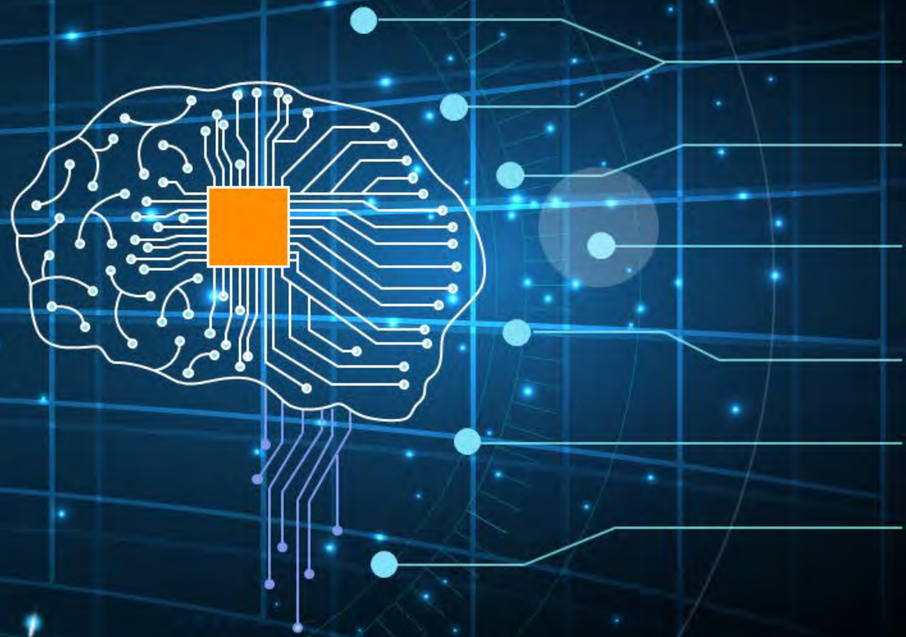
The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

ANDANTE

AI
ECSEL JU

AI4DI

TEMPO



Vienna, Austria

25-26 July 2022

The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

Efficient Edge Deployment Demonstrated on YOLOv5 and Coral Edge TPU

Ruben Proksha (M.Sc.)

Ostbayerische Technische Hochschule Amberg-Weiden

Vienna, Austria 25-26 July 2022

Authors

M.Sc.
Ruben Prokscha
0000-0002-7452-3416

M.Eng.
Mathias Schneider
0000-0002-7562-1122

Prof. Dr.-Ing.
Alfred Höß
0000-0002-1962-0096

Automotive Research

Department of Electrical Engineering, Media Technology and Computer Sciences,

Ostbayerische Technische Hochschule Amberg-Weiden, 92224 Amberg, Germany

Email: r.prokscha, mat.schneider, a.hoess@oth-aw.de

Presentation Outline

- Motivation
- Related Work
- YOLOv5
- Test Setup
- Model Optimizations
- Speed-Accuracy Comparison
- USB Performance
- Software Stack
- Proof of Concept
- Summary and Future Work
- Acknowledgement

Motivation

- Demonstrate the deployment of AI algorithms in an edge context
- Show the potential of edge AI for future Smart City applications
- Expose optimization potential from the hard- and software side
- Show potential pitfalls of deploying AI into resource restricted environment
- Present a lightweight open-source software stack modern edge computer vision software can be based upon

