Pipelined architecture for sparse DNNs

This repository contains the VHDL code of the accelerator described in the paper "Analysis of a Pipelined Architecture for Sparse DNNs on Embedded Systems", available on [https://doi.org/10.1109/TVLSI.2020.3005451](https://doi.org/10.1109/TVLSI.2020.3005451)

Analysis of a pipelined architecture for sparse DNNs on embedded systems

Deep neural networks (DNNs) are increasing their presence in a wide range of applications, and their computationally intensive and memory-demanding nature poses challenges, especially for embedded systems. Pruning techniques turn DNN models into sparse by setting most weights to zero, offering optimization opportunities if specific support is included. We propose a novel pipelined architecture for DNNs that avoids all useless operations during the inference process. It has been implemented in a field-programmable gate array (FPGA), and the performance, energy efficiency, and area have been characterized. Exploiting sparsity yields remarkable speedups but also produces area overheads. We have evaluated this tradeoff in order to identify in which scenarios it is better to use that area to exploit sparsity, or to include more computational resources in a conventional DNN architecture. We have also explored different arithmetic bitwidths. Our sparse architecture is clearly superior on 32-bit arithmetic or highly sparse networks. However, on 8-bit arithmetic or networks with low sparsity it is more profitable to deploy 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.