Title:
Hybrid ultra-low power edge computing

Contributors:
Alessandro Aimar
Manu Nair

Institutions (both authors):
Institute of Neuroinformatics
ETH Zurich and University of Zurich
Zurich, Switzerland

Synthara Technologies
Zurich, Switzerland

1. Problem

Edge-computing devices such as drones, industrial sensors, biomedical sensors, wearables, etc. operate in highly energy-constrained environments. Artificial intelligence (AI) is often used in these applications to enable personalization, analytics and value-added services. AI requires advanced processing capabilities that cannot be delivered by conventional low power processors such as an ARM Cortex or other traditional von Neumann designs.

2. Technical Approach and its Novelty

We have developed a hybrid 2-core fully-CMOS solution comprising a neuromorphic core and a deep-learning core. Our neuromorphic core is a low-footprint module that can run recurrent and fully-connected neural networks models with a power budget less than a mW. The deep learning core exploits sparsity in neural networks allowing it to dramatically lower memory bandwidth and power. Our current deep-learning core achieves an energy efficiency of 5+ TOP/s at 500 GOPs/s.

3. Results

When deployed, the ultra-low power neuromorphic core monitors the environment looking for anomalies and triggers the deep learning core when it detects an interesting event to run more sophisticated network models leading to an actionable decision. For example, trigger phrase detection task such “Hey Siri” can run on the neuromorphic unit with no battery impact which wakes up the deep learning core to process the more complex user commands.

4. Significance for tinyML Community

Our approach enables the development of powerful AI systems that can operate for a long time in an energy-constrained environment. Our technologies are fully compatible with standard machine learning libraries, allowing developers to create powerful models using the existing tool chains. These constraints are imposed by the neuromorphic core in exchange of extreme energy-efficiency.

5. Reflect and highlight TinyML aspects of the work

Our dual core solution is designed to enable complex decision making on tiny, mobile or in general edge devices. Our solutions enable machine learning in newer applications where this was not possible due to power constraints.