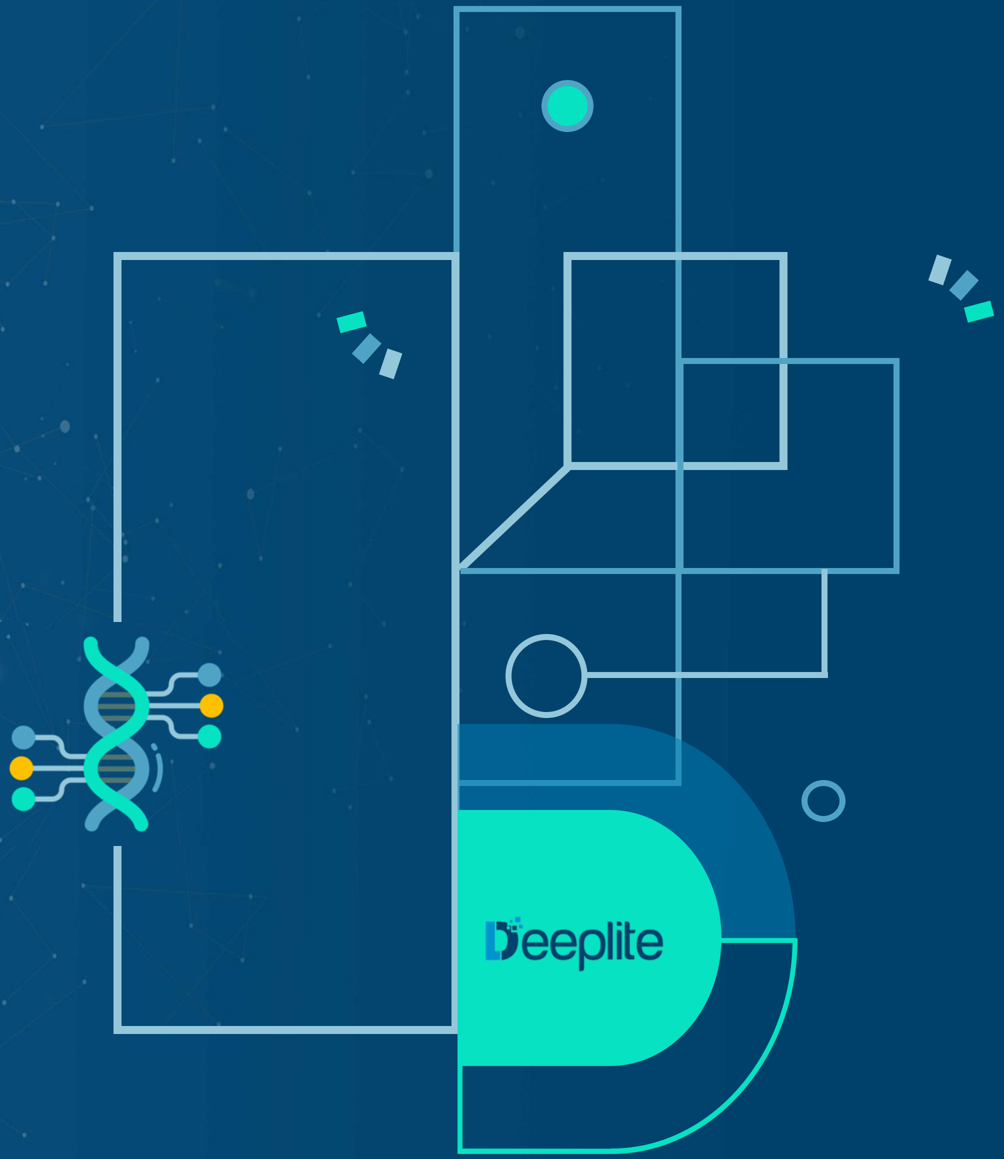
Deeplite

Optimized AI Models for the Edge

MCUBench: A Benchmark of Tiny Object Detectors on MCUs

Sudhakar Sah, Darshan C. Ganji, Matteo Grimaldi, Ravish Kumar, Alexander Hoffman, Honnesh Rohmetra, & Ehsan Saboori

