ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXV Cycle

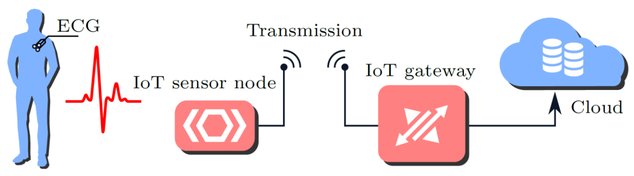
Map-Reduce alternatives to MAC for efficient DNN implementation

Luciano Prono

Supervisor: Prof. Gianluca Setti

Research context and motivation

- Nowadays, the processing of information at the edge of an acquisition system has become increasingly important for **Internet of Things (IoT)**, meaning that the study of low power architectures for signal processing has become fundamental.
- In particular, the implementation of low resources **Deep Neural Networks (DNNs)** for edge computing has become essential. Examples of DNNs applied to edge computing can be found with Computer Vision, Natural Language Processing or biomedical signal elaboration.



- The search for small and portable DNN structures has been addressed in many different ways, ranging from the adoption of new technologies (such as Phase Change Memories for in-memory computing) to optimization techniques such as parameters quantization and pruning.

Addressed research questions/problems

- **Non-conventional DNNs architectures** such as XNOR-net, that uses XNOR and bit-count operations, or logarithmic DNNs, that do not use multipliers, are promising ways to lower DNNs computational cost.
- With this premise, alternatives to the classic **Multiply and Accumulate (MAC)** paradigm have been investigated, with the objective of reducing the energy consumption and the memory footprint of the DNN structures.
- Moreover, **Spiking Neural Networks (SNN)** are of interest for the inherent spatial and time sparsity of the elaborated data.

Novel contributions

- Alternatives to MAC map-reduce paradigm have been developed and studied:
  - **Sum and Max** paradigm (**SAM**);
  - **Multiply and Max&Min** paradigm (**MAM²**).
- The possibility of using these structures for multiplier-free devices has been investigated.
- Pruning of these novel structures has been studied, with promising results.
- A framework for driving a hardware digital accelerator for **Recurrent SNN (RSNN)**, i.e., **ReckOn**, has been developed to test the performance of RSSNs in a real-world scenario.

Adopted methodologies

- **SAM and MAM²** as custom TensorFlow operations have been trained with **image vision tasks** (MNIST, CIFAR-10, CIFAR-100, ImageNet), as well as with other particular tasks such as **signal decoding**. They have been used in **fully-connected layers** both in small dense and in larger convolutional DNNs (AlexNet, VGG-16).
- ReckOn RSNN accelerator design has been loaded on an FPGA and driven by an ARM Cortex M7 MCU. The inference of an **IMU Event classification task** has been tested, as well as **online training with novel e-prop technique** with a cue accumulation task.

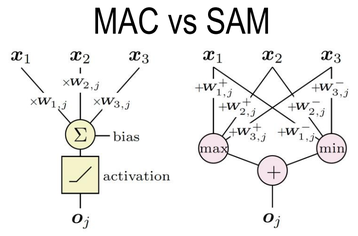
List of attended classes

- 02LWHRV – Communication (02/12/2019, 1 CFU)
- 01RISRV – Public speaking (11/01/2020, 1 CFU)
- 01LCPRV – Experimental modeling: costruzione di modelli da dati sperimentali (04/02/2020, 6 CFU)
- 01SYBRV – Research integrity (20/02/2020, 1 CFU)
- 01SWPRV – Time management (03/03/2020, 1 CFU)
- 01QORRV – Writing Scientific Papers in English (26/03/2020, 3 CFU)
- 08IXTRV – Project management (04/04/2020, 1 CFU)
- 01SFURV – Programmazione scientifica avanzata in matlab (25/05/2020, 4 CFU)
- 01UJBRV – Adversarial training of neural networks (01/07/2020, 3 CFU)
- 01UJARV – Data science for networks (23/07/2020, 4 CFU)
- 01SWQRV – Responsible research and innovation, the impact on social challenges (29/07/2020, 1 CFU)
- 02QUBRS – Statistical data processing (04/02/2021, 4 CFU)
- 01UKAIU – Advanced techniques for digital testing (06/05/2021, 4 CFU)

SAM and MAM² map-reduce paradigms

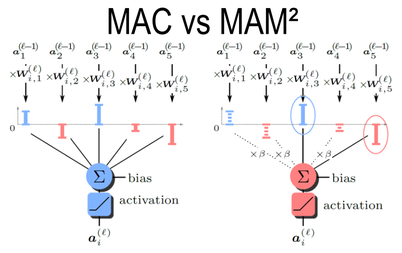
SAM paradigm:
- no multipliers needed (low-power)
- no activation function needed (already non-linear)

MAC vs SAM

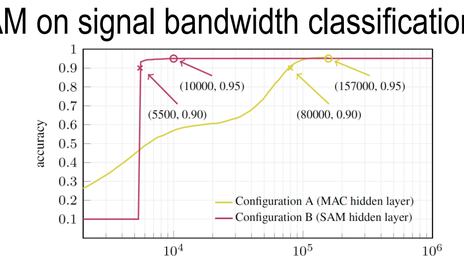


MAC is transitioned to **MAM²** during training with a variable $\beta = 1$ to $\beta = 0$:
- accuracy equivalent to MAC
- potentially multiplier-free (log encoding)

