Hardware Accelerated CNN for Digit Recognition

Liam Bishop, Dan Cooke, Felix Hanau, Ryan Kennedy and Richard Mouradian
We used a very compact and simple CNN model featuring 2 convolution layers, 2 pooling layers and 1 Fully Connected layer.
The input to the network was a 28x28 gray scale image.
The output is a probability vector.
# Model Size and Parameters

The input to the CNN is a grayscale 28x28 image.

<table>
<thead>
<tr>
<th>Layer Number</th>
<th>Type</th>
<th>Kernel Size</th>
<th>Output Size</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Convolution</td>
<td>6 5x5 (150)</td>
<td>6x24x24 (3456)</td>
<td>156</td>
</tr>
<tr>
<td>2</td>
<td>Avg Pooling</td>
<td>6 2x2 (24)</td>
<td>6x12x12 (864)</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>Convolution</td>
<td>6 groups of 12 5x5 (1800)</td>
<td>12x8x8 (768)</td>
<td>1812</td>
</tr>
<tr>
<td>4</td>
<td>Avg Pooling</td>
<td>12 2x2 (48)</td>
<td>12x4x4 (192)</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>Fully Connected</td>
<td>None</td>
<td>50</td>
<td>9610</td>
</tr>
</tbody>
</table>
Software Implementation

- Implementation of CNN written in C to:
  - Verify functionality of configuration
  - Implemented using 16-bit fixed-point approximation of trained weights
  - Debug output of hardware implementation
    - Also used verilator to generate waveforms and compare results of different layers
- Trained model (weight and biases) with Keras
- 98% accuracy when tested with recognizing handwritten digits
Block Diagram

- **Software:**
  - Sends image data for processing
  - Controls hardware by telling it which layer to compute and waits for hardware to return acknowledgement

- **Hardware:**
  - 4 main Sections
    - MAC and After MAC
    - Average Pooling
    - Memory
    - Control
  - Data stream is looped through modules to perform different layer operations of CNN
Software Driver Interface

- 14 Registers
  - Control Input
  - Control Output
  - Input address
  - Input data
  - Output registers (10)
- Control states
MAC and After MAC

- Used to compute Convolution layers and Fully Connected Layer
- MAC:
  - Convolution:
    - 24 multiply and accumulate can be performed simultaneously.
  - Fully Connected:
    - All 10 outputs are computed simultaneously
- After MAC:
  - Processes layers:
    - Convolution:
      - Performs ReLU, adds bias to outputs of MACs, and shift outputs for proper scaling
    - Fully Connected:
      - Shifts outputs from MAC for proper scaling
Memory

- Stores input image, layer weights, and between layer values
  - Originally Quartus Mega-Wizard Memory blocks but changed to implied memory for ease of use and debugging
  - Redundant memory blocks used to allow for more than 2 accesses at a time
  - Each section of memory controlled by read and write counters

Pooling

- Used to compute Pooling Layer outputs
  - Average of the 4 inputs

<table>
<thead>
<tr>
<th>Layer</th>
<th>Data (Bits)</th>
<th>Weights + Bias (Bits)</th>
<th>Data Mem Blocks</th>
<th>Parameter Mem Blocks</th>
<th>Memory Needed (Bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>28x28x16</td>
<td>0</td>
<td>2</td>
<td>N/A</td>
<td>12544</td>
</tr>
<tr>
<td>1</td>
<td>6x24x12x16</td>
<td>6x5x5x16 + 6x16</td>
<td>24</td>
<td>3</td>
<td>168384</td>
</tr>
<tr>
<td>2</td>
<td>6x12x12x16</td>
<td>0</td>
<td>12</td>
<td>N/A</td>
<td>13824</td>
</tr>
<tr>
<td>3</td>
<td>12x8x8x16</td>
<td>12x5x5x16 + 12x16</td>
<td>24</td>
<td>12</td>
<td>53568</td>
</tr>
<tr>
<td>4</td>
<td>12x4x4x16</td>
<td>0</td>
<td>12</td>
<td>N/A</td>
<td>3072</td>
</tr>
<tr>
<td>5</td>
<td>50x16</td>
<td>10x192x16</td>
<td>N/A</td>
<td>5</td>
<td>3242</td>
</tr>
<tr>
<td>Total:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>230634</td>
</tr>
</tbody>
</table>
BENCHMARKING AND RESULTS

- Software implementation runtime: 7.71 ms
- Hardware Implementation runtime (Verilator): 0.205 ms
  - Capable of processing 4878 images per second
- 37 times faster than software
- FPGA resource Utilization
  - Total block memory bits: 371,328 (9%)
  - Total DSP blocks: 24 (28%)
  - Total Registers: 2534