Exploring Learning Techniques for Edge AI Taking Advantage of Non-Volatile Memories

Michele MARTEMUCCI [(1)], François RUMMENS [(1)], Elisa VIANELLO [(1)], Sylvain SAIGHI [(2)]

[(1)] CEA, Univ. Grenoble-Alpes, Grenoble, France. [(2)] Laboratoire de l'Intégration du Matériau au Système, Univ. Bordeaux, Bordeaux INP, CNRS, France

The implementation of Artificial Neural Networks (ANNs) is divided into two phases: learning and inference. The learning phase consists in modifying a set of parameters of the network, namely the synaptic weights, according to a learning algorithm to make them converge towards values such that the network accomplishes the task to which it is trained, with a sufficiently high accuracy. The inference phase consists in applying the previously learned task to new input data. Therefore, the synaptic weights are modified several times during the learning phase to converge to an optimum set of values for the desired task, whereas they are fixed during the inference operation. The relatively recent development and remarkable results of ANNs are due to the construction of gigantic databases and algorithmic innovations requiring large hardware resources, which results in equally substantial energy consumptions. As Artificial Intelligence (AI) is now being embedded more and more into various connected objects, ranging from medical implants to autonomous cars, it is clear that the algorithmic and hardware solutions available in data centers will not be able to cover all the AI integration needs.

The field of microelectronics has been working for several years now on the development of emerging memory technologies with the aim of integrating Non-Volatile Memory (NVM) within computing units. In a conventional processor architecture, such co-integration between the computation units and the memory would simplify the memory hierarchy, but also increase the bandwidth between computation and data access. From a technological point of view, the PhD project is based on two non-volatile memory technologies, HfO_2-based OxRAM and FeRAM. The two technologies appear as suitable candidates to enable an on-chip learning system. In particular, OxRAM is a resistive memory technology, i.e. storing the information in the conductance of the device, which can be controlled by means of electrical pulses. OxRAM devices can be used as multilevel cells, as multiple conductance states can be tuned into the same device. The FeRAM technology relies on the possibility to change the polarization of electrical dipoles inside a ferroelectric material by means of electrical pulses. The polarization remains unchanged if the power is shut off. Programming a FeRAM device consists in switching the polarization of the ferroelectric material in one direction or another. Therefore, FeRAM devices present themselves as intrinsically binary devices. As the two memory technologies share the core material, they can be easily co-integrated on the same substrate. Moreover, the quasi-infinite reading endurance of OxRAM devices and their poor writing endurance makes them suitable for inference only applications, whereas the reported large writing endurance of FeRAM device would effectively allow to move training on-chip as well. Eventually, the migration of inference and learning from data centers to edge devices will allow them to adapt to the evolution of input data, to specialize each device to its user, to retain private data and offer faster service.

Therefore, the objectives of this PhD are to study the compatibility of various algorithmic tracks for learning NN with the characteristics of the different NVM technologies developed at CEA LETI and the hardware constraints of on-board electronics, as well as to produce on silicon a demonstration circuit combining NVMs and CMOS technologies. The first year of the PhD led to the design of two memory circuits, with hybrid FeRAM and OxRAM technologies, embedded on the 130nm and 22nm CMOS manufacturing nodes. Also, one of the two circuits has been patented. Future testing of the designed systems will confirm the effectiveness of the proposed solutions.