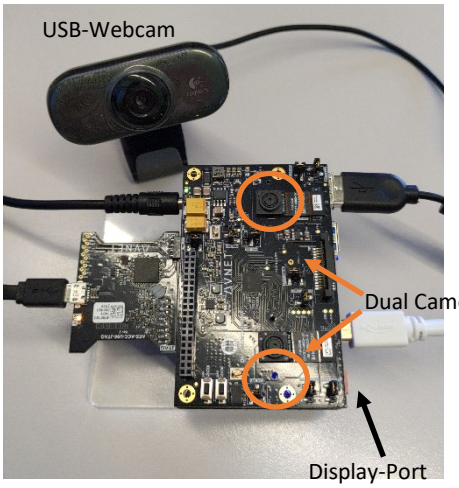
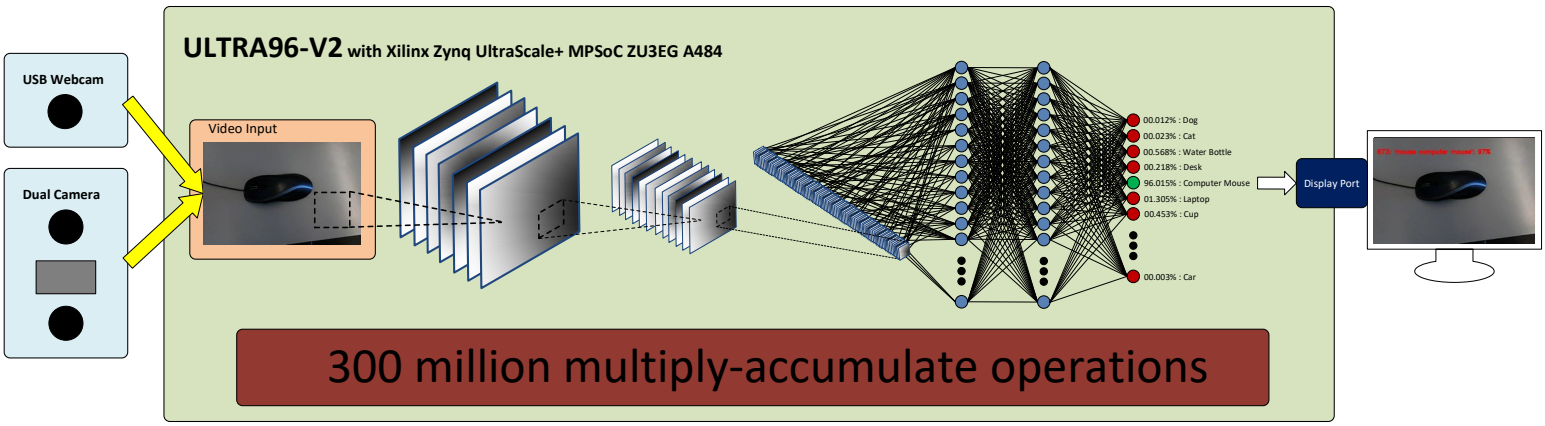
Edge-AI Showcase - Product Brief



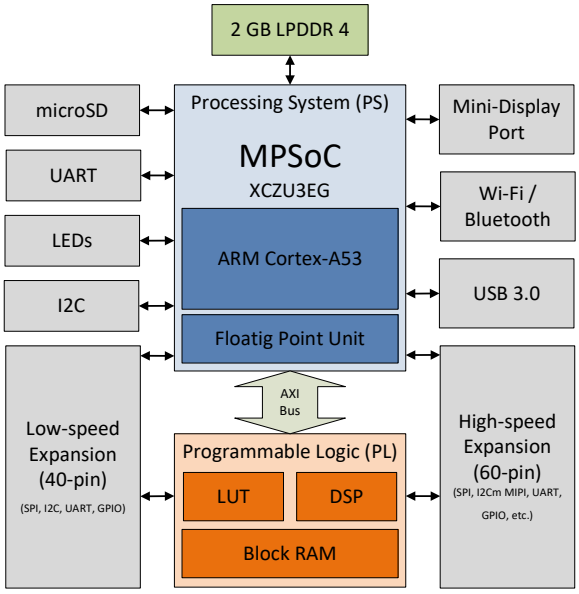
Ultra96-V2 Board with Dual Camera Mezzanine

In recent years, using Convolutional Neural Networks (CNNs) for solving complex machine learning problems has become state-of-the-art. However, for a single inference with CNNs usually several million multiply-accumulate (MAC) operations must be performed. This requires very powerful processing units with correspondingly high energy consumption. For real-time systems with limited resources and energy availability such as embedded systems or IoT-Devices, access to a powerful processing unit is not always possible. *Edge-AI* addresses this problem with several approaches, such as efficient network architectures and implementations.
This project demonstrates how a CNN can be efficiently run on an embedded multiprocessor system with a power consumption of less than 10 watts.
Therefore, two different classic computer vision tasks were implemented using CNNs: Image Classification and Object Detection.
In a first development stage, the network models were implemented to run on the devices Central Processing Unit (CPU). To increase performance, the computationally intensive MAC operations can be accelerated using the devices Field Programmable Gate Array (FPGA). The FPGA enables parallel computation, resulting in a massive speedup and thus a much higher throughput of the system. In a second stage the hardware acceleration was done using a Deep Learning Processor Unit (DPU) from Xilinx. In a third and last stage, the acceleration was implemented using the proprietary BinArray hardware accelerator, which has been developed at HSLU.