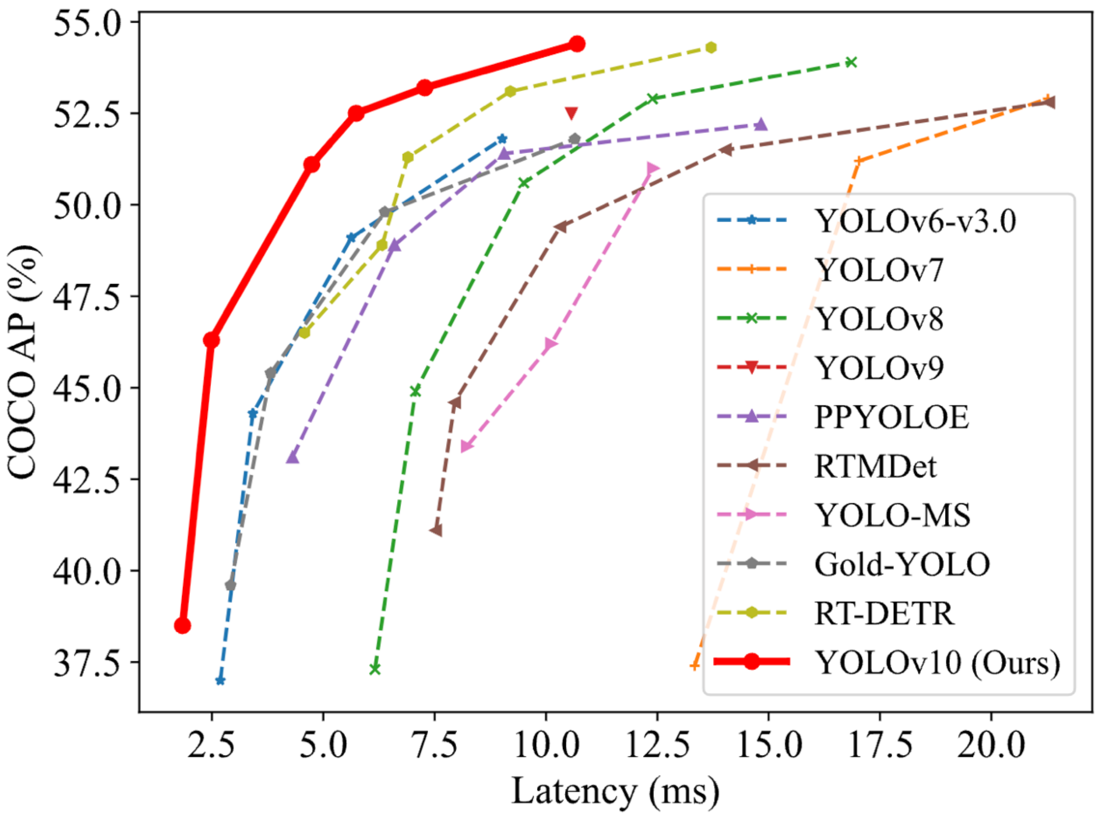
Toronto, Canada
sudhakar@deeplite.ai



CADL Workshop ECCV 2024

Sep 30, 2024

Introduction



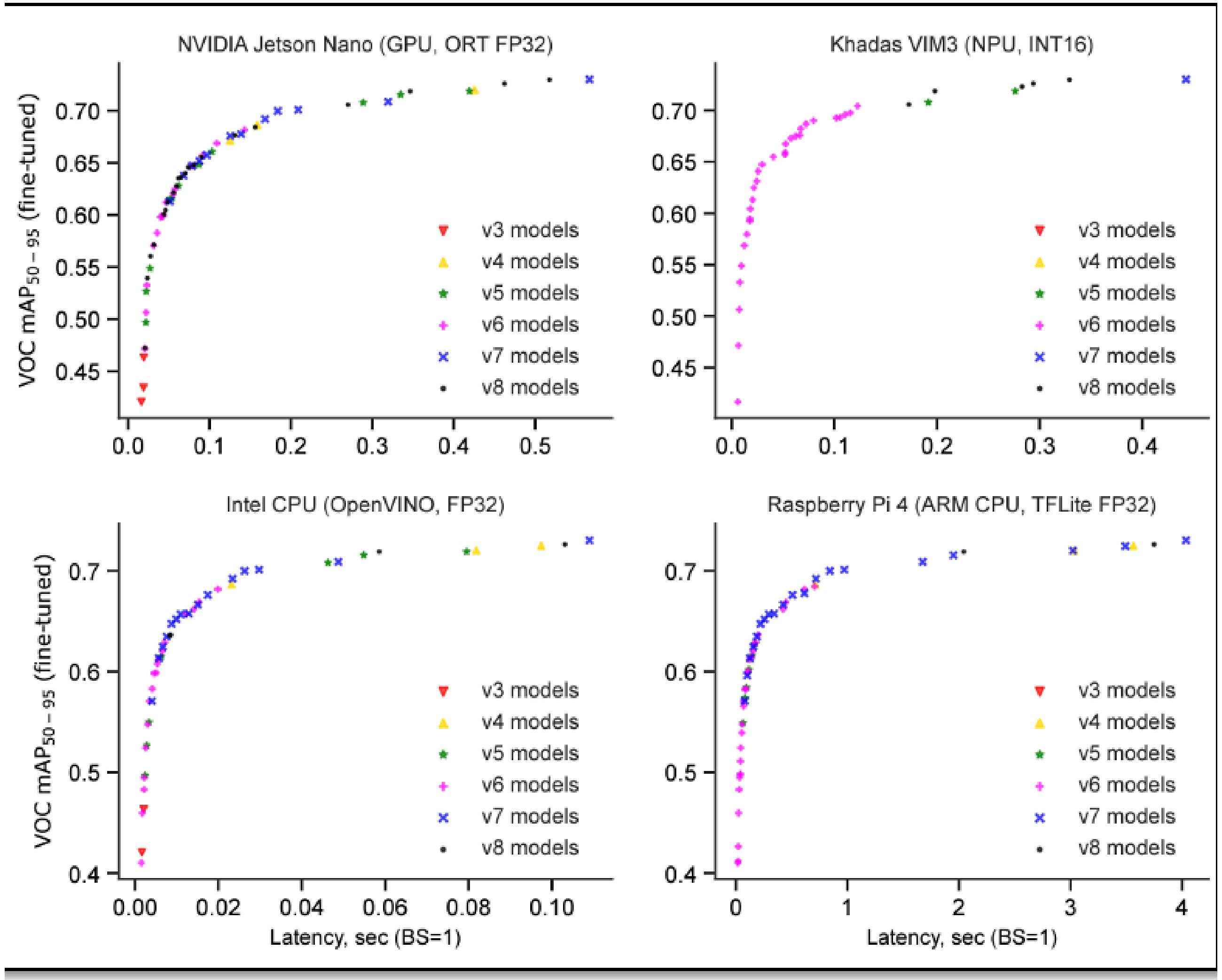
- YOLOv3…YOLOv10 and counting
- New family demonstrate better latency-mAP tradeoff
- Latency-mAP curve holds for GPUs (not for Edge devices)

Motivation

- **Challenge** - Best YOLO model for edge devices given RAM, flash, latency constraint

- **Benchmarking and optimizing AI models** for edge computing can improve the performance and efficiency of IoT devices.

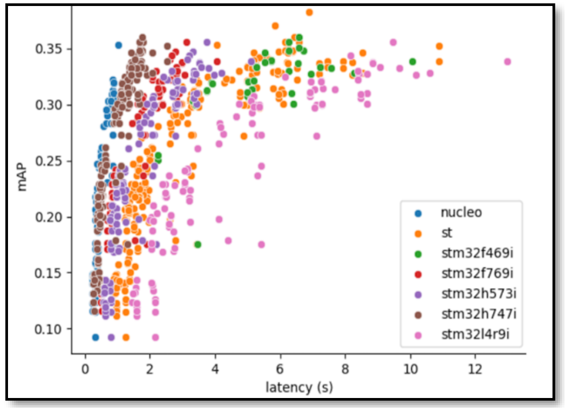
- **MCUBench** is a comprehensive benchmark of over 100 tiny YOLO-based OD models specifically designed for MCU-grade hardware.



Lazarevich, Ivan, et al. "YOLOBench: benchmarking efficient object detectors on embedded systems." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Methodology