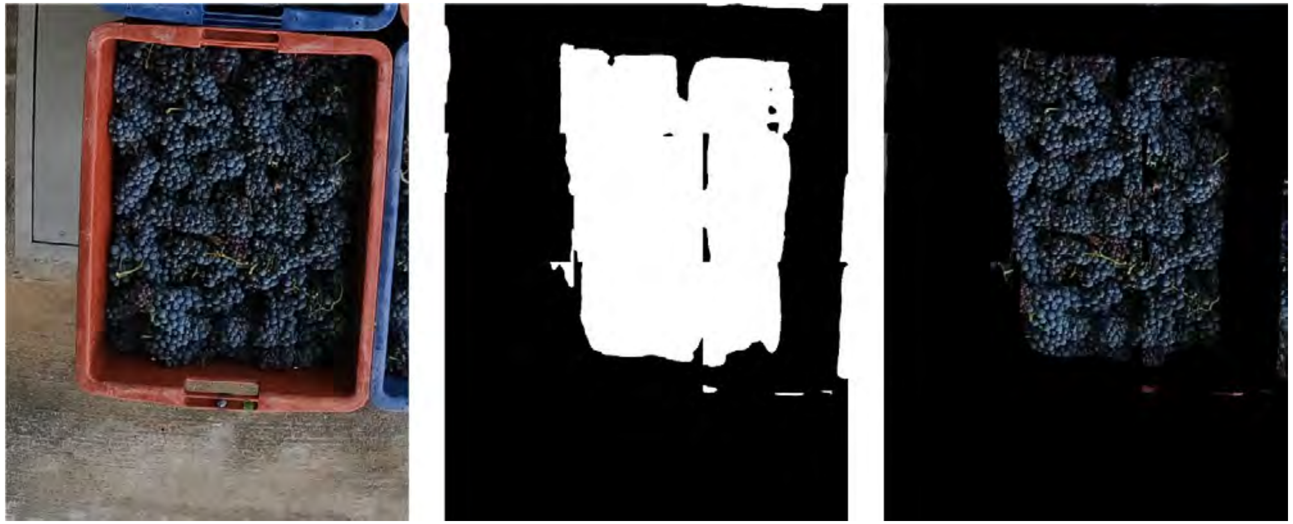
Related Work - Applications

General Application:

- Predictive maintenance [1]
- Intrusion detection [2]

Computer Vision:

- Indoor person-following systems [3]
- Trash and litter detection [4]
- Odometry estimation [5]



Segmentation of grapes [6]

[1] Carlos Resende et al. “TIP4.0: Industrial Internet of Things Platform for Predictive Maintenance”. en. In: Sensors 21.14 (July 2021), p. 4676.

[2] Seyedehfaezeh Hosseininoorbin et al. “Exploring Edge TPU for Network Intrusion Detection in IoT”. en. In: arXiv:2103.16295 [cs] (Mar. 2021).

[3] Anna Boschi et al. “A Cost-Effective Person-Following System for Assistive Unmanned Vehicles with Deep Learning at the Edge”. en. In: Machines 8.3 (Aug. 2020), p. 49.

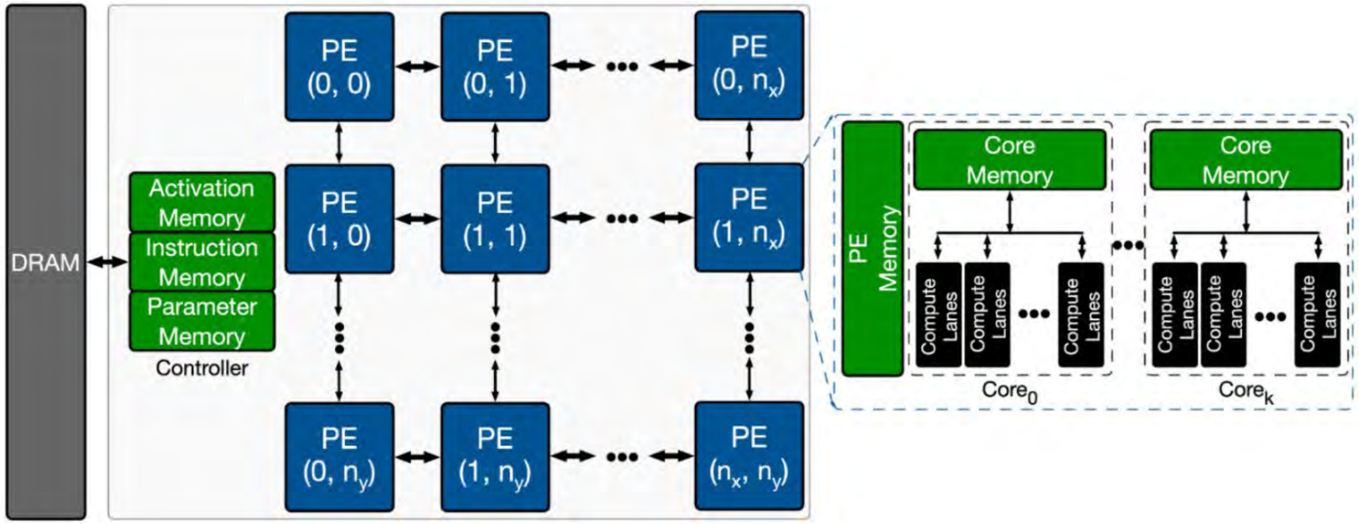
[4] Marek Kraft et al. “Autonomous, Onboard Vision-Based Trash and Litter Detection in Low Altitude Aerial Images Collected by an Unmanned Aerial Vehicle”. en. In: Remote Sensing 13.5 (Mar. 2021), p. 965.

