Contribution ID: **163** Type: **not specified**

Weighting Resistive Matrix technique (WRM) for high speed, critical decision-making in data evaluation

A device based on critical decision-making must evaluate data in a time scale short enough for the decision to be useful in the application context. Exemplary cases are the autonomous vehicles and robots, or medical robotic prosthetics, with the inherent needs to fast and meaningfully react to the environment. The very same need is also present in frontier science experiments, where sensors produce higher than manageable data flows, and data driven selective fast decision is needed, also known as either trigger or compressive sensing.
The above-mentioned problems are often combinatorial, because data produced by individual sensors assume a meaningful interpretation only in correlation with the other sensors, in a non-predictable way. Their computational complexity represents a limit to the present state of the art systems, when the tolerated latency prevents to use a distributed computing architecture (such as a cloud) or when the application dictates severe limitation in local weight, volume and power consumption. In facts, the classic approach is based on a rigid sequence: $sensor \rightarrow digitization \rightarrow serialization \rightarrow transmission \rightarrow deserialization \rightarrow computing$.
This scheme, where any intelligence is centralized far away from the sensor, is generally adopted due to the digital standardization advantages, but is in principle inefficient since it accumulates latency and concentrate the data before any processing, which is the opposite of what would suggest a divide and conquer strategy.
A representative attempt to implement a fast decision process are the Associative Memories, content addressable memories that contain a set of precomputed solutions for a fixed set of inputs. These are very fast solutions to manage a big data flow, but their potential scalability is strongly limited by the fact that the address space grows with a factorial progression with dimension of the input parameter space.
Deep neural networks (DNN) are considered instead the mainstream approach to complex data. They consist in a concatenation of simple neural network trained with data, behaving like tailored filters, as flexible as the number of layers of concatenated networks, each representing a possible degree of freedom for the data matching. DNN are more and more often optimized by running on GPUs and/or FPGAs, to increment their speed, at the expenses, though, of power consumption. The amount of flexibility is appreciable but limited by the fact that a DNN is essentially a black box, every time learning from scratch, thus not eligible for an engineering based on incremental knowledge trough problem modelling.
An attempt in overcoming these limitations, is the Weighting Resistive Matrix technique (WRM), conceived for discriminating vertices in real time in HEP experiments; by relying on an analog computing architecture it implements the equivalent of a probabilistic regression at nanosecond scale, without actually computing but exploiting the physics of a specific network. Thanks to this it extracts directly from data the most likely fit parameter values, using the energy of the input signal in a single clock cycle independently from the input. For this feature the WRM is an extremely fast and low power consumption device. A WRM chip has been already produced for HEP and adapted to artificial vision applications. The extensive studies performed in this application, from the full software simulation to the real demonstrator, have shown that WRM potential has never been fully exploited as it is essentially hindered by the digital design inheritance of the present WRM, and by the limited dimensionality of the matrix, constrained by the standard IC technology. The study finally revealed an impressive and very inspiring similarity between the WRM technique and the neural information processing, from the retina up to the visual cortex.
We propose to develop a new IC architecture, inspired to the WRM principles and overcoming its limits, integrating sensors, front end and data processing in the same device, to work ahead of the digitization so to achieve an extreme computing power intensity per electric power unit. This will be done by studying a fully analog neuro-inspired matrix, relying on a high level of internal connectivity, which can be dynamically configured through a memristors-like network. We intend to investigate how this technology can be employed to enable data driven decision at the front end, across successive levels of data abstraction, enabling the mobile things awareness concept. We envisage as realistically reachable target a new generation of reliable and

lightweight self-driving vehicle and autonomous robots safely interacting with the environment.

Summary

A device is proposed based on critical decision-making in evaluating data in a time scale short enough for the decision to be useful in the application context. Exemplary cases are the autonomous vehicles and robots, or medical robotic prosthetics, with the inherent needs to fast and meaningfully react to the environment.

Author: AIELLI, Giulio (Universita e INFN Roma Tor Vergata (IT))

Presenter: AIELLI, Giulio (Universita e INFN Roma Tor Vergata (IT))