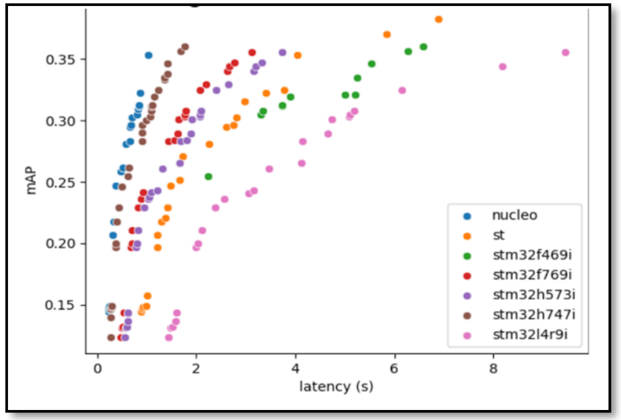
Build Initial Search Space



240 Models

- Width, depth, activation …

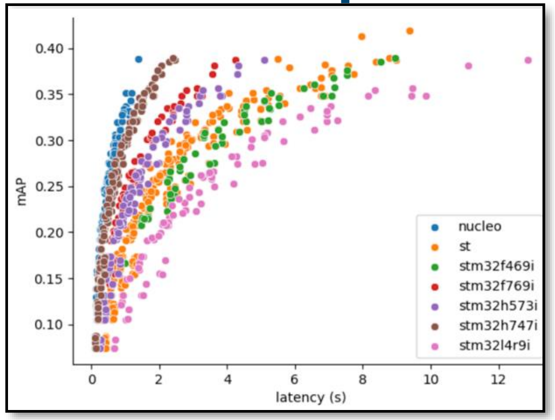
Model Pre-selection



72 Models

- Benchmark on MCUs
- Compute per device pareto models
- Combine pareto models

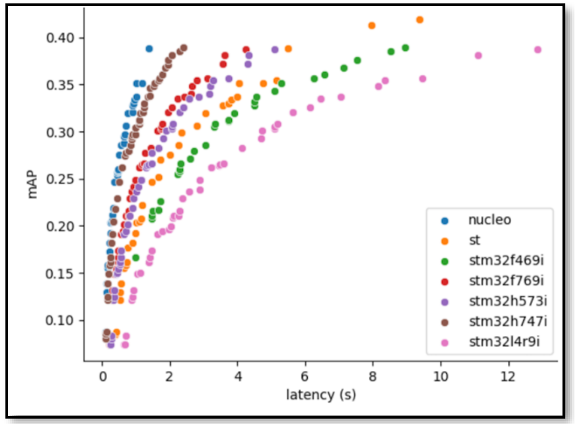
Expand Search Space



288 Models

- Resolution variation
- Benchmark on MCUs

Final Model Ranking

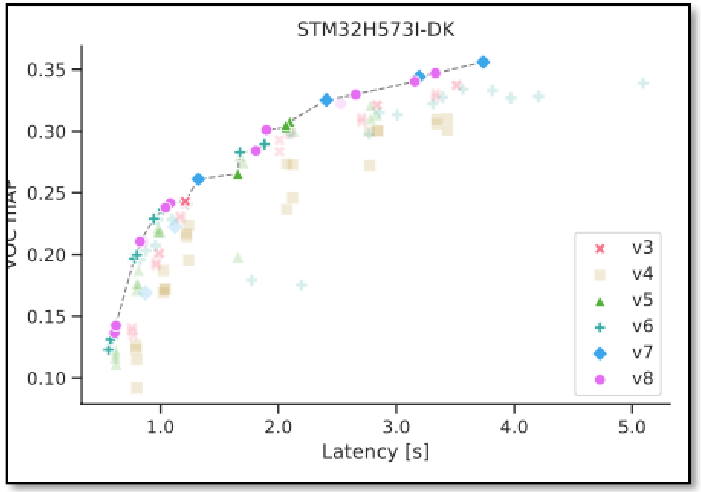


131 Models

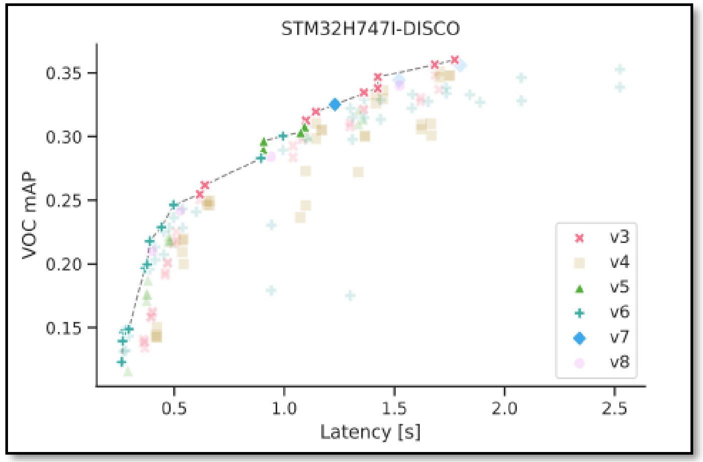
- Compute per device Pareto
- Combine pareto models

