Chair for Integrated Systems
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Forschungspraxis

Graph Neural Network-based Pruning

Convolutional neural networks (CNNs) are the defacto standard for many computer vision (CV) applications. These range from medical technology, robotics applications to autonomous driving. However, most modern CNNs are very memory and compute intensive, particularly when they are dimensioned for complex CV problems.

Compressing neural networks is essential for a variety of real-world applications. Pruning is a widely used technique for reducing the complexity of a neural network by removing redundant and superfluous parameters. One characteristic of this approach is the pruning granularity, which describes the substructures that should be removed from the neural network. Another aspect is the method for finding the redundant and unused structures, which plays a central role in effective pruning without loss of task-related accuracy. The optimization goal determines which elements (kernel, filter, channel) can be removed from the topology of the CNN.

The goal of this work is to learn the internal relationships between the channels, filters, kernels of the layers by means of a graph neural network, and identify their relevance to the classification task of the CNN. The learned relationships are then used for pruning the neural network.

Prerequisites

Prerequisites

To successfully complete this project, you should have the following skills and experiences:

- Good programming skills in Python and Tensorflow
- Good knowledge of neural networks, particularly convolutional neural networks

The student is expected to be highly motivated and independent.

Contact

Nael Fasfous
Department of Electrical and Computer Engineering
Chair of Integrated Systems

Phone: +49.89.289.23858
Building: N1 (Theresienstr. 90)
Room: N2116
Email: nael.fasfous@tum.de

Advisors

Nael Yousef Abdullah Al-Fasfous, Alexander Frickenstein