Master's Thesis / Semester Project:
Improving Network Efficiency for Embedded Platforms.

Computer Vision Lab, D-ITET

Description

Autonomous robotic systems heavily rely on a robust visual perception to understand their environment. One crucial step towards enabling high-level visual understanding is the detection of relevant objects. State of the art deep learning based methods show impressive results in detecting a multitude of objects in diverse environments [3]. Many of these methods [5] are capable of real-time inference on dedicated hardware, such as GPUs and TPUs. In many robotic applications such as the RoboCup Soccer Standard Platform League (SPL), however, this type of hardware is not available and several tasks besides object detection, e.g. planning and locomotion control, must be performed in real-time on a single embedded CPU. Therefore, minimizing the latency of the models used for visual perception is of paramount importance.
In recent years, there have been several works exploring techniques to reduce the size and complexity of deep learning models. Quantization [4] can reduce the computational cost of neural network operators by replacing costly floating point operations with integer-arithmetic ones. Network Pruning [2, 1, 6] aims to reduce a heavy network into a lightweight one by removing redundant parameters.
In this project we aim to leverage these techniques to create accurate and fast object detection models for RoboCup SPL.

Goal

The goal of this project is the development of machine learning models for object detection capable of real-time inference on the embedded CPU of the NAO V6. The first step will be retraining of the current object detection network used by Team NomadZ using quantization-aware training schemes and a comparison of the performance of the quantized model to the current one. Secondly, the impact of the network architecture should be investigated using larger models to understand the trade-off between performance and latency[7]. Finally, network pruning can be used to shrink larger models and make them amenable to inference on the target hardware.
The research aspects include object detection, network compression and evaluation of efficient networks. High-quality domain specific object detection datasets are provided.

Requirements

- Solid grasp of deep learning fundamentals.
- Hands-on experience with PyTorch.
- Basic experience with C++.
- Experience with TinyML and quantization is a plus.

Supervisor:

Jan-Nico Zaech ETF C115 jan-nico.zaech@vision.ee.ethz.ch
Dr. Ajad Chhatkuli ETF D113.2 ajad.chhatkuli@vision.ee.ethz.ch

Professor:

Prof. Luc Van Gool ETF C117 vangool@vision.ee.ethz.ch

References

[1] Xuanyi Dong and Yi Yang. Network pruning via transformable architecture search. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv: Learning*, 2018.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988, 2017.

[4] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

[5] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788, 2016.

[6] Victor Sanh, Thomas Wolf, and Alexander Rush. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems*, 33:20378–20389, 2020.

[7] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.