Towards Further Compression of Low-Bitwidth DNNs with Permuted Diagonal Matrices
Mohsenin, Tinoosh, University of Maryland Baltimore County

We propose Cyclic Sparsely Connected (CSC) layers,with memory/computation complexity order of O(NlogN), that can be used as an overlay for fully connected (FC) layers whose number of parameters,O(N2), can dominate the parameters of the entire DNN model. The FPGA and ASIC hardware implementation results in 65-nm CMOS show the proposed CSC hardware outperforms the conventional pruned architecture with an equal compression rate by 2× in power, energy, area and resource utilization when running at the same frequency. Compared to previous work, this design is 37×to 55×higher energy efficient.