Master Thesis

Efficient Implementation of Partial Operator Folding for Low-memory Implementation of Artificial Neural Networks on Embedded Microcontrollers

A previous M.Sc project [Hajer, Chebil TUM 2020] has shown the potential for “partial folding” of convolutional neural network operations (interleaving their execution using sliding-window buffering schemes) to allow memory-efficient execution on embedded microcontrollers. The project aims to develop and refine this to achieve an efficient, practically realizable, implementation architecture compatible with industry-standard neural-network C++ deployment frameworks.

The project will comprise:

- Surveying the current state-of-the-art using the previous student’s work as a starting point.
- Design and proof-of-concept implementation of a backward-compatible extension to the architecture of the TF-lite(micro) embedded neural network framework to add support partially folded execution of operations.
- Design and implementation of reference implementations of convolution operators in the extended framework sufficient to allow testing of one or two existing benchmark applications.
- Use of instruction-set simulators to estimate the practical performance/memory/power impact of using the extended framework the benchmark applications on representative microcontroller CPUs.

If you are interested please contact Rafael Stahl (r.stahl@tum.de; Room 2922).