration, and adaptive quantization-aware training. The complete list of publications is provided in the bibliography below.

Following software is issued from the results of the project:

- **MPtorch**: open-source framework for fine-grained simulation of mixed-precision arithmetic.
- **Aster**: prototype tool implementing sound MILP-based mixed-precision fixed-point quantization.
- **AdaQAT**: a Python library enabling the mixed-precision QAT framework

Also, FPGA-based prototypes were developed to validate the proposed numerical methods and to assess their impact on latency and resource usage using high-level synthesis toolflows.

The project funded an organization of a national workshop on Mixed-Precision Computing, held in Paris in 2023 that provided a platform for PhD students to exchange experience on AI- and Scientific Computing-related mixed-precision algorithms.

The outcomes of LeanAI on hardware-aware learning are currently being improved and extended in the ongoing PEPR IA Holigrail, in which all project inverstigators participate.

5 Human resources and staffing evolution

The LeanAI project initially involved doctoral positions at both partner sites. During the course of the project, the staffing situation evolved due to several departures and specific personal or structural constraints, which are documented below in the interest of transparency.

A PhD candidate based in Nantes left the project and academic research for personal reasons, which also involved relocating outside Nantes and France. This departure was unrelated to the scientific content or organization of the project.

A second PhD candidate, based in Rennes, left academia after two years of doctoral work in order to pursue a position as a machine learning engineer in industry. This decision reflects the particularly strong competition between academia and industry in the field of artificial intelligence, where salary disparities and employment conditions can be significant. Unfortunately, no peer-reviewed publication resulted directly from this doctoral work prior to departure.

Despite these changes, the project continued successfully. The corresponding funding was reallocated and converted into research engineer positions at both partner institutions. This adaptation proved effective and enabled the continuation of the project along its original scientific axes, leading to several significant results in mixed-precision QAT, software tooling, and hardware-oriented methods as reported in the previous sections.

In addition, the project was impacted by temporary changes in permanent staff availability. Since January 2025, Silviu Filip has been detached from Inria. Furthermore, Anastasia Volkova was on maternity and parental leave from August 2024 to March 2025.

Overall, while the project experienced non-trivial staffing challenges, appropriate mitigation measures were implemented, allowing LeanAI to reach its scientific goals and deliver its main expected outcomes.

References

[1] M. Tatsumi, Y. Xie, C. White, S. I. Filip, O. Sentieys, et al., "MPtorch and MPArchimedes: Open Source Frameworks to Explore Custom Mixed-Precision Operations for DNN Training on Edge Devices," in *ROAD4NN 2021 – 2nd ROAD4NN Workshop: Research Open Automatic Design for Neural Networks*, San Francisco, United States, Dec. 2021.

[2] E. Dupuis, S. I. Filip, O. Sentieys, D. Novo, I. O'Connor, et al., "Approximations in Deep Learning," in *Approximate Computing Techniques – From Component to Application Level*, Springer, 2022.

[3] T. Habermann, J. Kühle, M. Kumm, A. Volkova, "Hardware-Aware Quantization for Multiplierless Neural Network Controllers," in *IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Shenzhen, China, Nov. 2022.

[4] M. Tatsumi, S. I. Filip, C. White, O. Sentieys, G. Lemieux, "Mixing Low-Precision Formats in Multiply-Accumulate Units for DNN Training," in *2022 International Conference on Field-Programmable Technology (ICFPT)*, Hong Kong SAR, China, Dec. 2022.

[5] P. Kashikar, O. Sentieys, S. Sinha, "Lossless Neural Network Model Compression Through Exponent Sharing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 11, pp. 1816–1825, 2023.

[6] R. Garcia, A. Volkova, "Toward the Multiple Constant Multiplication at Minimal Hardware Cost," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 5, pp. 1976–1988, May 2023.

[7] C. Gernigon, S. I. Filip, O. Sentieys, C. Coggiola, M. Bruno, "Low-Precision Floating-Point for Efficient On-Board Deep Neural Network Processing," in *European Data Handling and Data Processing Conference (EDHPC)*, Oct. 2023.

[8] V.-P. Ha, O. Sentieys, "Maximizing Computing Accuracy on Resource-Constrained Architectures," in *IEEE/ACM Design, Automation and Test in Europe (DATE)*, Apr. 2023, Antwerp, Belgium.

[9] N. Brisebarre, S. I. Filip, "Towards Machine-Efficient Rational Approximations of Mathematical Functions," in *IEEE Symposium on Computer Arithmetic (ARITH)*, Sep. 2023, Portland, USA.

[10] S. Ben Ali, S. I. Filip, O. Sentieys, "A Stochastic Rounding-Enabled Low-Precision Floating-Point MAC for DNN Training," in *IEEE/ACM Design, Automation and Test in Europe (DATE)*, Mar. 2024, Valencia, Spain.

[11] C. Gernigon, S. I. Filip, O. Sentieys, C. Coggiola, M. Bruno, "AdaQAT: Adaptive Bit-Width Quantization-Aware Training," in *6th IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, Apr. 2024, Abu Dhabi, UAE.

[12] L. Pradels, S. I. Filip, O. Sentieys, D. Chillet, T. Le Calloch, "FPGA-based CNN Acceleration using Pattern-Aware Pruning," in *6th IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, Apr. 2024, Abu Dhabi, UAE.

[13] E.-M. El Arar, M. Fasi, S. I. Filip, M. Mikaitis, "Probabilistic Error Analysis of Limited-Precision Stochastic Rounding," preprint, 2024.