UNIVERSITÉ CÔTE D'AZUR

LEAT

PhD thesis Project

An energy proportional neuromorphic solution for processing streaming data with SpikeNets

Context

Spiking neural networks are considered as the third generation of neural networks and could thus replace the conventional networks used in machine learning in order to reduce energy consumption of AI, especially in Edge applications.

But taking advantage of SNN needs to efficiently parallelize their execution onto multiple neuromorphic (event-based) processors (NPU) according to the effective profile of activity in terms of spikes generated in each layer of the network.

The literature didn't address the question of energy proportional execution of SNN on parallel architectures such as Loihi 2 (Orchard 2021), Spinnaker (Mayr 2019) or SPLEAT (Abderrahmane, 2021).

The goal of this PhD project is to develop such a new neuromorphic architecture by integrating 3 main architectural features: parallel processing in a dedicated memory hierarchy, computation based on asynchronous graded spikes, and dynamic control of stream of data. The research will mainly address the domain of image processing for embedded and low-power devices.

State-of-the-art

The binary nature of spikes leads to considerable information loss, i.e. quantization errors, causing performance degradation compared to ANNs using floating-point operations. The SNNs quantization error can be reduced by increasing latency over the network. However, with a longer conversion time, more spikes are generated, thus increasing the energy consumption as well. Several techniques have been proposed to minimize both the quantization error and the latency of SNNs. These approaches can be applied to either the ANN-to-SNN conversion or directly during the SNN training using the surrogate gradient (SG) method. In Li et al. (2022) and Rathi and Roy (2021) the authors adopt an ANN-to-SNN conversion scheme and optimize the firing threshold of the spiking neurons after conversion to better match the distribution of the membrane potential. In Castagnetti et al. (2023b) the SNN is trained using SG and the Adaptive Integrate-and-Fire (ATIF) neuron. This ATIF neuron has been proposed as an alternative to the original Integrate and Fire (IF) neuron since the firing threshold (Vth) is a learnable parameter rather than an empirical hyper-parameter. In Guo et al. (2022), the authors also use SG to train the SNN, but introduce a distribution loss to shift the membrane potential distribution into the conversion range of the spiking neurons. With these approaches it is possible to get SNNs with almost no accuracy loss when compared to the equivalent ANNs, using only few timesteps. To further decrease the latency, recent approaches propose to go beyond binary spikes and introduce multi-level spiking neurons, or graded spikes as in Orchard et al. (2021). This mechanism expands the output of spiking neurons from a single bit to multiple bits, thus increasing the information that can be communicated at each timestep, Shrestha et al. (2024). In Guo et al. (2023) the authors propose a ternary spiking neuron that transmits information with {-1, 0, 1} spikes. Moreover, in multi-level spiking neurons the spike is extended to a fixed-point unsigned binary number with m integer bits Feng et al. (2022) and possibly n fractional bits, Xiao et al. (2024). But most of the previous works only focus on the SNN latency. However, it has been shown Lemaire et al. (2023); Castagnetti et al. (2023a) Dampfhoffer et al. (2023) that besides latency, another important parameter that has to be optimized to improve the energy efficiency is the sparsity of the network, in other words the number of spikes, either binary or multi-level, generated during an inference Castagnetti et al. (2025). In this PhD project we will compare binary and multi-level SNNs from the energy-efficiency point of view on its impact onto the neuromorphic architecture SPLEAT and how this feature can help to reach energy proportional computation according to the neural activity.

Project mission

The PhD project will be organized in several periods.

During the first period, we will address the question of the optimization of the existing version of the architecture on FPGA devices. After a complete review of existing neuromorphic hardware architectures and an introduction to training spiking neural networks, we will focus on the profiling of activity and energy of SNN over different datasets and the potential gain offered by integrating graded spikes into the architecture (Castagnetto et al. 2025). This phase will provide conclusions on the precision of high-level estimation method of energy consumption of SNN and the effectiveness of parallel execution of SNN on FPGA.

