Block-wise Training for Systolic Arrays of Digital Neural Network Accelerators

In recent years, deep neural networks (DNNs) have been widely applied in various fields, e.g., image/speech recognition. In DNNs, there are a large number of multiply-accumulate (MAC) operations. To accelerate MAC operations in DNNs, systolic arrays are introduced as an attractive platform due to their high degree of concurrent computation and high data reuse rate. Recently, various state of the art hardware accelerators using systolic arrays or properties of systolic arrays have been proposed. TPU is the most well-known accelerator based on systolic arrays. Systolic arrays have a regular structure where Processing Elements (PEs) are replicated and connected together to process data in a pipelined fashion. Figure 1 shows the structure of the systolic array. However, weights of neural networks after unstructured pruning usually exhibit irregular patterns, as shown in Figure 2. Implementing MAC operations with such irregular weight patterns on systolic arrays with regular designs, might result in an underutilization of hardware resources.

In this master thesis, a block-wise neural network training method will be explored to fully exploit the benefits of systolic arrays.

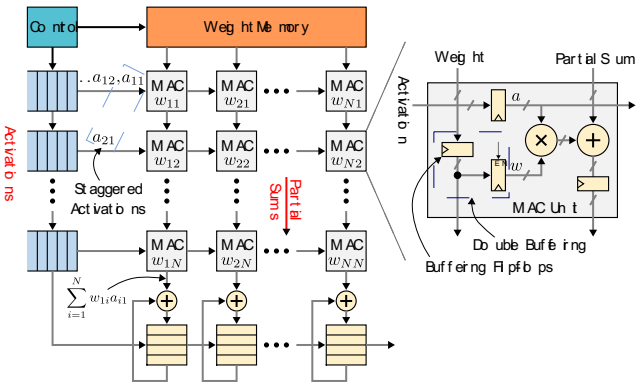

Figure1: The structure of systolic array.

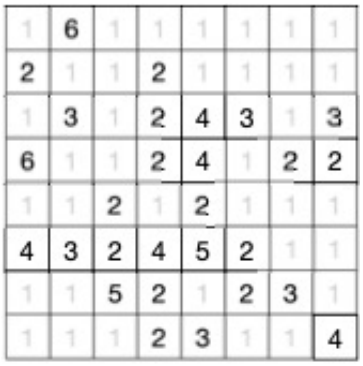

Figure2: Weights after unstructured pruning.

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.