

The challenges of data transfer and processing on low resource hardware for AI applications

BILLOT Louis, ISIA Lab researcher

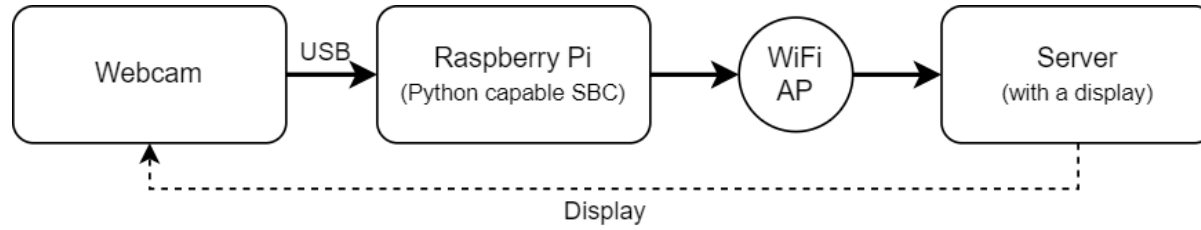
Problems

Object detection and pose estimation applications in real-time using edge devices

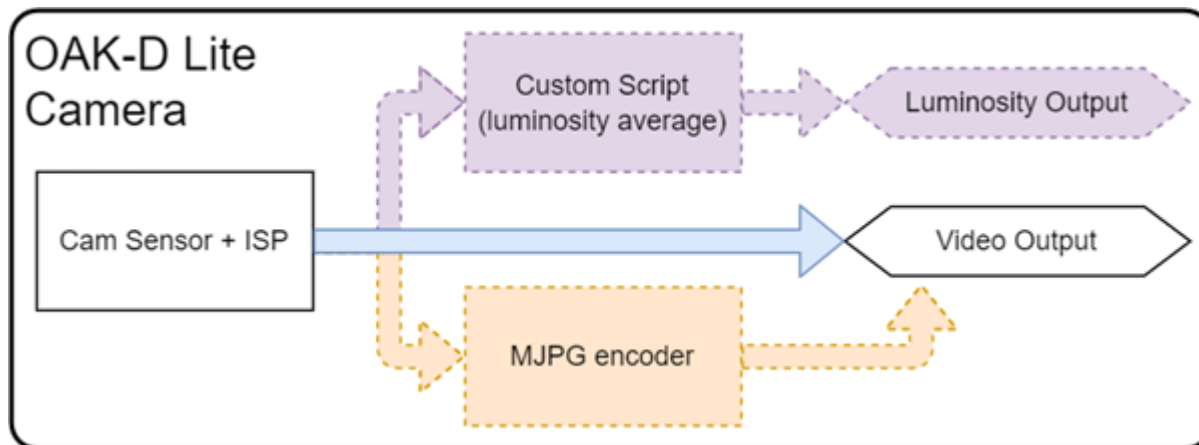
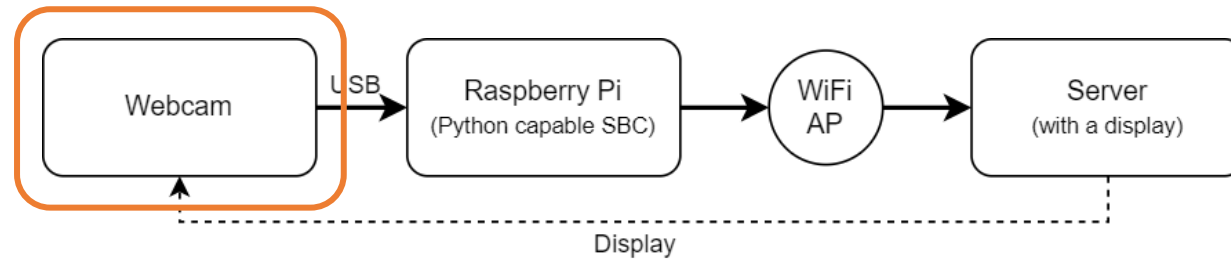
- Best accuracy obtained with heavy models and data with high dimensions
- Edge devices have limited resources (computation, storage, bandwidth, etc.)
- Need for an optimal distribution of the computational load



