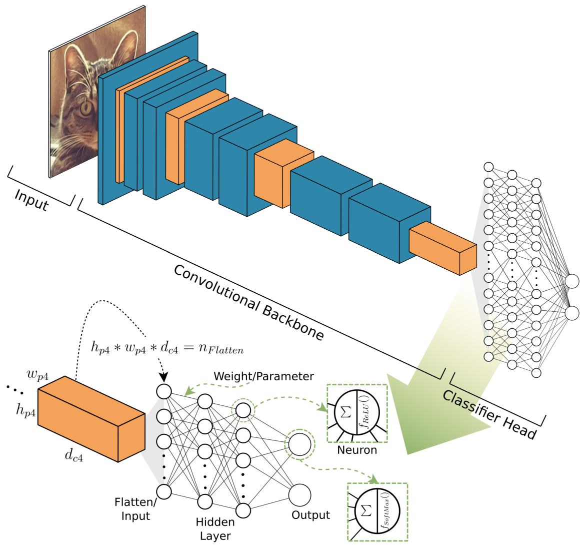
TU WIEN TECHNISCHE UNIVERSITÄT WIEN

Institut für Computertechnik
Institute of Computer Technology

Christian Doppler Forschungsgesellschaft

Master Thesis: Depth-wise Concatenation of Two Pruned Networks

The Embedded Machine Learning (EML) team is part of the Christian Doppler Laboratory and does research on Deep Neural Networks (DNNs) in resource-constrained embedded devices. It studies how energy consumption and resource usage can be minimized while keeping high accuracy. The solution space is characterized by architecture parameters, DNN optimization and transformations, implementation platform configurations, and mapping options. This design space is huge, poorly understood, and rapidly evolving.



A Convolutional Neural Network (CNN) for object detection generally consists of two parts: (1) a feature extractor, also called an encoder, compressing the original image and providing a low dimensional representation of the original image. (2) A classifier (a fully connected layer) and a localizer (a bounding box regressor), providing the object class and the location of the object, respectively. In the literature, state-of-the-art (SOTA) CNNs have been trained for different classification tasks. They are available as commercial-off-the-shelf (CTOS) components ready to be deployed in different applications. However, if more than one CNN is required in the given embedded application, such CTOS SOTA CNNs might not be suitable due to limited available resources.

This thesis project aims to take two CTOS SOTA CNNs trained on two different classification tasks and fuse them into a single CNN with complexity nearly similar to individual CTOS CNN without degrading the accuracy significantly. This thesis project consists of the following steps:

- Select two state-of-the-art CNNs trained on two different tasks, e.g., YOLO-NAS
- Prune these CNNs to 25-50% and fuse them channel-wise via depth-wise convolution
- Evaluate the relative performance of these CNNs compared to original CNNs
- Optimize this network for hardware acceleration

This thesis offers you an excellent opportunity to get into the hot topic of deep learning. It allows you to become an expert in configuring neural networks. Moreover, you acquire critical skills in using neural networks in embedded systems and resource constraints. Some of the M.Sc. projects may be combined with a part-time position. For details, please consult the following:

- Maximilian Götzinger (maximilian.goetzinger@tuwien.ac.at)
- David Breuss (david.breuss@tuwien.ac.at)

MISSION EMBEDDED
MEMBER OF THE FREQUENTIS GROUP