Platform Features:

- **ULTRA96-V2 board Xilinx System-on-Chip (SOC)**
  - Xilinx Zynq UltraScale+ MPSoC ZU3EG A484
  - Quad-core Arm Cortex-A53 (up to 1.5GHz)
  - Programmable Logic (FPGA) with 154k System Logic Cells
  - 2 GB LPDDR4 Memory
  - Mini DisplayPort
  - Microchip Wi-Fi / Bluetooth
  - Petalinux Operating System
  - Vitis / Vivado Development environment
- **Dual Camera Board**
  - Two IAS sensor modules
  - AP1302 imaging coprocessor
  - 1920×1080 image from one or both sensor
- **(Optional) USB Camera**



MPSoC Architecture with CPU-based Processing Unit (PS) and Programmable Logic (PL) connected via AXI-Bus system

Development in three stages

To demonstrate several implementations and their advantages, the development was done in three stages. Each consecutive stage can demonstrate some improvement and highlight specific competences in *Edge AI*.

Stage 1:
CPU processing - TensorFlow light
TensorFlow light is a software library to process neural networks. This is easy to implement, but the processing of large images is very slow.

Stage 2:
FPGA acceleration - DPU
The DPU is a parametrizable hardware block (soft IP-core) provided by Xilinx.
The DPU can be integrated in any custom hardware design but is only compatible with Xilinx SoCs.

Stage 3:
FPGA acceleration - BinArray
BinArray is a hardware accelerator to run binary approximated CNNs efficiently in an FPGA. It is fully implemented in VHDL. The development originates from master projects at HSLU.

simple implementation → accelerated processing → generally usable

Two Edge-AI Use Cases

Each development stage can demonstrate two different *Edge AI* Use Cases in image processing. The neural networks implemented can classify images or detecting individual object in an image. They consist of millions of MAC operations. Both are suitable for real-world applications, such as automation, self-driving cars and many more.

Image Classification

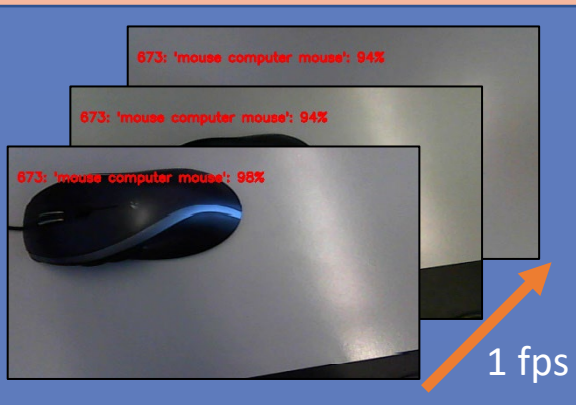
Assign an object class to an image

- only one object per image
- no localization of the object
- 1000 different classes (ImageNet)

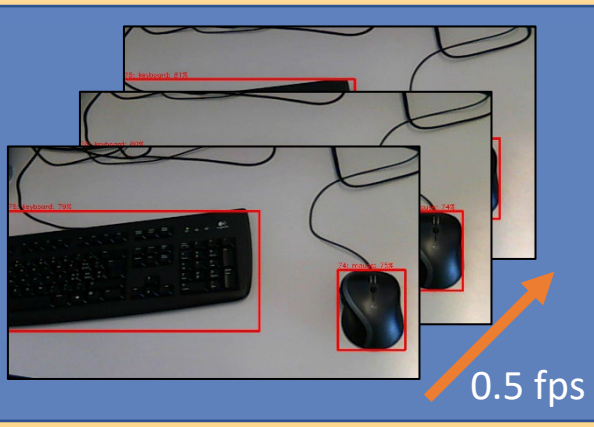
Object Detection

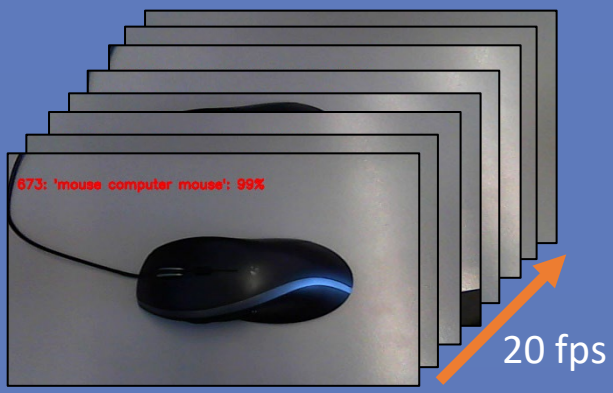
Detect and locate multiple objects in an image

- multiple objects in one image
- objects are marked by a bounding box
- 100 different classes for objects (COCO 2017)



Software Processing (CPU)





Accelerated Processing (FPGA)