TinyAI: from edgeAI tools to neuro-spiking architectures

Fabien Clermidy
CEATech, MINATEC, 38054, Grenoble, FRANCE
fabien.clermidy@cea.fr

Artificial Intelligence is nowadays a key feature of smart systems. While the data deluge increases and we must master our global energy consumption, it becomes clear that intelligence-in-the-cloud and "stupid objects" sending raw data is no more acceptable. On the other side, embedding more smartness close to their user solves some issues such as system reactivity, reliability proof or privacy.

However, embedding AI on edge comes with many challenges we are addressing with different solutions.

First, edge computing comes with reduced resources and porting Neural Network Inference solutions is challenging. We have developed for years a specific tool, called N2D2, for porting of DNN on resource-constrained architectures such as microcontrollers or dedicated low-power architectures. This tool can target embedded processors, FPGA or specific tinyAI architectures. N2D2 is compatible with standard environments such as Caffe or TensorFlow and supports Spiking architectures. It is also used to evaluate the gain of specific technologies such as RRAM or 3D stacking.

Second, we are proposing a tinyAI modular and scalable architecture called PNEURO for inference of DNN. A 28nm FDSOI implementation of this architecture, compatible with low-level image processing, can run a SqueezeNet NN on a 224x224 pixels image in 20,6ms, outperforming a GPU-based computing (NVIDIA-TX1) by a factor of 16x in terms of energy consumption for equivalent performance.

Third, we are developing a spiking approach compatible with event-based sensors for integration in Cyber-Physical-System. This approach, based on embedded-Non-Volatile-Memories (eNVM) aimed at solving the global power consumption issue of edgeAI thanks to the combination of event-based (spiking) neural processing triggered on sparse sensors events and multi-level eNVM for solving the challenging memory issue of NN.

Finally, beyond these realizations, we are opening the way to new breakthroughs in the field thanks to perspectives works such as: In-Memory-Computing (IMC), NN hardware design using 3D monolithic integration, AI for cybersecurity or in-line learning solutions.