[5] Nitin J. Sanket et al. “PRGFlow: Benchmarking SWAP-Aware Unified Deep Visual Inertial Odometry”. en. In: arXiv:2006.06753 [cs] (June 2020).

[6] Mathias Roesler et al. “Deploying Deep Neural Networks on Edge Devices for Grape Segmentation”. en. In: Smart and Sustainable Agriculture. Ed. by Selma Boumerdassi, Mounir Ghogho, and Éric Renault. Vol. 1470. Cham: Springer International Publishing, 2021, pp. 30–43.

Related Work –Architecture and Performance

- Empirical performance measurements [7, 8]
- Architecture specific optimizations [9, 10]



Principle TPU matrix architecture [9]

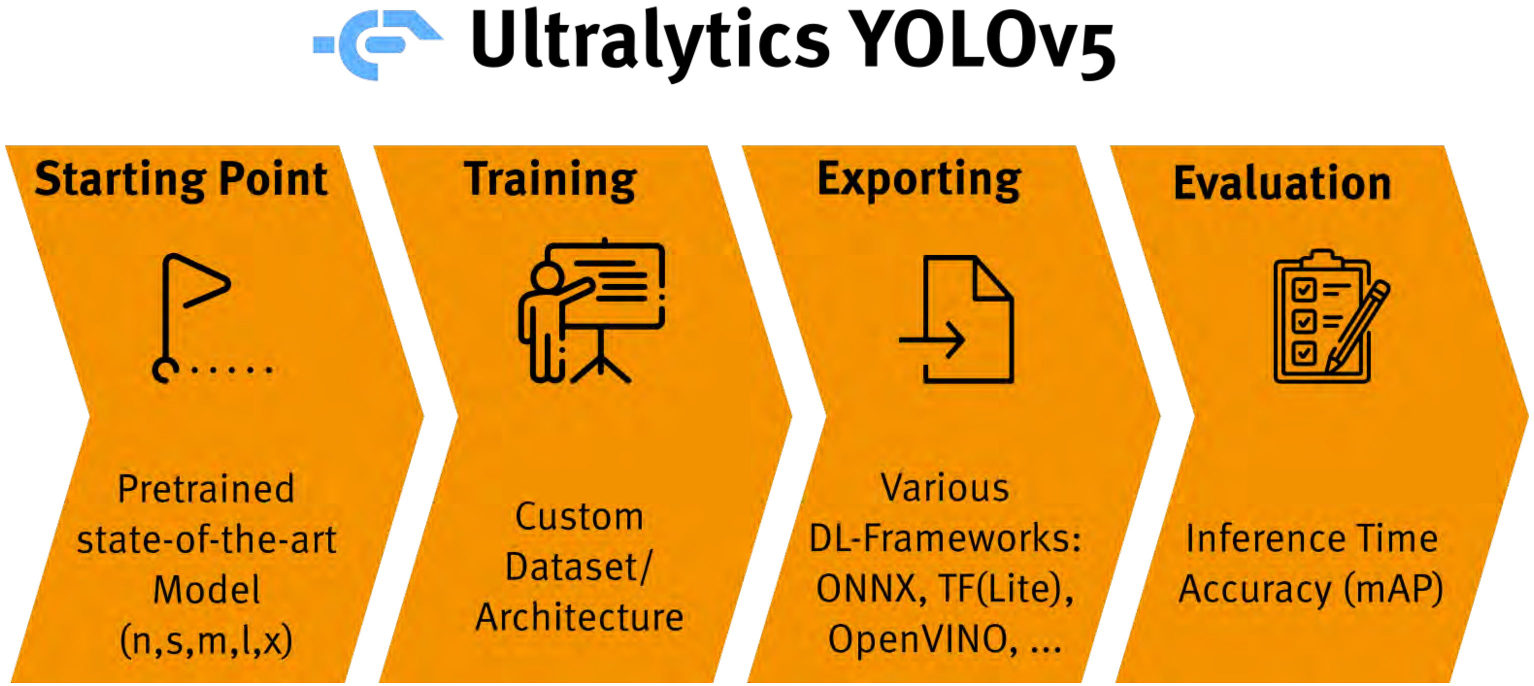
[7] Mattia Antonini et al. "Resource Characterisation of Personal-Scale Sensing Models on Edge Accelerators". en. In: Proceedings of the First International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things. New York NY USA: ACM, Nov. 2019, pp. 49–55.

