PhD title:

Design of an AI hardware accelerator for edge computing

Thesis supervisors:

Haralampos-G. Stratigopoulos, DR, CNRS, (email : haralampos.stratigopoulos@lip6.fr)
Hassan Aboushady, MCF, Sorbonne Université, (email : hassan.aboushady@lip6.fr)

Sorbonne Université, CNRS, LIP6

PhD context and objectives:

Artificial Intelligence (AI) and Machine Learning (ML) algorithms have been a subject of interest for several decades now. Although AI and ML have gone through hype cycles of disappointment and enthusiasm, recent algorithmic advancements, in particular Deep Neural Networks (DNNs) [1], as well as the availability of big data and the rapid growth of computing power, have renewed interest leading nowadays to applications in numerous distinct fields, i.e., robotics, medicine, autonomous vehicles, computer vision, speech recognition, natural language processing, gaming, etc.

DNN models are computational intensive taking up a number of operations in the order of millions. From a hardware perspective, this poses severe challenges of data storage, data frequent movement, and processing speed on conventional Central Processing Units (CPUs) having a traditional Von Neumann computer architecture, commonly known as the “memory wall” problem. To this end, there is a pressing need for designing dedicated customized processors for AI, referred to as AI hardware accelerators, which belong to the larger family of domain-specific computing paradigms. Widely used AI hardware accelerators today are Graphics Processing Unit (GPUs) and Field-Programmable Gate Arrays (FPGAs), but orders of magnitude of energy-speed improvement can be achieved by Application-Specific Integrated Circuits (ASICs).

Another high incentive for designing AI hardware accelerators is to push the execution of AI algorithms from the cloud closer to the sources of data onto edge devices. This is driven by energy, bandwidth, speed, availability, and privacy requirements. More specifically, edge computing reduces the data transfer requirement saving energy. Given the forecast of several tens of billions of edge devices in the near future being connected to the internet, edge computing would save bandwidth. Several applications, i.e., autonomous vehicles, require low-latency, real-time computation which is slowed down due to the communication with the cloud. Also, several applications require availability, thus they need to be independent of the internet. Finally, handling data locally offers privacy as opposed to transmitting sensitive data over the cloud. Edge AI is a challenging objective since edge devices have limited resources and are often battery-operated.

Design efforts towards embedded or application-specific AI hardware accelerators are intense and on-going. There are several design flavors. Analog and mixed-signal implementations can offer orders of magnitude lower power consumption compared to their digital counterparts, thus they may be better-suited for edge computing being capable of acting directly on sensory data from world-machine interfaces [2]. One way to reduce the energy consumption is approximate computing which refers to using approximate arithmetic units in the processing elements of the hardware neural network [3] or performing network compression or quantization, which means reducing the precision of the weights and neuron activation values by transforming floating point numbers into narrow few-bit integers [4]. Another design paradigm with tremendous is in-memory computing

where the matrix-vector multiplications of the neural network are performed within the memory itself [5]. In-memory computing has two main embodiments, namely performing arithmetic and logic operations within the on-chip SRAM or on memristive crossbar arrays. Finally, another trend is spiking neural networks which are the third generation of neural networks aiming at bridging the gap between biological neural networks and machine learning in terms of speed and energy consumption [6].

The objective of this thesis will be the design of a lightweight, low-energy AI hardware accelerator to be embedded onto edge devices. The particular application we are interested in is spectrum sensing [7]. The edge device will be connected to other devices as well as to the cloud. It is important to analyze in real-time the spectrum of the signals it is receiving for two principal reasons: (a) optimize the spectrum utilization, i.e., give priority to under-utilized frequency bands, so as not to congest the wireless network; and (b) detect incoming signals that present suspicious behavior for security purposes, i.e., jamming signals or signals of a side-channel attack attempting at stealing sensitive data out of the chip or bringing the chip into denial-of-service. First, a DNN model will be designed for incoming signal classification. Then, a dedicated AI hardware accelerator will be designed on which the DNN model will be mapped. The thesis will focus on the circuit-level implementation of the DNN accelerator and will reach up to chip fabrication in an advanced technology.

The thesis will also study the security properties of the DNN accelerator, including resilience to adversarial attacks [8], backdoor attacks [9], DNN model theft [10], and fault injection attacks [11].

The prospective student should be highly motivated and should have good background knowledge on analog and digital integrated circuit design. Knowledge on deep neural networks and cyber-security is a plus.

[1] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
[2] A. Rubino, C. Livanelioglu, N. Qiao, M. Payvand, and G. Indiveri, “Ultra-low-power FDSOI neural circuits for extreme-edge neuromorphic intelligence,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 68, no. 1, pp. 45–56, Nov. 2021.
[3] J. Han and M. Orshansky, “Approximate computing: An emerging paradigm for energy-efficient design,” in Proc. IEEE Eur. Test Symp. (ETS), May 2013.
[4] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1,” arXiv:1602.02830 [cs.LG], Mar. 2016.
[5] A. Ankit, I. Chakraborty, A. Agrawal, M. Ali, and K. Roy, “Circuits and architectures for in-memory computing-based machine learning accelerators,” IEEE Micro, vol. 40, no. 6, pp. 8–22, Nov./Dec. 2020.
[6] M. Bouvier, A. Valentian, T. Mesquida, F. Rummens, M. Reyboz, E. Vianello, and E. Beigne, “Spiking neural networks hardware implementations and challenges: A survey,” ACM J. Emerg. Technol. Comput. Syst., vol. 15, no. 2, Apr. 2019.
[7] A. Emad, H. Mohamed, A. Farid, M. Hassan, R. Sayed, H. Aboushady, and H. Mostafa, “Deep Learning Modulation Recognition for RF Spectrum Monitoring," Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), 2021.
[8] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” IEEE Access, vol. 6, pp. 14410–14430, Feb. 2018.
[9] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, “BadNets: Evaluating backdooring attacks on deep neural networks,” IEEE Access, vol. 7, pp. 47230–47244, Apr. 2019.
[10] W. Hua, Z. Zhang, and G. E. Suh, “Reverse engineering convolutional neural networks through side-channel information leaks,” in Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC), Jun. 2018.
[11] Y. Liu, L. Wei, B. Luo, and Q. Xu, “Fault injection attack on deep neural network,” in Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD), Nov. 2017, pp. 131–138.