Pipelined architecture for sparse DNNs

Analysis of a pipelined architecture for sparse DNNs on embedded systems

Deep neural networks (DNN) are increasing their presence in a wide range of applications, and their computationally-intensive and memory-demanding nature poses challenges, especially on embedded systems, where performance and energy are tightly constrained. Pruning techniques turn DNN models into sparse by setting most weights to zero, offering optimization opportunities if specific support is included. Exploiting sparsity yields remarkable speedups and energy savings but also an area overhead. We propose a pipelined architecture that avoids all useless operations in the inference process on DNNs. It has been implemented in a Xilinx UltraScale+ FPGA, where performance, energy efficiency and area were characterized. We also propose an assessment methodology where sparse and dense architectures are similar in area. The benefits of exploiting sparsity depend not only on the sparsity but also on the arithmetic precision. Our architecture is clearly superior on 32-bit arithmetic or highly-sparse networks. However, on 8-bit arithmetic or networks with low sparsity the benefits are smaller and it is more profitable deploying a dense architecture with more arithmetic resources than including support for sparsity. We consider that FPGAs are the natural target for DNN sparse accelerators since they can be loaded at run-time with the best-fitting accelerator.