[8] Ahmad Ammar Asyraaf Jainuddin et al. "Performance Analysis of Deep Neural Networks for Object Classification with Edge TPU". In: 2020 8th International Conference on Information Technology and Multimedia (ICIMU). Aug. 2020, pp. 323–328.

[9] Amir Yazdanbakhsh et al. An Evaluation of Edge TPU Accelerators for Convolutional Neural Networks. Feb. 2021.

[10] Amirali Boroumand et al. "Google Neural Network Models for Edge Devices: Analyzing and Mitigating Machine Learning Inference Bottlenecks". en. In: arXiv:2109.14320 [cs] (Sept. 2021).

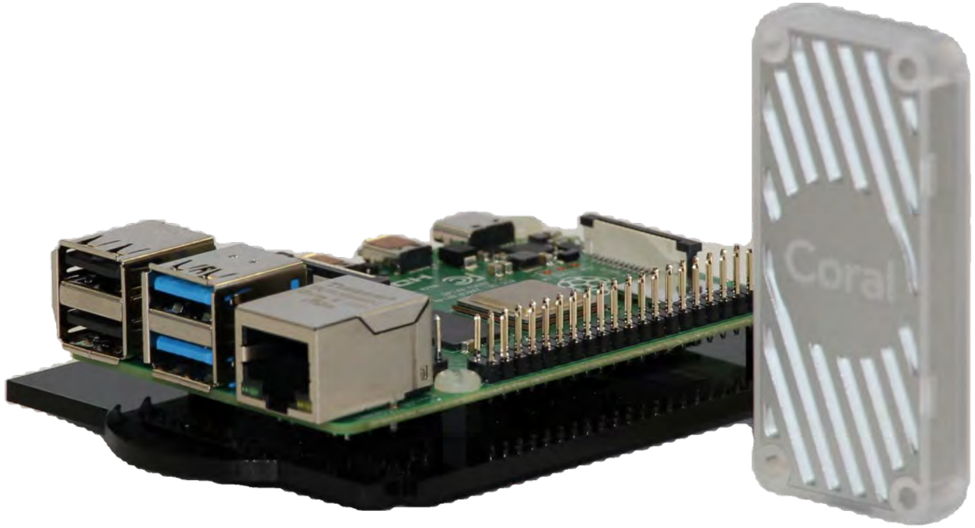
YOLOv5



Yolov5 deployment pipeline

Test Setup

- Google benchmark model tool for performance evaluation [11]
- USB2 and USB3 port speed comparison
- Accuracy evaluation with pycocotools and the Common Objects in Context (COCO) evaluation dataset [12]
- Comparative models from Coral model zoo [13]



