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

Seminar

Optimized CUDA Kernels for Efficient Processing of Binary and Sparse Neural Networks

Going beyond the efficient inference libraries provided by NVIDIA, researchers have started developing customized kernels, which better exploit their applied neural network compression techniques. In this seminar, these computation kernels will be studied and analyzed for the benefits they could offer to binary and sparse neural network execution.

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

Manoj Rohit Vemparala, Nael Yousef Abdullah Al-Fasfous