System under test

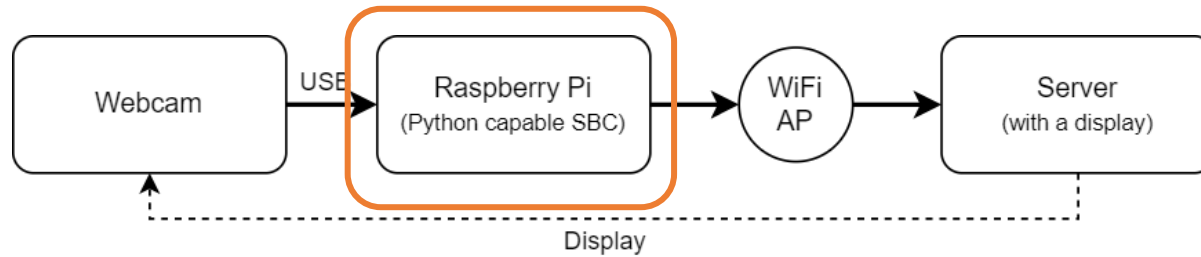


System under test - Camera



- Model implementation limited and complicated
- Very limited resources
- Not all cameras allow embedded computation

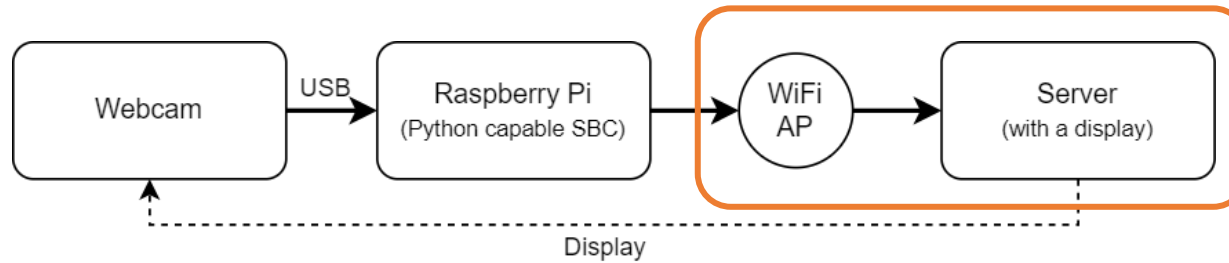
System under test – Edge processing



- CPU (quadcore ARM @1.8GHz)
- RAM (a few gigabytes)
- No GPU (only for low power multimedia)
- WiFi (limited to ~100Mbps)



System under test – Network and server



- No resource problem except network bandwidth!
- Problem with the latency
- Scalability may become an issue



Results

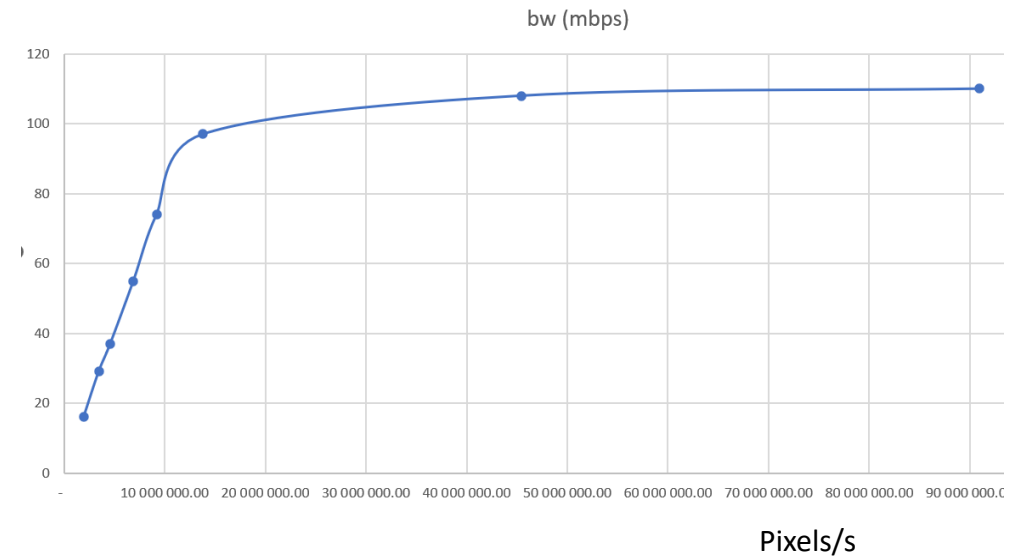
Development of Python tool for latency measurement across the pipeline (released soon™)

Run models at several stages of the pipeline

- On the webcam (OAK-D Lite) new models releasing
- On the raspberry Pi

Dimensionality reduction for faster data transfer

- ROI cropping
- Image encoding



Not encoded

Width	Height	FPS	Bandwidth (Mbps)	Latency (ms)
320	240	10	19.14	101
320	240	20	38.23	101
640	480	7.64	58.42	600

Encoded

320	240	20	1.16	100
640	480	10	2.58	100
640	480	20	5.5	100
640	480	20	6.22	100