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

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

Description

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 Master thesis, these computation kernels will be studied, analyzed, and implemented to extract the benefits of binary and sparse neural network execution on inference GPUs.

Requirements

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

- Very good programming skills in C/C++
- Experience with CUDA programming
- Good knowledge of neural networks, particularly convolutional neural networks

The student is expected to be highly motivated and independent. By completing this project, you will be able to:

- Understand the impact of sparsity and binarization on inference GPUs
- Analyze to effect of runtime neural network inference approximation
- Evaluate trade-offs between redundancy, structured parallelism and approximate computing

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