Neural Network Evaluation and Enhancement Considering Quantization

Neural networks as shown in Figure 1 have successfully been applied to solve complex problems such as speech/image processing. To improve computing accuracy, the depth of neural networks has steadily increased significantly, leading to deep neural networks (DNNs). The increasing complexity has put massive demands on computing power and triggered intensive research on hardware acceleration for neuromorphic computing in recent years.

The computation function at a neuron in a neural network can be considered as an incompletely specified truth table. The known entries in such a table are determined by the training data. Since training data are usually a small subset of all entries in the truth table, we need to estimate the other entries to realize the logic design with the truth table. The concept of this technique is illustrated in Figure 2, where the neural network is quantized so that the inputs and the outputs of neurons are represented by binary values.

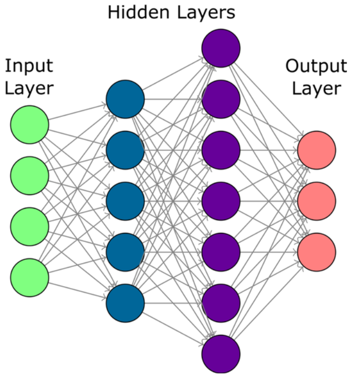
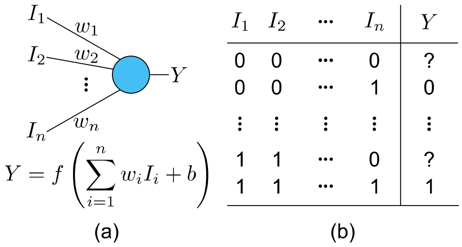


Figure 1: Neural network with multiple layers.

In the quantization, fewer bits reduce resource usage, but the accuracy of the computation may also degrades. In this thesis, a balance between hardware resource and computation accuracy of neural networks will be explored. To compensate the accuracy degradation caused by quantization, the structures of neural networks may also be modified together with quantization to achieve an overall good area efficiency and computation accuracy. The major tasks of this thesis may include:

- Quantize inputs and outputs of neurons into different number of bits to evaluate the relation between the number of bits and the accuracy of the neural networks.
- Modification of neural network structures for a tradeoff between accuracy, number of quantization bits and required number of computation operations.



$$Y = f\left(\sum_{i=1}^{n} w_i I_i + b\right)$$

(a)

I_1	I_2	···	I_n	Y
0	0	···	0	?
0	0	···	1	0
⋮	⋮	⋮	⋮	⋮
1	1	···	0	?
1	1	···	1	1

(b)

Figure 2: A neuron and its truth table from training data. (a) A neuron with n inputs and one output. (b) The

If you are interested in this topic for master thesis, please contact:
Dr.-Ing. Li Zhang (grace-li.zhang@tum.de) with your CV and transcripts.