Contribution

- **Comprehensive Benchmark**:
  - Over 100 tiny YOLO-based benchmark models
- **Datasets and Analysis**:
  - VOC/COCO,
  - 7 MCUs
- **Fixed Training Pipeline**:
  - Fixed training loop, head,
  - vary backbones, necks, activations, and resolutions.
- **Performance Analysis**
  - V8 models are not always the best models
- **Open Source**
  - Trained weights
  - Hugging Face
  - Deeplite Studio
  - Benchmark values on MCUs



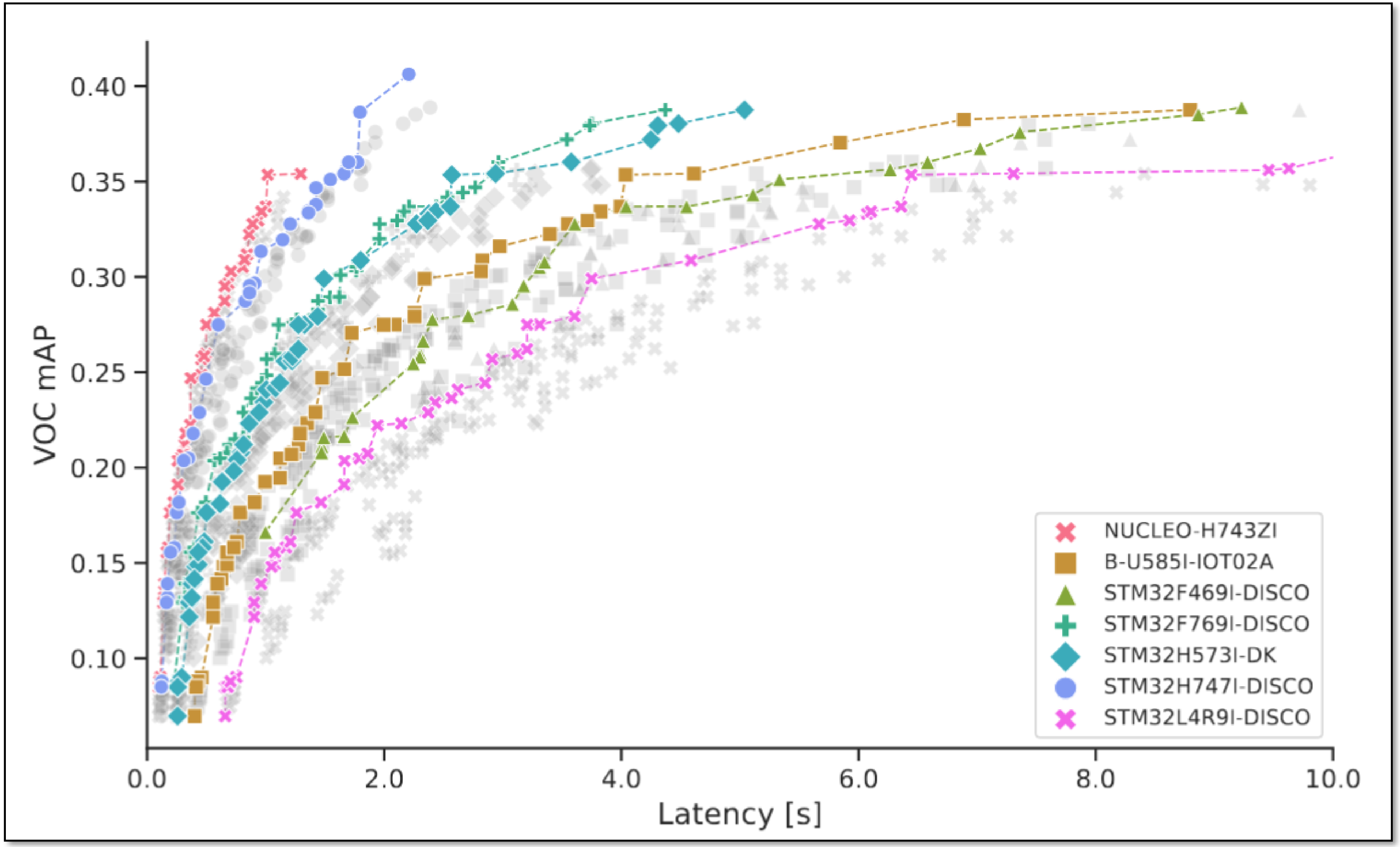
Most of the pareto models are from V8, v7 and v5 family