The second period will be dedicated to the optimization of the memory hierarchy, both by designing of a specific memory hierarchy optimized for efficient usage of memory blocks on FPGA, and by explicit control of clock gating signal per clock region (Varasala 2024)

The third and last period will be dedicated to the definition of a continual prediction paradigm to reduce energy consumption on stream of data and to the application of the proposed solutions to realistic applications on stream of visual data.

References

Abderrahmane, Miramond et al. SPLEAT: SPiking Low-power Event-based ArchiTecture for in-orbit processing of satellite imagery, IJCNN 2022.

Christian Mayr, Sebastian Hoeppner, Steve Furber, SpiNNaker 2: A 10 Million Core Processor System for Brain Simulation and Machine Learning, Arxiv:1911.02385, 2019

Garrick Orchard, E. Paxon Frady, Daniel Ben Dayan Rubin, Sophia Sanborn, Sumit Bam Shrestha, Friedrich T. Sommer, Mike Davies, Efficient Neuromorphic Signal Processing with Loihi 2, arXiv:2111.03746, 2021

A Castagnetti, A Pegatoquet, B Miramond, All in One Timestep: Enhancing Sparsity and Energy Efficiency in Multi-Level Spiking Neural Networks, SSRN 5379149, 2025

Varasala, Karapa & Maddu, Intelligent Clock Gating for FPGA-based RISC Architectures: A Novel Approach to Switching Activity and Dynamic Power Reduction. International Journal of Computer (IJC), vol. 51(1), 2024.

Conference on Computer Vision and Pattern Recognition (CVPR), pages 326–335, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00042. URL https://ieeexplore.ieee.org/document/9880053/.

Y. Guo, Y. Chen, X. Liu,W. Peng, Y. Zhang, X. Huang, and Z. Ma. Ternary Spike: Learning Ternary Spikes for Spiking Neural Networks, Dec. 2023. URL http://arxiv.org/abs/2312.06372. arXiv:2312.06372 [cs].

E. Lemaire, L. Cordone, A. Castagnetti, P.-E. Novac, J. Courtois, and B. Miramond. An Analytical Estimation of Spiking Neural Networks Energy Efficiency. In M. Tanveer, S. Agarwal, S. Ozawa, A. Ekbal, and A. Jatowt, editors, *Neural Information Processing*, pages 574–587, Cham, 2023. Springer International Publishing. ISBN 978-3-031-30105-6. doi: 10.1007/978-3-031-30105-6_48.

C. Li, L. Ma, and S. Furber. Quantization Framework for Fast Spiking Neural Networks. *Frontiers in Neuroscience*, 16, 2022. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2022.918793.

G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies. Efficient neuromorphic signal processing with loihi 2. In *2021 IEEEWorkshop on Signal Processing Systems (SiPS)*, pages 254–259, 2021. doi: 10.1109/SiPS52927.2021.00053.

N. Rathi and K. Roy. DIET-SNN: A Low-Latency Spiking Neural Network With Direct Input Encoding and Leakage and Threshold Optimization. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–9, 2021. ISSN 2162-2388. doi: 10.1109/TNNLS.2021.3111897. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

S. B. Shrestha, J. Timcheck, P. Frady, L. Campos-Macias, and M. Davies. Efficient video and audio processing with loihi 2. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13481–13485, 2024. doi: 10.1109/ICASSP48485.2024.10448003.

Y. Xiao, X. Tian, Y. Ding, P. He, M. Jing, and L. Zuo. Multi-Bit Mechanism: A Novel Information Transmission Paradigm for Spiking Neural Networks, July 2024. URL http://arxiv.org/abs/2407.05739. arXiv:2407.05739 [cs] version: 1.

Practical information

Location : LEAT Lab / SophiaTech Campus, Sophia Antipolis, LIRMM, Montpellier
Duration : 36 months from November 2025
Profile : electronic engineer, VHDL programming, FPGA, embedded programming, machine learning
Research keywords : FPGA, Embedded systems, Edge AI, Spiking neural networks

Contact and supervision

Alain Pegatoquet, Benoit Miramond
LEAT Lab – University Cote d'Azur / CNRS
Polytech Nice Sophia
04.89.15.44.39. / firstname.name@univ-cotedazur.fr