



TÉCNICO
LISBOA



Politecnico
di Torino

Edge Computing Platform for AI

Development of a general, ultra-low-power computing architecture for data acquisition and AI/ML processing on edge, based on spiking neural networks (SNN).

Ph.D. Research Proposal

Link to [FCT - Unite! ULisboa Ph.D. Scholarship Call Notice on EURAXESS \(Scientific Area A2\)](#)

Advisers

Luís Guerra e Silva

Assistant Professor

Department of Computer Science and Engineering

Instituto Superior Técnico - University of Lisbon

luis.g.silva@tecnico.ulisboa.pt

Stefano di Carlo

Full Professor

Control and Computer Engineering Department

Politecnico di Torino

stefano.dicarlo@polito.it

Motivation

Smart sensors are becoming increasingly ubiquitous and commonplace in everyday life. Such sensors should be extremely power-efficient, enabling their small batteries to last for several years. However, this requirement often collides with the need to continuously collect and transmit large amounts of data to be processed by computationally-intensive AI/ML algorithms in the cloud. Not only the power consumption of wireless data transmission is prohibitively high (even for low-power wireless protocols), but also the sheer number of devices that will be deployed in the near future, and the amount of data that they will generate are going to overload even the most powerful cloud infrastructure. Therefore, research is starting to focus on power-efficient solutions that incorporate not only data collection but particularly online AI/ML data processing on edge devices, circumventing the need for expensive data transmission and data-intensive cloud infrastructures. Following this new on-the-edge computational paradigm, systems will not only reduce the need for intensive cloud computing but also ensure a far more secure data privacy management. In fact, by computing on the edge, only a small amount of already processed data will be shared on the cloud, making it possible to avoid privacy leakages.

Most AI/ML algorithms used nowadays are based on applying backpropagation on deep artificial neural networks (ANNs) [9-10]. While these networks, and their underlying algorithms, perform extremely well for a wide range of applications, they are not particularly power-efficient or effective for all tasks. Unlike ANNs, spiking neural networks (SNNs) [1] are not suited to be used on a wide range of applications. Still, they are extremely power-efficient and effective in handling time-code or rate-coded information. SNNs are the computational model that better approximates the key known principles of operation of biological neurons and the human brain. In SNNs, timed sequences of electrical impulses (spikes) are propagated across a network of neurons, which are responsible for processing them. The information is encoded in the timing between spikes or spike rates rather than in the spikes themselves, which are alike. Unlike regular ANNs, SNNs excel at recognizing spatiotemporal patterns. Moreover, they are extremely power-efficient since the network is idle when no stimuli are applied. Recent technology developments have enabled significant advances in hardware implementations of SNNs, fostering their use in a wealth of new applications [2].

While adaptations of backpropagation can still be used to train SNNs, simpler methods based on synaptic plasticity rules, such as spike-timing-dependent plasticity (STDP) [3,4], that mimic the process of learning employed by most biological neural networks, have proven to be as effective, and dramatically more power-efficient, for specific applications, that deals with the identification of timing patterns in continuous data streams, while requiring a significantly smaller number of neurons. Examples of such applications are image categorization [5], unsupervised anomaly detection in continuous sensor data [6], and trajectory prediction in computer vision [7], among several others.

The realization of any edge device for AI/ML, besides hardware for implementing the neural network machinery, requires a general-purpose processing unit for overall control and communication. RISC-V is a free instruction set architecture (ISA) [8] which, unlike commercial ISAs (e.g., ARM and Intel/AMD), is open-source and can be freely used and extended to develop software and hardware that support it. RISC-V is based on reduced instruction set computer (RISC) principles, meaning that the baseline set of instructions is simple and very small, which reduces processor complexity and enables hardware implementations to be extremely power-efficient. The simplicity, power efficiency, and flexibility of RISC-V make it an ideal platform for building intelligent edge devices.

Objectives

The main objective of this work is to research novel, flexible, ultra-low-power computing architectures and underlying algorithms, for data acquisition and AI/ML processing on edge, based on SNNs. Such novel architectures should leverage not only the simplicity and power efficiency enabled by the RISC-V ISA but also the proven ability of even SNNs with a very small number of neurons to effectively identify spatiotemporal patterns using simple learning rules such as STDP.

Fulfilling this objective will require the development of a streamlined version of the RISC-V platform that retains all the capabilities necessary for the target application while ensuring specific performance requirements and extremely low power operation. Additionally, various types of SNN implementations, either as external hardware accelerators or in-memory computing units, among others, should be investigated, as well as their seamless integration (physical interface, protocols, ISA extensions, etc.) with the RISC-V core.