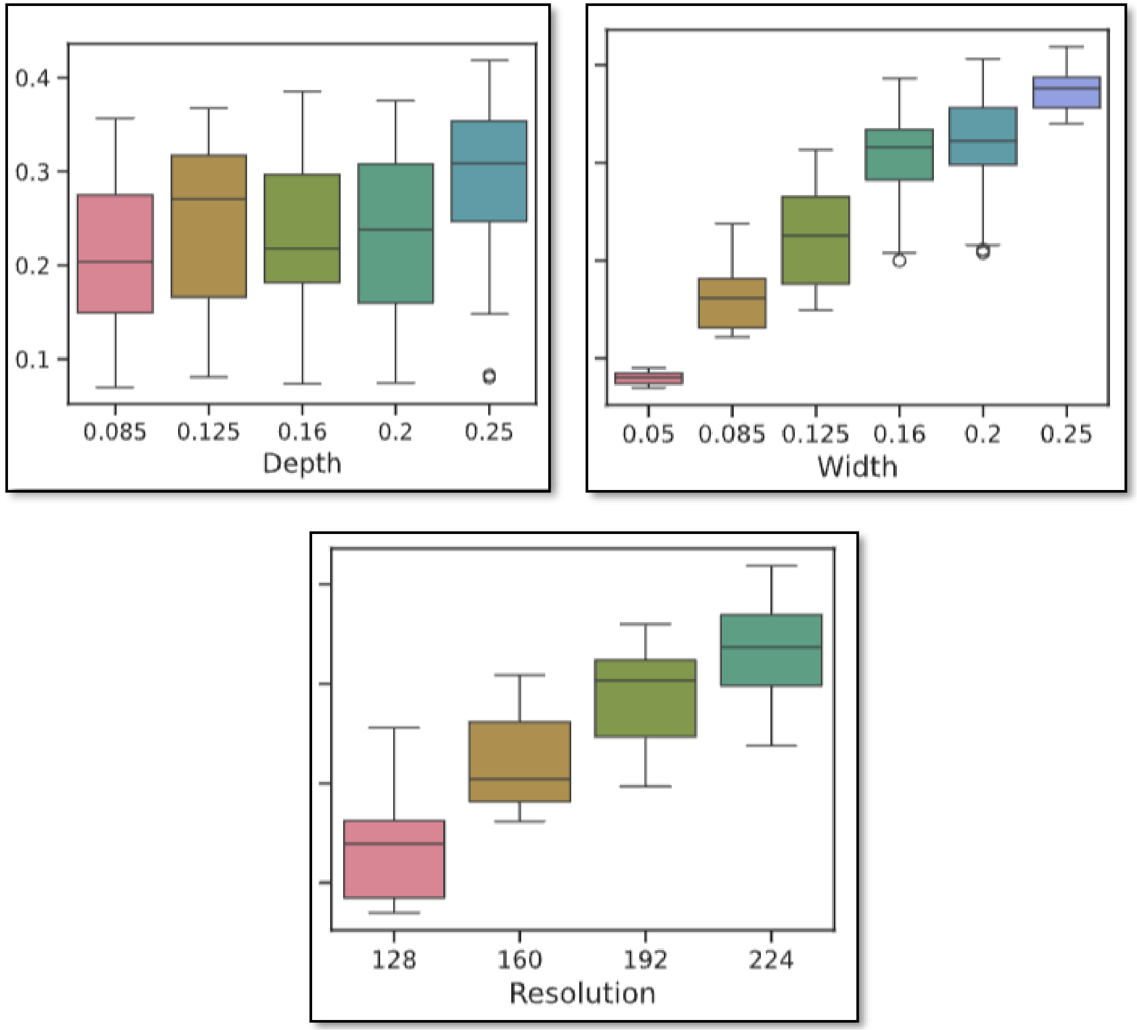
Most of the pareto models are v3 and v6 families

Combined Benchmark

Deeplite



Key Findings



- **Latency-mAP Plot**
  - depends up the hardware
- **Device-Specific paretos**
  - vary drastically
- **Latency Influences**
  - Internal RAM & model complexity..
- **mAP Influencers**
  - resolution and width
- **Depth and Activations**:
  - Variable impact on mAP
  - Activations only affect latency

