Tech Talk: Energy-Efficient Convolutional Neural Network Accelerators for Edge Intelligence

REGISTER HERE

Politecnico di Torino, Room 3I
November 22, 2021 - 5.30 PM

Alessandro AIMAR
Synthara AG

SYNTHARA
Adaptive AI Chips

Bio

Alessandro Aimar received the B.Sc. degree in Physical Engineering from Politecnico di Torino (Italy) and the M.Sc. degree in Nanotechnologies from a joint program of Politecnico di Torino, INP Grenoble (France), and EPFL (Switzerland). After working as an engineer at Imagination Technologies (UK) on the PowerVR architecture, powering at that time Apple iPhones and smartwatches, he joined the Institute of Neuroinformatics (ETH Zürich and University of Zürich) for his Ph.D. studies. In 2017 he founded Synthara, a startup designing ultra-low-power and high energy efficiency hardware accelerators for deep learning algorithms. Today, he is the CTO of Synthara, leading the technology development and the research efforts of the company, focusing on both digital and mixed-signal electronics.

Abstract

Over the last ten years, the rise of deep learning has redefined the state-of-the-art in many computer vision and natural language processing tasks, with applications ranging from automated personal assistants and social network filtering to self-driving cars and drug development. The growth in popularity of these algorithms has its root in the exponential increase of computing power available for their training consequent to the diffusion of GPUs. The achieved increase in accuracy created the demand for faster, more power-efficient hardware suited for deployment on edge devices. In this talk, I'll present a set of innovations and technologies belonging to one of the many research lines sparkled by such demand, focusing on energy-efficient hardware for convolutional neural networks. The talk will start from Nullhop, an accelerator pioneering the use of feature map sparsity, typical of convolutional neural networks, and quantization to boost the hardware capabilities. Nullhop's novelty is its ability to skip all multiplications including a zero-valued activation. It reaches a power efficiency of 3 TOP/s/W with a throughput of almost 0.5 TOP/s in 6.3 mm^2. The talk will continue with a short overview of Elements, a convolutional neural network accelerator architecture that supports variable weight bit precision as well as sparsity, reaching an energy efficiency of over 4 TOP/s/W using only 3.3 mm^2. Finally, the presentation will end with a third accelerator named TwoNullhop, able to skip over zeros of both feature maps and kernels. We tested the TwoNullhop architecture in multiple configurations, the most relevant of them being Carbon, an accelerator that, despite having only 128 multiply-accumulate units and running at a frequency of only 500 MHz, achieves more than 2.4 TOP/s with an energy efficiency of 10.2 TOP/s/W in 4 mm^2.

IEEE
Advancing Technology for Humanity

IEEE PoliTO Student Branch
Politecnico di Torino
IEEE Student Branch