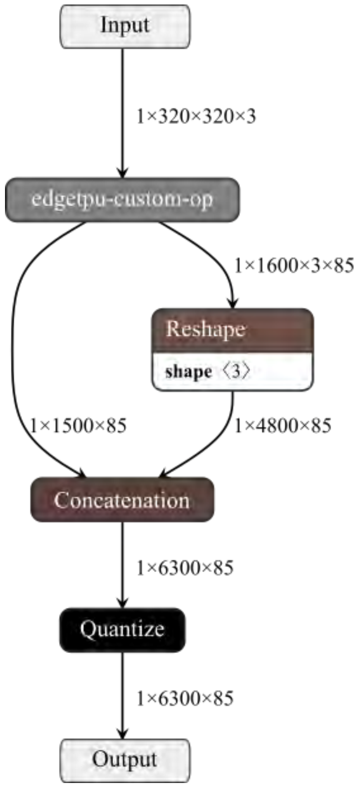
Raspberry Pi 4B with Google Coral edge TPU USB accelerator

[11] Performance measurement — TensorFlow Lite, en. [Online]. Available: https://www.tensorflow.org/lite/performance/measurement (visited on 03/30/2022).

[12] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common Objects in Context," en, in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 740–755.

[13] Models - Object Detection, en-us. [Online]. Available: https://coral.ai/models/object-detection/.

Model Optimizations



Quantized YOLOv5s edge TPU model

- Edge TPU can only process a reduced set of instructions
- A dedicated compiler is used to compile contiguous TFLite ops into an “edgetpu-custom-op”
- All incompatible operations are performed on CPU
- Reshape and Transpose ops become ‘incompatible’ if their input tensor exceeds an unspecified threshold

