A ULP 22 nm System-on-Chip with Dual-engine Hardware Acceleration for Edge ML Inference

P. Jokic, E. Azarkhish, R. Cattenoz, E. Türetken, V. Moser, P. Nussbaum, S. Emery

Neural network-based object detection algorithms disrupted the field of computer vision, achieving unprecedented detection accuracies in various application domains ranging from large-scale automotive to miniaturized IoT devices. This advance was enabled by the introduction of increasingly deeper and thus more computationally intensive network architectures, challenging the processing hardware. IoT platforms are restricted in size and power, requiring efficient hardware engines to enable on-board processing of such neural network algorithms. We present a system-on-chip, fabricated in an advanced 22 nm CMOS process to provide end-to-end embedded machine learning inference capabilities at the edge.

Smart vision systems provide embedded image analysis capabilities, allowing to implement miniaturized IoT applications for sub-mW face detection as shown in Figure 1. To analyze the acquired images for detecting faces in the field of view, we develop an efficient ultra-low power (ULP) machine learning (ML) inference processor with two ML accelerators.

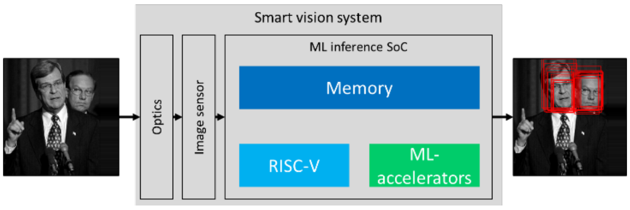

Figure 1: Smart vision system for face detection applications.

The system-on-chip (SoC) is built around CSEM's RISC-V based icyflex-V ecosystem and features 1.2MB of on-chip SRAM memory as well as two ML inference accelerators: one for computing binary decision trees (BDT) and another one for efficiently implementing convolutional neural networks (CNN). A rich set of peripherals allows the SoC to directly connect to an image sensor, allowing to build complete end-to-end solutions without having to add other active components or external memory. The system architecture is illustrated in Figure 2 below.

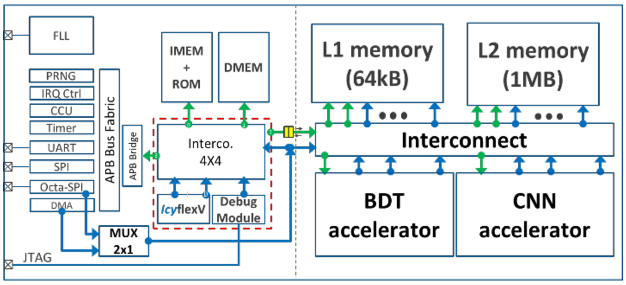

Figure 2: System architecture block diagram.

While CNNs have gained a lot of attention in the last decade, following their success in various object recognition challenges, BDT algorithms have been used for multiple decades. Their simple and scalable computation scheme is based on a cascade of weak classifiers (EC) that allow BDT algorithms to be dynamically adapted (during run-time) to a wide range of algorithmic complexities. Controlling the number of evaluated window positions, the covered range of orientation angles as well as the depth (maximum number of WCs) of the classifier allows to scale both the energy per frame and the analysis accuracy. In the smart vision system use-case this can be exploited by running the BDT algorithm in a low-power setting until a more detailed analysis becomes necessary.

CNNs do not feature such dynamic configurability for trading-off energy per frame versus accuracy but have been shown to tolerate various levels of parameter quantization that enable power reductions by simplifying the processing hardware. Thus, the CNN accelerator of the presented SoC features multi-precision processing, allowing networks with both 1-bit (binary) or 16-bit precision weights to be computed. Binary weights can reduce the memory footprint by 16x and the dominating multiply-and-accumulate (MAC) operation is simplified to an addition/ subtraction, reducing the energy per operation while allowing larger networks to fit onto the on-chip memory. Each layer can be configured individually, such that networks can consist of layers with different parameter precisions.

The ML inference SoC was fabricated in the GlobalFoundries 22 nm FDX process, enabling a low power consumption at up to 180 MHz clock frequency and a small die area of 3.4mm^2. Figure 3 below shows the packaged chip with the following main features:

- CSEM icyflex-V core, achieving 3.2 CoreMark/MHz at 2.23 µW/MHz
- Octo-SPI interface with up to 8 parallel data lanes for up to 180 MB/s sensor data transmission
- Dual-accelerator ML cluster (CNN, BDT) reaching a high computational throughput of up to 5.8 GOP/s
- More than 1 MB of shared memory with direct access from the core, the DMA engine and both ML accelerators

The SoC achieves 150 µW average power consumption while running 320x320 pixel face detection at 1 frame per second with more than 98% accuracy. This enables complex ML applications to be implemented in smart IoT devices which are restricted in size and power.

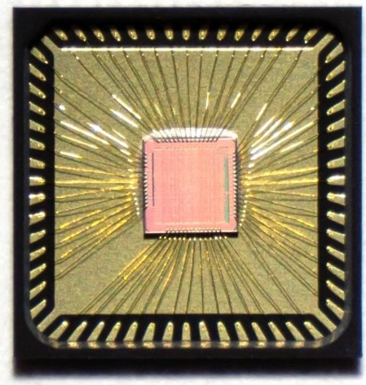

Figure 3: Image of the packaged SoC die.