PhD Thesis proposal in low-power consumption edge AI

I. Thesis subject

Optimization of hardware multilayer spiking neural networks.

II. Location

University of Bordeaux, IMS Laboratory, (Building A31, 351 Cours de la Libération, 33405, Talence), Hybrid Hardware Computation (2HC) research group.

IMS laboratory offers a multifaceted scientific positioning in systems engineering: the integration of hardware, intelligence and knowledge in communicating and human-centred systems. IMS supports fundamental research as well as project-based interdisciplinary research. More than one hundred research grants are currently running in IMS, targeting domains such as transportation, telecommunications, health, environment and energy.

The 2HC research team focuses on frugal Artificial Intelligence (AI). This involves developing computing paradigms, mainly event-based, that require few resources and can be implemented in lightweight hardware systems. This optimization of energy and hardware resources has applications in embedded systems and/or edge computing. To achieve these objectives, the 2HC team uses tools from many different fields and investigates several candidate solutions like device-, circuit-, and system-level optimization of the design by simulations, integration of nanodevices (memristors, spintronic devices), implementation on purely digital targets, and design of mixed circuits and systems. In particular, this wide range of technical solutions has made it possible to develop hardware-friendly event-driven neural networks with both supervised and unsupervised learning capabilities.

III. Context

In recent years, artificial intelligence (AI) has become increasingly intertwined with our daily lives. However, AI such as that currently supported by most major players in the industry like GAFAM is decentralized to servers. Since the electricity consumption of Internet infrastructures represents about 5% of the world's entire electricity production and because Internet traffic can be expected to triple every three years[1], we are in great need of alternative, energy-saving methods of calculation, so that the large-scale rise in AI does not lead to widespread disillusionment. In addition, embedded systems requiring AI are not necessarily permanently connected to the grid. **The need to develop an energy-efficient hardware for the implantation of AI in nomadic systems is becoming increasingly urgent**.

Major players in the industry like Qualcomm[2], Intel[3] and Google[4], Meta[5] have already proposed CMOS chips for the implementation of AI. However, these dedicated integrated circuits are currently limited to the implantation of continuous-valued neural networks (e.g. multi-layer formal neural networks). **The development of a new hardware substrate must be accompanied by a more ambitious technological solution, e.g., event-based computing, which is particularly suitable for low-latency and low-power systems**.

In this promising computational paradigm, information is created, processed or transmitted only when a change occurs either at the level of the sensor or the calculator. Such a system has thus extremely low power consumption if the activity is null. An illustration of this concept is the event-based camera developed by Prophesee[6] or Samsung[7]. Video streams in conventional systems are produced at about 25 frames per second. A processor then reduces the amount of information by eliminating redundant pixels from one image to another, i.e. if the pixel has not changed it is not stored after compression. Therefore, the scenario until now has been that

1 Jones, *Nature*, 2018
2 Zeroth Processor, Qualcomm, 2013
3 Myriad IC, Intel Movidius
4 Jouppi, et al., *IEEE ISCA*, 2017
5 https://ai.facebook.com/blog/meta-training-inference-accelerator-AI-MTIA/
6 Prophesee's website
7 Son et al. , *IEEE ISSCC*, 2017

redundant information is unnecessarily produced from the outset. On the other hand, the information created in event-based sensors is more meaningful from the very start. Similarly, in calculators based on spiking neural networks (SNNs), computation takes place only when an event occurs. **Beyond reducing the amount of incoming data to process, event-based computing requires fewer operations per second during the inference phase compared to classical artificial neural networks**[8]. Both characteristics – reduction of data and sparse computation – make **event-based computing a promising framework for designing and building energy-efficient hardware for AI**.

This computation paradigm is at the heart of innovative CMOS chips developed by IBM (TrueNorth[9]) or Intel (Loihi[10]). In Europe, several companies already use this principle of event-based calculus[5,11,12,13,14]. The current stakeholders are involved either in neuromorphic sensors or in neuromorphic computing. However, the data processing depends on the nature of the input data. **This PhD work will be a part of the Emergences project**[15] that belongs to the PEPR AI funded by the French Research Agency (ANR).

The Emergences project aims at advancing the state-of-the art on near-physics emerging models by collaboratively exploring various computation models leveraging physical devices properties. **This PhD work will focus on event-based models dedicated to both inference and learning on embedded systems for Edge AI applications, which requires to perform correctly with limited hardware resources while also increasing energy efficiency.**

IV. Thesis objectives

In this PhD work, the student will focus on designing, optimizing and implementing hardware-friendly learning rules for multilayer spiking neural networks.

The application cases of the work will initially be devoted to 1D signals such as audio or biomedical electrode recordings. The first results will then be exploited to address 2D signals such as video. **Another important objective of this thesis is the hyper parameters optimization of multilayer spiking neural networks to reach the highest energy-efficiency possible while striving to get similar performance levels as with state-of-the-art neural networks.**

A cornerstone of the neuro-inspired architectures and algorithms that will be developed and investigated will be to take into account the feasibility of hardware integration, whether for analogue, digital or mixed-mode implementations.

V. Salaries

1800 € per month including social insurance and paid leave.

VI. Skills

The candidate should have a Master or similar degree in electrical engineering, physics or computer science. As the PhD thesis proposal lies at the intersection of artificial intelligence and hardware implementation, the candidate should have a strong background in at least some of these topics. Programming skills are also required.

VII. Contact

Pr Sylvain Saïghi sylvain.saighi@u-bordeaux.fr
Dr Pierre Lewden pierre.lewden@u-bordeaux.fr

8 Tavanaei, *Neural Networks*, 2019
9 Merolla, et al., *Science*, 2014
10 Intel, Loihi
11 Yumain
12 IniVation
13 Brainchip
14 aiCtX
15 https://www.pepr-ia.fr/en/projet/emergence-2/