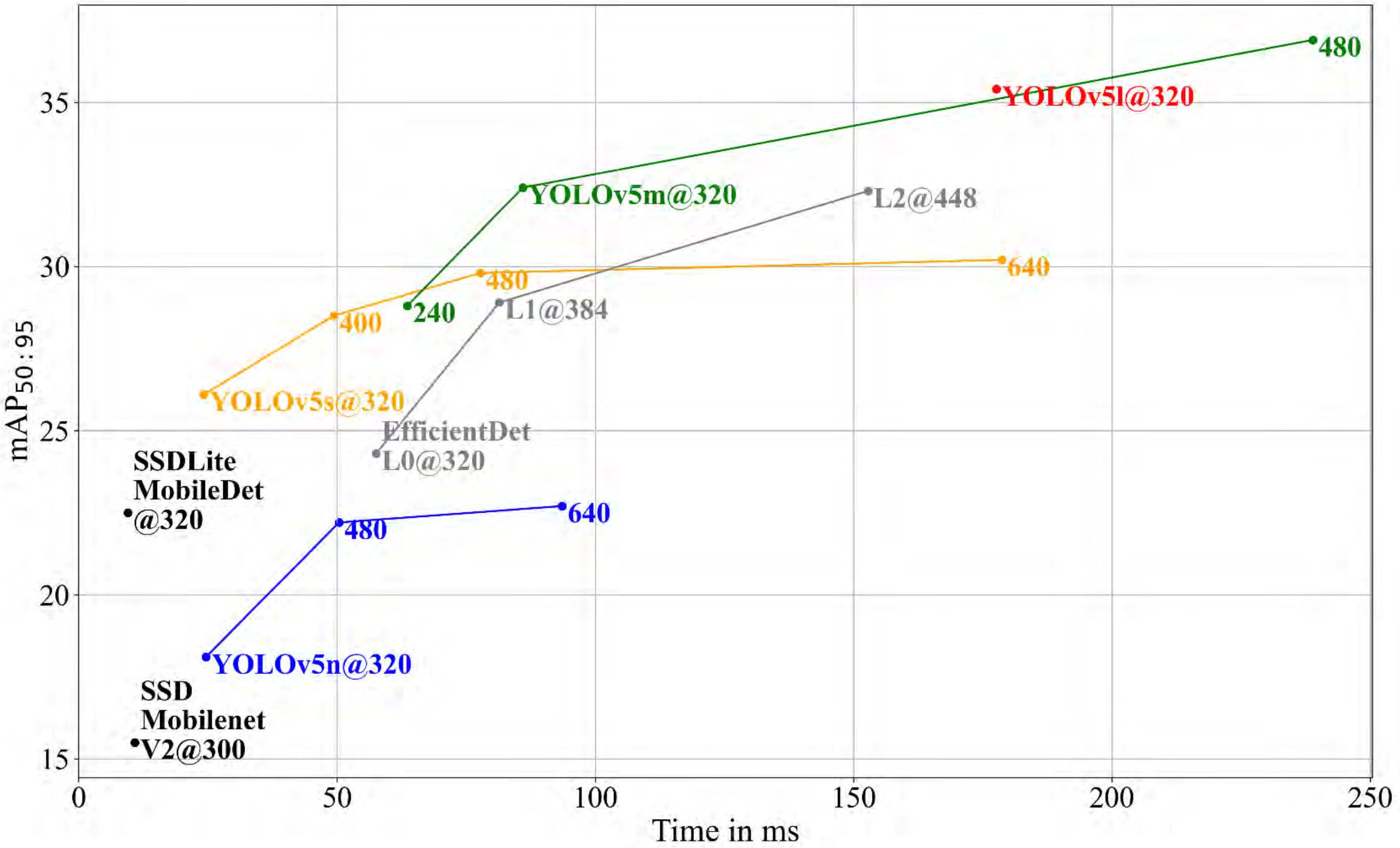
Model Optimizations

- Reduce input size and number of classes
- Split, process and merge tensors (divide and conquer)
- Move incompatible ops to the bottom of the graph
- Transform data flow to avoid reshaping and transposing
- Perform mathematical transformations offline on the weight tensors

Input Size	YOLOv5s (6.1)		YOLOv5s (6.2dev)		Speedup
	TPU/CPU	USB3 Speed	TPU/CPU	USB3 Speed	
320x320	245/40	30.10ms	253/3	24.27ms	19.37%
640x640	225/59	427.48ms	240/16	178.67ms	58.20%

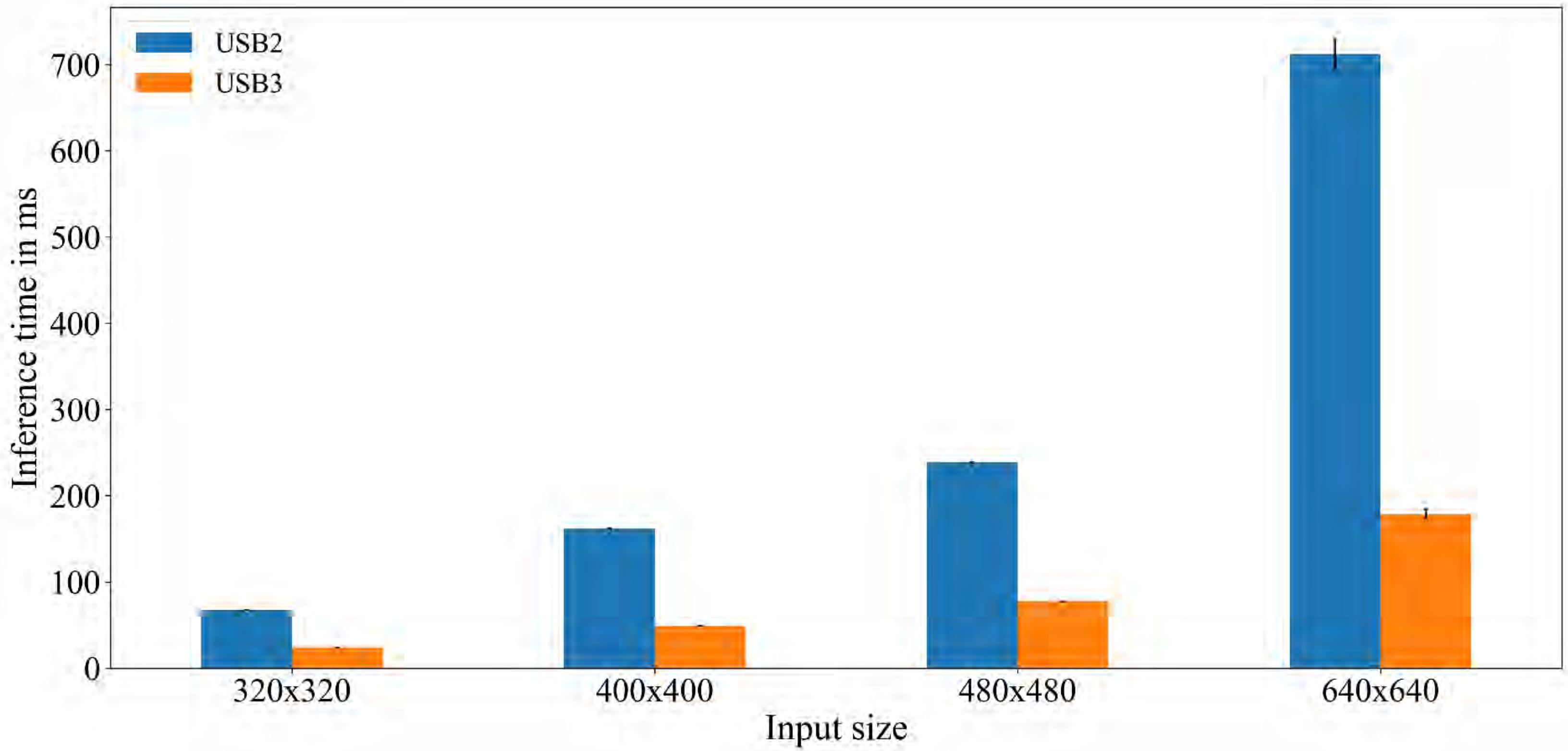
Comparison of YOLOv5s model before and after optimizations

