Titre : Mixed-Signal Accelerator for AIoT

Supervisor:

Van-Tam Nguyen (van-tam.nguyen@telecom-paris.fr)

Summary:

This master 2 internship will explore hardware accelerators for AIoT by embracing mixed signal circuits for deep neural network learning and inference.

Description:

The number of IoT devices is increasing very rapidly, enabling numerous applications including smart manufacturing, personalized healthcare, precision agriculture, automated retail, smart manufacturing, autonomous driving, etc. By embedding AI on these tiny devices or Artificial Intelligence of Things (AIoT), we can directly perform data analytics near the sensor, thus dramatically expand the scope of AI applications. However, they have a very limited resource budget, especially memory and storage showing a big gap between the desired and available hardware capacity.

Deep neural networks have proven remarkably effective in a variety of machine learning tasks including image classification, speech recognition, natural language processing, etc. But they contain millions of parameters and require billions of computations for a single inference, making it difficult to run them on resource-constrained devices. One solution to this problem lies in embracing domain-specific hardware accelerators, which are customized toward the workloads of a certain application class. These accelerators provide specialized operations and parallel computing to improve speed and energy consumption by orders of magnitude relative to general purpose processors. Progress in this field heavily relies on hardware - algorithm codesign.

This internship will explore an additional customization vector by embracing mixed signal circuits for deep neural network learning and inference. This direction is motivated by two observations. First, DNN inference is inherently resilient to small computation errors, and can be desensitized further by injecting noise during training. This property enables bit-width reductions down to four bits and below and can be exploited to absorb mixed signal processing errors. Second, it is well understood that mixed signal computation can be more efficient than digital for low bit precisions due to its structural simplicity.

At the first phase, we will consider mainstay kernels used in traditional digital accelerators. In particular, we focus on convolution layers, which tend to dominate the size and workload of modern deep neural networks with millions of parameters. Within this specific context, the potential merits of mixed signal computation deserve further investigation. Most significantly, it is well understood that the energy consumption of digital deep neural networks is dominated by data movement and memory access and it is important to realize that construction and density of the compute elements plays a significant role in a deep neural network’s data flow. In other words, the motivation for including mixed signal is not limited to improving arithmetic energy but extends to introducing new degrees of freedom for reduced data movement and memory access.

A conventional deep neural network accelerator resembles the structure of a multicore processor and contains an array of tens to hundreds of processing elements. A

processing element typically includes a small amount of buffer memory plus strictly separated multiply-accumulate arithmetic. Each multiply-accumulate operation triggers several accesses to the local memory, introducing unwanted energy overhead. In addition, once the buffers have been filled or exhausted, data movement to a larger global buffer (with higher access energy) is required. A key objective for the circuits investigated in this internship is to reduce the overhead caused by these memory accesses. Digital as well as analog memory and compute elements will be densely interspersed to aim for a “best of both worlds” mixed-signal solution. Dense digital memory is kept in the mix for fast data caching, and so is digital arithmetic for operations that are difficult to emulate in the analog domain. Analog compute is primarily added to enable low-energy spatial accumulation. Instead of writing a local compute result to memory, it can be accumulated via current or charge summation on a wire. Finally, if dense multilevel nonvolatile memory is available, it can find dual use for weight storage and synaptic operations via Ohm’s law. In all possible combinations, close attention will be paid to the efficiency of the required D/A and A/D interfaces.

Depending on the obtained result from the first phase, we will further investigate the in-memory computing approach where compute function will be integrated on the scale of a memory cell. We will also investigate neuromorphic approaches and bio-inspired circuits and architectures in order to further enhance the energy efficiency.