Post-doc position in digital hardware for AI, SLS Team, TIMA Lab, Grenoble

In short : According to the GAFAM, AI tasks need a 100x power reduction in the coming years to be sustainable. This post-doc position focuses on highly energy efficient hardware/software architectures integration of AI and deep-learning to take up this challenge.

Context : The main challenges address the tight integration of AI accelerators in software-intensive systems bearing in mind non-functional requirements: low to very low power consumption, easy system-level co-processor usage, results reproducibility, real-time and low latency computations, AI functions virtualization for deployment on diverse execution platforms, compatibility with academic or industrial machine learning frameworks, etc. All production grade hardware IA accelerators (TPU, Nervana, DGX, Inferencia, etc) or IPs available for integration use classical digital CMOS, which will for some time still be the dominant hardware integration technology. Gaining a 100x in energy efficiency in this context will come from the architectural side: mathematical changes implied by the hardware constraints, hardware/software integration, and digital hardware implementation. The work takes place in the context of the European project AI4I.

Subject : The goal of the research is to study the applicability of networks with ternary weights and activations. We have designed and developed ternary CNN with some success. However many other network structure classes have been designed. Given its ease of implementation, it is worth studying whether “ternary” can be applied to, for example, recurrent neural networks, long short term memory, temporal convolutional network, etc. We also target an application in this project: manufacturing package detection. It need high-throughput and low-latency, to sort-out faulty packages from batches, so the neural network in charge of taking the decision must be wisely optimized.

The work by itself can be seen in two pieces: an initial phase focusing on the learning process for the type of application we target, and hands-on on the state-of-the-art learning frameworks (pytorch, keras, tensorflow, …). Indeed, the classical learning processes cannot be used as such to produce weights, and coarsely quantized values lead to poor accuracy, therefore work on the subject needs to be done. Note that there is some literature regarding learning with ternary weights and activations, but mainly focused on CNN. A second phase will concern the actual implementation of a proof-of-concept design in hardware. The focus will be on high-throughput hardware design under energy efficiency constraints.

Duration : 1 year, possibly renewable 1 once.

Prerequisits :

PhD in computer science or computer engineering, then either :

- Good knowledge in digital hardware architectures, and in digital design on FPGA/ASIC
- Some knowledge in AI for inference and learning and in current AI frameworks

or

- Some knowledge in digital hardware architectures, and in digital design
- Good knowledge in AI for inference and learning and in state-of-the-art AI frameworks

Contacts :

Frédéric Pétrot, laboratoire TIMA, frederic.petrot@univ-grenoble-alpes.fr
Liliana Andrade, Laboratoire TIMA, liliana.andrade@univ-grenoble-alpes.fr

Lab, team and place:

System Level Synthesis team, TIMA lab, Univ. Grenoble Alpes, Grenoble-INP.
46, avenue Félix Viallet, 38031 Grenoble Cedex