An important and necessary outcome of this work is a proof-of-concept implementation of an ultra-low power AI/ML-enabled edge device, addressing a specific use-case application to be selected that will incorporate and demonstrate the most relevant research outcomes of this work.

Main Tasks

The main tasks to be carried out during the Ph.D. work will be the following (not necessarily in chronological order):

- Complete the necessary coursework (30 ECTS) during the first 12 months.
- Survey the state-of-the-art. Understand the RISC-V ISA and review the existing RISC-V ecosystem, particularly emphasizing ultra-low power implementations targeting edge devices and integration with external accelerators. Review existing successful use cases of SNN applications, focusing on those employing simple learning algorithms (e.g., synaptic plasticity). Review relevant SNN hardware implementation technologies.
- Write and defend the thesis proposal, containing a review of the state-of-the-art, a blueprint of the research work to be conducted, associated risks, and the expected outcomes. To be completed during the first 24 months.
- Research and propose a novel, general, ultra-low-power system architecture for SNN-based edge computing platforms considering the emerging RISC-V ecosystem and underlying algorithms. The proposed architecture should target small, extremely power-efficient SNNs.
- Implement an instance of the general architecture targeting a specific use-case, proof-of-concept application to be selected.
 - Obtain/generate data sets for the specific application
 - Generate the specific system architecture from the general model
 - Select and adapt the learning/inference algorithm on SNNs
 - Simulate the hardware + software description and evaluate
 - Collect experimental data
- Write and defend the Ph.D. thesis.
- Write conference and journal papers to disseminate the research outcomes of the Ph.D. work, whenever deemed appropriate. Present the research results in international fora (e.g., workshops, conferences, meetings, etc).

Other Information

- The student will be required to enroll in a Ph.D. programme on EE or CSE at Instituto Superior Técnico, University of Lisbon, Portugal.
- Check the [Application Tips](#), to help you out in submitting your application to this scholarship.
- The host institution where the research work will be carried out will be [INESC-ID Lisboa](#) (Lisbon, Portugal), a research institute affiliated with Instituto Superior Técnico. Additionally, the student will also spend a significant amount of time in [Politecnico di Torino](#) (Turin, Italy). The student will be able to spend, at most, a total of 2 years abroad (i.e. outside Portugal).
- The scholarship includes the Ph.D. programme tuition, a monthly allowance of 1199.64€ (while in Portugal) or 2008.65€ (while abroad), all tax free, and some travel expenses.
- The advisers collaborate in the EU project NEUROPULS (<https://neuropuls.eu>) on the development and applications of photonic NNs and intend to draw synergies from technologies and results produced in the project for this work. Therefore, the student will have the opportunity to interact with an international multidisciplinary research team.
- For further information on how to apply to the scholarship, please check [FCT - Unite! ULisboa Ph.D. Scholarship Call Notice on EURAXESS \(Scientific Area A2\)](#). The English version is found at the end of the page.

References

- [1] Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659– 1671, 1997.
- [2] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019.
- [3] Wulfram Gerstner, Richard Kempter, J Leo van Hemmen, and Hermann Wagner. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383(6595):76–78, 1996.
- [4] Henry Markram, Wulfram Gerstner, and Per Jesper Sjöström. A history of spike-timing-dependent plasticity. *Frontiers in Synaptic Neuroscience*, 3:4, 2011.
- [5] Diehl, P. U. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers of Computational Neuroscience* 9, 99, 2015.
- [6] A. Amirshahi and M. Hashemi, “ECG classification algorithm based on STDP and R-STDP neural networks for real-time monitoring on ultra low-power personal wearable devices,” *IEEE Transactions on Biomedical Circuits and Systems*, pp. 1–1, 2019
- [7] Debat, G. et al. Event-Based Trajectory Prediction Using Spiking Neural Networks. *Frontiers of Computational Neuroscience* 15, 658764, 2021.
- [8] Waterman, A., Lee, Y., Patterson, D. A. & Asanovi, K. The RISC-V Instruction Set Manual. Volume 1: User-Level ISA, Version 2.0, 2014.
- [9] A. Marchisio et al., “Deep learning for edge computing: Current trends, cross-layer optimizations, and open research challenges,” in 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE, 2019, pp. 553–559.
- [10] M. Shafique et al., “An overview of next-generation architectures for machine learning: Roadmap, opportunities and challenges in the iot era,” in 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2018, pp. 827–832.