MAC vs MAM²

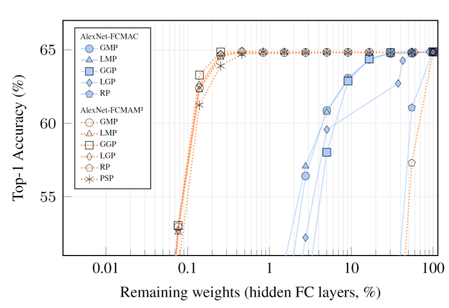


- The computational complexity of Max/Min operations is similar to accumulate operations
  - Both structures are **highly prunable** compared to MAC
- SAM on signal bandwidth classification task



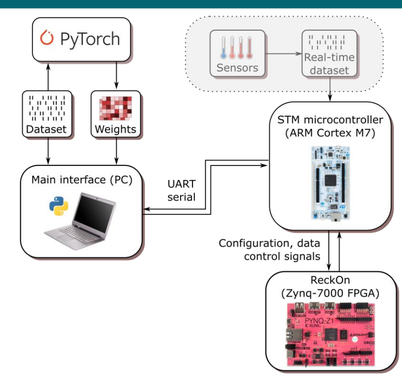
Configuration	Parameters	Multiplications	Additions
A (MAC-based)	$80.0 \cdot 10^3$	$80 \cdot 10^3$	$80 \cdot 10^3$
B (SAM-based)	$5.5 \cdot 10^3$	0	$11 \cdot 10^3$

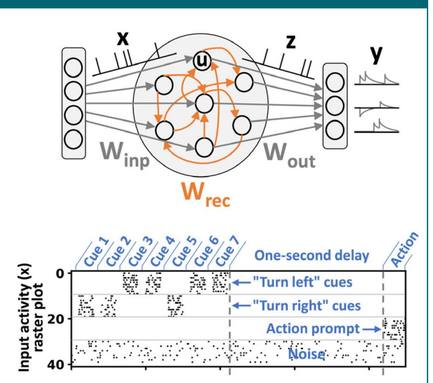
- MAM² on AlexNet + CIFAR-100



ReckOn RSNN hardware accelerator

- **ReckOn** is an RSNN hardware accelerator developed by Charlotte Frenkel (UZH)
- An MCU-based framework has been developed to test it (inference + online learning with e-prop)





Future work

- Implementation and test of **SAM and MAM² in convolutional layers**.
- **Hardware implementation** (FPGA, MCU) of SAM and MAM² to study and measure memory footprint and computational cost (with sparse representations of parameters matrices for non-structured pruning).
- Test **on-device transfer learning** capabilities with ReckOn and e-prop.
- **On-the-fly sensor-to-spikes conversion** with inference/transfer learning on ReckOn.

Submitted and published works

- L. Prono, A. Marchioni, M. Mangia, F. Pareschi, R. Rovatti, G. Setti, "A High-level Implementation Framework for Non-Recurrent Artificial Neural Networks on FPGA" 215th Conference on Ph.D Research in Microelectronics and Electronics (PRIME2019), Lausanne, Switzerland, July 2019, DOI: 10.1109/PRIME.2019.8787830
- M. Mangia, L. Prono, A. Marchioni, F. Pareschi, R. Rovatti and G. Setti, "Deep Neural Oracles for Short-window Optimized Compressed Sensing of Biosignals," in IEEE Transactions on Biomedical Circuits and Systems, vol. 14, no. 3, pp. 545-557, June 2020, DOI: 10.1109/TBCAS.2020.2982824.
- C. Paolino, L. Prono, F. Pareschi, M. Mangia, R. Rovatti, G. Setti, "A Passive and Low-complexity Compressed Sensing Architecture Based on a Charge-redistribution SAR ADC", Integration vol 75, pp. 40-51, November 2020, DOI: 10.1016/j.vlsi.2020.05.007
- M. Mangia, A. Marchioni, L. Prono, F. Pareschi, R. Rovatti and G. Setti, "Low-power ECG acquisition by Compressed Sensing with Deep Neural Oracles," 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS2020), Genova, Italy, 2020, pp. 158-162, DOI: 10.1109/AICAS48895.2020.9073945
- L. Prono, M. Mangia, A. Marchioni, F. Pareschi, R. Rovatti and G. Setti, "Low-Power Fixed-Point Compressed Sensing Decoder with Support Oracle," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), 2020, pp. 1-5, doi: 10.1109/ISCAS45731.2020.9180502.
- L. Prono, M. Mangia, A. Marchioni, F. Pareschi, R. Rovatti and G. Setti, "Deep Neural Oracle With Support Identification in the Compressed Domain," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 10, no. 4, pp. 458-468, Dec. 2020, doi: 10.1109/JETCAS.2020.3039731.
- L. Prono, A. Marchioni, M. Mangia, F. Pareschi, R. Rovatti and G. Setti, "An MCU Implementation of PCA/PSA Streaming Algorithms for EEG Features Extraction," 2021 IEEE Biomedical Circuits and Systems Conference (BIOCAS), 2021, pp. 01-05, doi: 10.1109/BioCAS49922.2021.9645035.
- L. Prono, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "A Non-conventional Sum-and-Max based Neural Network layer for Low Power Classification," in 2022 IEEE International Symposium on Circuits and Systems (ISCAS), May 2022.
- P. Bich, L. Prono, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "Aggressively prunable MAM²-based Deep Neural Oracle for ECG acquisitions by Compressed Sensing," accepted to 2022 IEEE Biomedical Circuits and Systems Conference (BIOCAS), Oct 2022.
- A. Marchioni, L. Prono, M. Mangia, F. Pareschi, R. Rovatti and G. Setti, "Comparison of Streaming algorithms for Subspace Analysis: implementation on IoT devices," submitted to IEEE Internet of Things Journal.

POLITECNICO DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering