Non-Linear Weight Expansion for Neural Networks (Master's thesis)

Scope

Artificial Intelligence (AI) subtle supports people in increasing number of ordinary tasks. In many applications people don't even recognized the contribution of AI. This becomes possible due to low latency on device neural network processing. Post decade dealing with neural networks on small edge devices was inconceivable, due to resource limitations and hardware cost. Now, due to more efficient and powerful hardware components, this previous storyteller's material becomes real. Nevertheless networks have to be heavily optimized to fit the restricted hardware resources edge devices can provide due to cost and power consumption, whereas hardware has to be optimized to fit the needs of optimized networks. So to say a HW/SW/AI codesign. One of the main cost driver is memory. To reduce cost utilizing memory more efficiently is key.

Objective

This thesis addresses efficient neural network weight compression/expansion techniques. Different approaches of compression shall be investigated and evaluated on their potential for mapping to hardware and the associated accuracy loss. The focus is on non-linear compression/expansion methods using partial linear mapping functions or lookup tables. Thereby precise mapping of high entropy ranges should be enhanced, whereas for low entropy ranges lower precision might be sufficient. Infineon's automated code generation framework will be used to implement and further generate different instances of the hardware expansion/decompression unit. The impact of the implementation on the quality of the inference shall be elaborated and analyzed concerning their accuracy impact (loss) and their area impact (gate count of chip/FPGA and memory footprint of the inference parameters).

Contact:

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