3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

Doctoral research topic

Benoît Miramond, benoit.miramond@univ-cotedazur.fr[1]

[1]**Université Côte d'Azur / LEAT / CNRS , Sophia-Antipolis, France**

TRAINABLE QUANTIZATION OF GRADED SPIKING NEURAL NETWORKS FOR STREAMING EVENT-BASED PROCESSING ON NEUROMORPHIC HARDWARE

Research axis of the 3IA: AI for Computional Biology and **Bio-inspired AI**

The laboratory:
Benoît Miramond is Full Professor in Electrical Engineering at **LEAT** laboratory from University Côte d'Azur (UCA). He leads the eBRAIN research group and develops a interdisciplinary research activity on embedded Bio-inspiRed AI and Neuromorphic architectures, especially based on SNNs. LEAT is a mixt research unit (UMR 7248) from UCA and CNRS.

Abstract

The "neuromorphic" event-based approach to vision and image sensing is recently gaining substantial attention as it proposes solutions to the problems encountered with conventional technology of image processing. The output of such a sensor is a time-continuous stream of pixels, delivered at unprecedented temporal resolution, containing zero redundancy and encoding orders of magnitude higher dynamic range than conventional image sensors. However, due to the lack of alternatives so far, the event-based, asynchronous output of these sensors have been processed using conventional computing devices such as CPUs and GPUs. This way of processing is obviously non-ideal and does not allow to fully benefit from the unique characteristics of such sensors. In this doctoral project, we will develop new training methods adapted to graded spiking neural networks in order to optimize the network sparsity in the case of streaming data captured by event-based sensors applied to a realistic contexts of moving objects. To reach this goal, we will explore and evaluate on-chip the efficiency of learning methods.

Keywords: bio-inspired computing, spiking neural networks, event-based sensors, supervised and unsupervised learning, neuromorphic systems, embedded applications

Application:
Apply by sending an email directly to the supervisor. The application will include:

- Letter of recommendation of the supervisor indicated above.
- Curriculum vitæ.
- Motivation Letter.
- At least, one letter of recommendation of the thesis supervisor.

1 Context

Spiking Neural Networks (SNN) models have been studied for several years as an interesting alternative to conventional Neural Networks both for their reduction of computational complexity in deep network topologies, and for their natural ability to support unsupervised and bio-inspired learning rules. In the context of interest of image processing, SNN are particularly suitable with event-based sensors and are therefore more suited to capture spatio-temporal regularities in a constant flow of events. The combination of event-based sensors and SNN networks make it possible to considerably increase the energy efficiency of neural-based AI while guaranteeing low reaction times in the context of real-time embedded processing such as autonomous systems, especially autonomous vehicles, drones or satellites. Much work has been proposed in this framework but very little of it addresses the formalization of the generalization of SNN to graded spikes, where events can be valuated. The eBRAIN group already developed an implementation of an embedded processing chain associating sensor and computation within an autonomous device equipped with a FPGA circuit. This chain will be used to provide realistic evaluations of the methods studied during the PhD. requires to overcome mainly three scientific and technical issues: i) to formalize the generalization of SNN to the case of graded-spikes ii) to confront the use of existing learning rules applicable to graded SNN and natively adapted to event-based data, iii) to develop new training methods optimizing the intrinsic sparsity of the network iv) make the method scalable to realistic usecases by tacking into account the quantization of SNN networks, v) to evaluate the methods on neuromorphic hardware to verify the resulting energy efficiency. This project proposes to leverage the scientific background of the eBRAIN research group from University Cote d'Azur and the current neuromorphic technologies to realize an unprecedented breaktrough on event-based processing with spiking neural networks.

2 Goals of the doctoral project

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 (V_{th}) 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 spikesOrchard 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 timestepShrestha 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 bitsXiao 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. In this project we will compare binary and multi-level SNNs from the energy-efficiency point of view. Our analysis will be based on the metric proposed in Lemaire et al. (2023), which is intended to be

independent from low-level implementation choices. The main objectives of this doctoral project is thus to:

- propose a multi-level model of an IF spiking neuron compatible with SNN direct training using SG.
- train and characterize the spiking activity of SNNs on different image and audio classification problems.
- Finally, compare the energy efficiency of binary and multi-level SNNs and highlight the different trade-off between latency/sparsity and accuracy for each configuration.

To reach this goal, we will mainly explore and compare the most recent learning methods adapted to spiking neural networks: a) conversion, b) Surrogate-Gradient (SG), c) STDP. This comparison will be made over state-of-the art neuromorphic technologies: i) Intel Loihi (Davies et al., 2018), ii) Brainchip Akida and iii) more specifically the SPLEAT architecture developed at LEAT Abderrahmane et al. (2022).

2.1 Skills

Master degree in one of the following domains artificial intelligence, image processing, neuromorphic engineering, spiking neural networks.

Background and experience in machine-learning, spiking neural networks and/or embedded systems.

Strong motivation, team working, fluent in english spoken and written.

Programming skills in python, keras, pytorch or equivalent.

References

N. Abderrahmane, B. Miramond, E. Kervennic, and A. Girard. Spleat: Spiking low-power event-based architecture for in-orbit processing of satellite imagery. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10, 2022. doi: 10.1109/IJCNN55064.2022.9892277.

A. Castagnetti, A. Pegatoquet, and B. Miramond. SPIDEN: deep Spiking Neural Networks for efficient image denoising. *Frontiers in Neuroscience*, 17, 2023a. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2023.1224457.

A. Castagnetti, A. Pegatoquet, and B. Miramond. Trainable quantization for Speedy Spiking Neural Networks. *Frontiers in Neuroscience*, 17, 2023b. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2023.1154241.

M. Dampfhoffer, T. Mesquida, A. Valentian, and L. Anghel. Are SNNs Really More Energy-Efficient Than ANNs? an In-Depth Hardware-Aware Study. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(3):731–741, June 2023. ISSN 2471-285X. doi: 10.1109/TETCI.2022.3214509. URL https://ieeexplore.ieee.org/document/9927729/?arnumber=9927729. Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence.

M. Davies, N. Srinivasa, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1), January 2018. ISSN 1937-4143. doi: 10.1109/MM.2018.112130359.

L. Feng, Q. Liu, H. Tang, D. Ma, and G. Pan. Multi-Level Firing with Spiking DS-ResNet: Enabling Better and Deeper Directly-Trained Spiking Neural Networks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 2471–2477, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/343. URL https://www.ijcai.org/proceedings/2022/343.

Y. Guo, X. Tong, Y. Chen, L. Zhang, X. Liu, Z. Ma, and X. Huang. RecDis-SNN: Rectifying Membrane Potential Distribution for Directly Training Spiking Neural Networks. In *2022 IEEE/CVF*

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 IEEE Workshop 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.