Speed-Accuracy Comparison



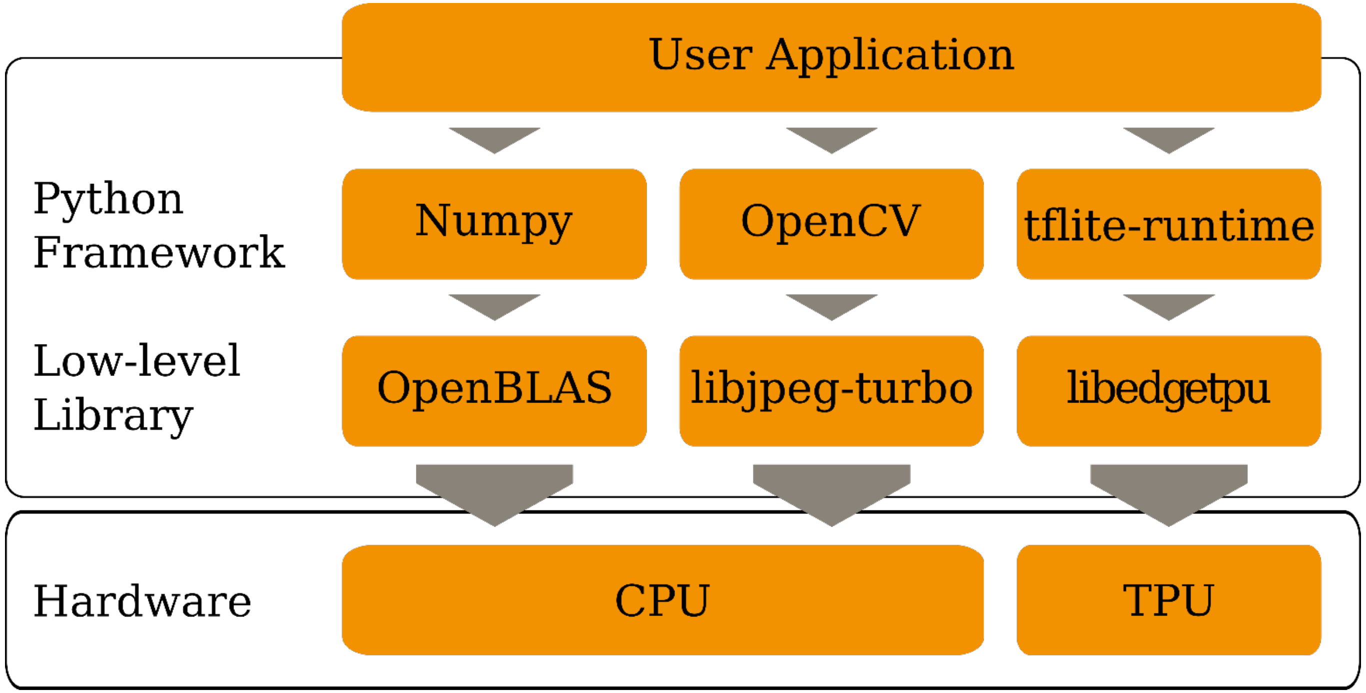
USB3 speed-accuracy comparison of different model types and configurations for edge TPU deployment