Conclusion

MCUBench code open source

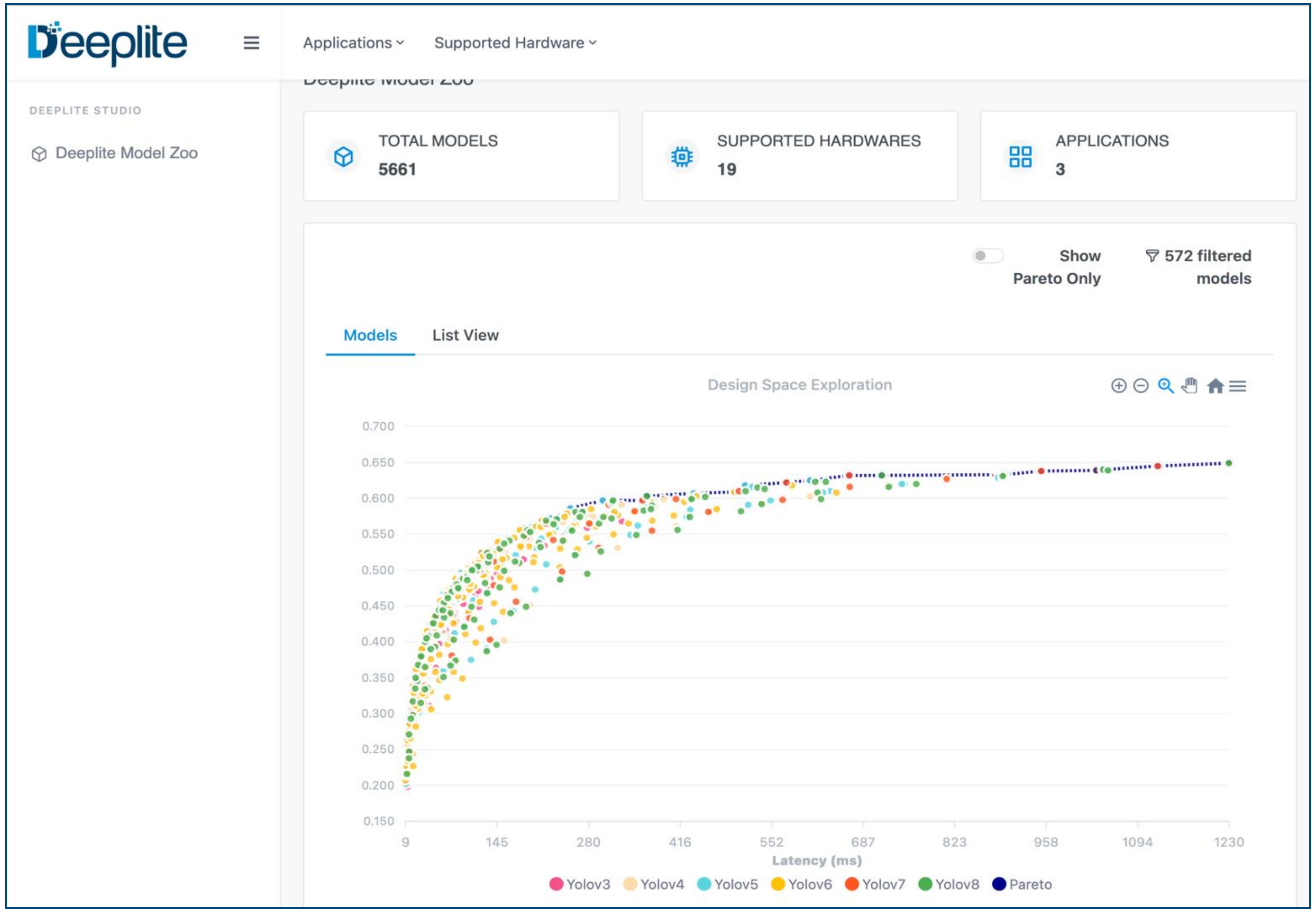
- https://github.com/deeplite/deeplite-torch-zoo/

Model weights (VOCO & COCO)

- Will be Available soon on Deeplite Studio (DLS)

Hugging-Face App

- Will be available soon



TOTAL MODELS 5661

SUPPORTED HARDWARES 19

APPLICATIONS 3

572 filtered models

Deeplite

TORONTO

100 Simcoe St. Suite 115,
Toronto M5H 3G2