Vau da Muntanialas: An Energy-Efficient Systolic Array of LSTM Accelerators

Gianna Paulin, Ph.D. Student, ETH Zurich, Switzerland

Lukas Cavigelli, Post-doctoral Researcher @ ETH Zürich, Switzerland
Francesco Conti, Post-doctoral Researcher @ ETH Zürich, Switzerland & University of Bologna, Italy
Luca Benini, Full Professor @ ETH Zürich, Switzerland & University of Bologna, Italy

The availability of vast amounts of training data and computing power has enabled increasingly sophisticated machine learning algorithms, particularly models based on deep learning, to master all kinds of tasks with astonishing results.

The phoneme recognition step, a first processing step in the time-series problem of speech recognition, alongside other tasks (e.g., language translation) is mostly approached using two particular RNN types: Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs). The vast improvement in accuracy of these methods over previous state-of-the-art has warranted a strong demand for embedded low-power accelerators.

Specialized hardware accelerators for low-power inference using non-recurrent neural networks have achieved energy efficiency gains in the range of 3 orders of magnitude. However, these cannot directly be applied to RNN inference, which comes with additional challenges such as an internal state that needs to be stored and regularly updated, and the densely connected layers which have a very large memory footprint and high bandwidth requirements since a new weight has to be loaded for each multiply-accumulate operation.

We have recently presented an energy-efficient LSTM accelerator called Chipmunk, and now we introduce its successor Muntaniala (Romansh for “marmot” and “marmot burrow”): an extension of Chipmunk for easier integration in a systolic array on a PCB for larger LSTM networks.

Additionally, we built Vau da Muntanialas, a grid of 2x2 Muntaniala LSTM accelerator chips on an FPGA-controlled PCB, collaboratively performing LSTM inference with a hidden state size of 192 in $0.33\ ms$ consuming $2.372\mu J$.

To the best of our knowledge, this system is the world's first hardware demonstration of a systolic multi-chip array concept for RNNs.