RISC-V Pooling Plug-In for Neural Networks (Bachelor's thesis / Master's thesis)

Scope

A recent trend in artificial intelligence (AI) is to bring neural networks (NN) to the Internet of Things (IoT) and extend the reach of deep learning applications to platforms that are heavily resource and energy constrained. Due to a lack of optimized accelerators, today's ultralow resource restricted IoT applications rely on cloud computing still, supported by powerful and resource intensive high performance AI processing units. Nevertheless the market for AI powered low end edge devices thrives. Since the number of devices grows rapidly and AI related network traffic explodes, not to mention the growing data privacy concerns, processing data locally to the device became key factor to make AI powered IoT a success.

Objective

This thesis addresses the search for a specialized hardware accelerator for supporting common NN's pooling layers. These layers make use of different pooling algorithms spatially applied as filters on tensors. Finding an efficient mapping to combinatorial logic with respect to area, power and performance for accelerating these kind of layer operations is the challenge. For design space exploration Infineon's automated code generation framework will be used to generate different solutions. The impact of the implementation of the pooling operation on the quality of the inference shall be elaborated and analyzed concerning their performance impact (execution cycles of software) and their area impact (gate count of chip/FPGA).

Contact:

Sebastian Prebeck:	Sebastian.Prebeck@infineon.com
Wolfgang Ecker:	Wolfgang.Ecker@infineon.com
Daniel Müller-Gritschneder:	daniel.mueller@tum.de