Master's thesis: Hardware-Aware Layer Fusion of Deep Neural Networks

Dataflow and mapping of Convolutional Neural Networks (CNN) influences their compute and energy efficiency on edge accelerators. Layer fusion is a concept which enables the processing of multiple CNN layers without resorting to costly off-chip memory accesses. In order to optimally implement layer fusion, different combinations of mapping and scheduling parameters need to be explored. We, at the BMW group, offer you a challenging master thesis position that aims to optimize the fusion strategy of a given CNN workload for maximal data reuse and resource utilization.

What awaits you?

- Identifying challenges in deploying state-of-the-art neural networks on resource-constrained embedded hardware.
- Implementation of state-of-the-art layer fusion and efficient resource partitioning techniques on a neural network accelerator.
- Experience in applying neural network optimization techniques, such as pruning and quantization.
- Engagement in a diverse team with experience in publishing at international peer-reviewed conferences.
- Presentation of the thesis results using the scientific method, both in written and oral form.

Please note that you must ensure that the thesis is supervised by a university.

What should you bring along?

- Strong knowledge in computer vision concepts, and convolutional neural networks.
- Hands-on experience with Xilinx FPGAs, Verilog/VHDL/HLS.
- Excellent programming skills in C, Python. Experience in Tensorflow 2, Git, Docker is a plus.
- Highly motivated and eager to collaborate in a team.
- Ability to speak and write in English fluently.

Are you passionate about new technologies and an innovative environment? Apply now!

Earliest starting date: Immediately

Duration: 6 months
Working hours: full-time

Contact:
Shambhavi.balamuthu-sampath@bmw.de

==================================================================