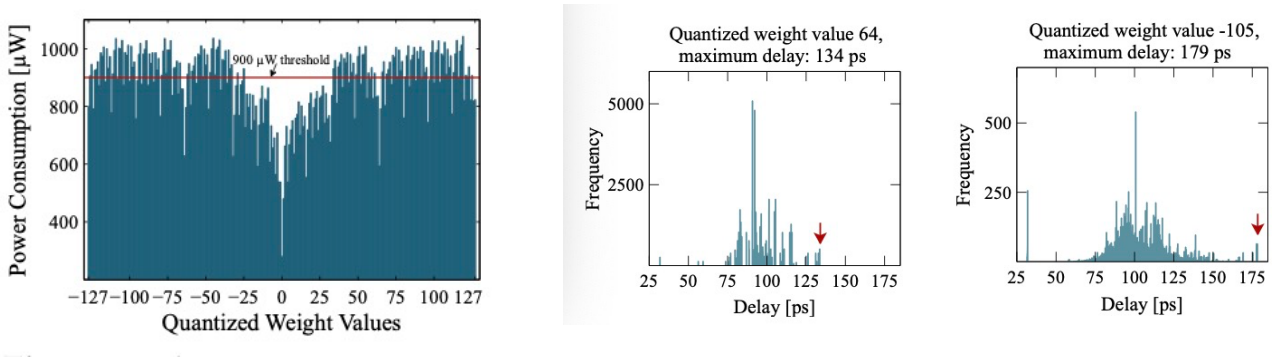
Layer-wise Selection of Weights and Activations for Power-Efficient Neural Network Acceleration

In this master thesis, the power consumption and timing characteristics of weights and input activations of neural networks will be analyzed, as shown in the following figure. Different weight values exhibit different power consumption and timing profiles. Besides, activations also have similar properties. Accordingly, such differences can be exploited to reduce the power consumption of digital accelerators for neural networks.

In this thesis, first, the weights and activations according to their hardware performance will be ranked. Second, the layer in a neural network that consumes the most power consumption will be identified. Third, this network will be trained with only weights and activations with good power and timing efficiency. Fourth, the remaining layers will also be trained in a similar way. In the last step, the power reduction will be evaluated.



Left figure: Average power consumption of quantized weight values. Right figure: Delay profiles of a MAC unit for two quantized weight values. The arrows point to the maximum delay of a given weight value with respect to all the activation transitions.

If you are interested in this topic for master thesis, please contact:

Prof. Dr.-Ing. Li Zhang (grace.zhang@tu-darmstadt.de) with your CV and transcripts.