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

Seminar

In-Train Quantization and Pruning Methods for Convolutional Neural Networks

Convolutional neural network (CNN) compression has become a standard approach of optimization before real-world deployment. Particularly in embedded scenarios, a high-accuracy CNN is usually heavily overparametrized for the target device. The most common techniques to compress a pretrained neural network are quantization and pruning. However, both techniques can also be applied at training-time, saving GPU hours and leading to one-shot training and optimization of CNNs.

Contact

Manoj Vemparala

Email: Manoj-Rohit.Vemparala@bmw.de

Advisors

Manoj Rohit Vemparala, Nael Yousef Abdullah Al-Fasfous