Binary Artificial Neural Network

Name:Akshay Kamalapuram Sridhar
Unityid: akkamala
StudentID:. 200272844

Delay (ns to run provided example). Clock period: 4.25ns # cycles": 7	Logic Area: (um^2) 206.15 Memory: N/A	1/(delay.area) ($ns^{-1}.um^{-2}$) 0.00114137336
Delay (TA provided example. TA to complete)		1/(delay.area) (TA)

Abstract

Binary Convolutional Neural Networks can significantly reduce the size of memory storage along with reducing the number of computational operations. In this project, a fixed size single stage of a binary convolutional neural network is implemented in hardware using Verilog. The 16 bit data read out of the Input Matrix SRAM is read as 4 convolution matrices and each of those matrices is xnor'ed with the Weight Kernal matrix read out of the Weight SRAM. The number of ones the xnor output is then calculated and used to determine the result for that specific convolution. The hardware was optimized to a clock period of 4.25ns, with a logic area of 206.15 um^2 and computed the final result in 7 clock cycles.

Binary Artificial Neural Network

Akshay Kamalapuram Sridhar

1. Introduction

The hardware designed in this project is aimed to replicate a single stage of a binary convolutional neural network for a 4x4 input matrix with a 3x3 weight kernel. All weights and input data are binary and take on values of -1 and 1, represented as 0 and 1 respectively. The inputs are read from the input SRAM as a 16 bit binary value and the weights are read from the weight matrix as a 16 bit binary value. The 4x4 final computed result is stored as a 16 bit value, with upper bits appended with 0, and stored in the Output SRAM.

The values read from the input matrix is sub-divided into 9 bits representing each convolution matrix. The specific set of bits used for computation is based on the control signal received from the controller. These are then xnor'ed with the values from the weight matrix and the bitcount of this output is used to write the result for the specific computation in the final output.

The final optimized hardware has a clock period of 4.25 ns and performs the entire computation in 7 cycles. It occupies a logic area off 206.15 um^2.The rest of the report dives further into the design, RTL, methods used for verification and the final results obtained

2. Micro-Architecture

The hardware utilizes a separate controller and datapath in design to implement the RTL in an efficient manner. The control signals for the controller are the select_line which is uses to select 1 out of the 4 convolutions, write_enable which goes high when the final output is ready to be written into the output SRAM, and busy to indicate that a computation is currently being performed. The hardware begins computation when a go signal goes high and gets resetted when reset_b is low

3. Verification

The design was verified using the waveforms in Verilog and using the test values provided along with the project.

4. Results Achieved

The final optimized hardware has a clock period of 4.25 ns and performs the entire computation in 7 cycles. It occupies a logic area off 206.15 um^2.

5. Conclusions

The project successfully implements a fixed size single stage of a binary convolutional neural network is implemented in hardware using Verilog. An efficient design is generated in the form of a netlist from the RTL code. The hardware generated as a very minimal area and satisfies slack to indicate a good synthesizable design.