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

Seminar

Accelerating Pruned Neural Networks

Convolutional neural networks (CNNs) have become the state-of-the-art in many computer vision tasks. This has come at the expense of high energy, latency and memory consumption. In recent years, methods of pruning CNNs to increase sparsity and reduce consumption for the deployment on mobile devices have become more popular. While these pruning methods improve the performance of the network, further speed up and reduced consumption can be achieved by improved zero detection, power gating and intelligent network-on-chip design. In this seminar, different pruning methods will be surveyed and compared in their accuracy, latency and compression rate. Additionally, different CNN acceleration designs for weight pruning will be investigated.

Contact

Manoj Vemparala

Email: Manoj-Rohit.Vemparala@bmw.de

Advisors

Manoj Rohit Vemparala, Nael Yousef Abdullah Al-Fasfous