USB Performance



YOLOv5s inference speed comparison between USB2 and USB3

Software Stack



Software µStack for edge AI applications

Proof of concept

- STM32MP1 board with ARM Cortex-A7 dual core at 800Mhz + Coral Edge TPU + Webcam
- YOLOv5s model with 320x256 input
- Microservices for inference and visualization
- Raw inference time: ~300ms
- Overall power consumption: ~3.5W
- 10 times speed up with TPU compared to CPU
- Only USB2 and no SIMD acceleration

Summary and Future Work

Efficient Edge Deployment Demonstrated on YOLOv5 and Coral Edge TPU

- Efficient AI at the edge is feasible utilizing novel specialized hardware
- Inference Accelerators require optimized models to expose their full potential
- Data transfer has a significant impact on AI performance
- Optimized low level libraries are the foundation of edge application

Future Considerations

- Reduction of USB2 bottleneck
- Expand model portfolio (e.g., Image Segmentation)
- Consider new low power ARMv8/9 platforms for future deployment with SIMD support
- Evaluate edge TPU performance for MCU applications

Acknowledgment

This work has been financially supported by the AI4DI project. AI4DI receives funding within the Electronic Components and Systems For European Leadership Joint Undertaking (ESCEL JU) in collaboration with the European Union's Horizon 2020 Framework Programme and National Authorities, under grant agreement n° 826060.

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

Bundesministerium
für Bildung
und Forschung

AI4DI

https://ai4di.eu/

Questions?

Upcoming presentations and Topics at EAI4IA 2022:

Monday July 25

16:45-17:00

Preetha Vijayan et al.: *Temporal Delta Layer: Exploiting Temporal Sparsity in Deep Neural Networks for Time-Series Data*

Tuesday July 26 - Topics

- Strategic Vision and Road mapping
- Explainable AI in the Embedded Electronics Industry Environment
- Trustworthy, Dependable AI for Digitizing Industry
- Emerging AI Technologies in Industrial Applications

Ruben Prokscha, Mathias Schneider and Alfred Höß

Department of Electrical Engineering, Media Technology and Computer Sciences,
Ostbayerische Technische Hochschule Amberg-Weiden, 92224 Amberg, Germany
Email: r.prokscha, mat.schneider, a.hoess@oth-aw.de

Related Publication:
Mathias Schneider et al.: **ECBA-MLI: Edge Computing Benchmark Architecture for Machine Learning Inference**

- Development of an edge AI benchmarking software
- Custom build energy monitor hardware
- SOTA object detection models benchmarked
- Model profiles for smart deployment

Event Organisers

The Key Digital Technologies Joint Undertaking - the Public-Private Partnership for research, development and innovation – funds projects for assuring world-class expertise in these key enabling technologies, essential for Europe's competitive leadership in the era of the digital economy. KDT JU is the successor to the ECSEL JU programme. www.kdt-ju.europa.eu

The AI4DI project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the national authorities. www.ai4di.eu

The TEMPO project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Switzerland. www.tempo-ecsel.eu

The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.andante-ai.eu

Thank You

For your attention

@ r.prokscha@oth-aw.de