Proposition de sujet de P-SAT – Année 2025-2026

Embedded Computer Vision

Encadrant.e.s au département informatique

Stefan Duffner

Résumé

Contexte scientifique :
State-of-the-art techniques in machine learning, particularly in deep learning for image analysis, are increasingly resource-intensive and energy-consuming. However, there are a variety of methods for reducing the complexity of these models in certain contexts while maintaining their accuracy. This is particularly important for edge computing when access to the cloud is not possible or not relevant, for example in mobile robotics.

Objectifs :
The objective of this P-SAT project is to implement an efficient technical solution for object detection in a video stream using state-of-the-art CNN models and an "intelligent" camera equipped with a small CPU (e.g. Raspberry PI, Arduino, Luxonis OAK).
Different solutions should be compared in terms of precision and complexity/energy consumption.

Méthodologie :
Different solutions exist already for optimising CNN models for object detection in images (quantisation, pruning, distillation). Also, different CNN-based deep learning models exist for object detection (YOLO, SSL, Faster-RCNN). For simplicity, we will treat the videos stream as a sequence of independant images.
The project consists of the following steps:
- Chosing and getting acquainted with the technical environment and pipeline to train and deploy a CNN model for the given hardware/camera.
- Optimising the model using standard techniques (quantisation, pruning, distillation) or novel techniques (specific architectures or ways of training).
- Evaluating and comparing the different implemented solutions in terms of precision and complexity (FLOPS) or estimated energy consumption (W).

Mots-clés

Neural Network compression, computer vision, object detection, embedded AI, frugal AI

Contexte de travail

This project is associated with the LIRIS lab. No other partners are involved. The overall objective is to create a demonstration platform for neural network compression algorithms. There may be interactions with other PhD or post-doc students of the team.

Références

[1] Berthelier A., Château T., Duffner S., Garcia C., Blanc, C. *Deep Model Compression and Architecture Optimization for Embedded Systems: A Survey* Journal of Signal Processing Systems, 2020
[2] https://onnxruntime.ai/docs/tutorials/iot-edge/rasp-pi-cv.html
[3] https://pytorch.org/docs/stable/quantization.html