Object detection as fast as possible (a.k.a. going yolo on YOLO)

Description: This practical project involves taking the YOLO architecture and applying various compression techniques, including quantization, pruning, and low-rank approximation, to optimize it to the extreme for embedded robotic systems. The goal is to implement and compare compressed models on a real robotic platform (e.g., Raspberry Pi + camera) to test their accuracy and efficiency in object detection. Alternatively, also ViT models can be considered.

Keywords: Model compression, YOLO, Embedded AI

Resources:

https://docs.ultralytics.com/
https://huggingface.co/docs/optimum/en/concept_guides/quantization
https://pytorch.org/docs/stable/quantization.html
https://github.com/NVIDIA/TensorRT
https://onnxruntime.ai/

Required level:
Master student with experience in computer vision, deep learning, and Pytorch.
You should have trained deep learning model before.

Contact: fabmo@dtu.dk