An Energy-Efficient Bit-Split-and-Combination Systolic Accelerator for NAS-Based Multi-Precision Convolution Neural Networks

presented by Liuyao Dai

Authors:
Liuyao Dai, Quan Cheng, Yuhang Wang, Gengbin Huang, Junzhuo Zhou, Kai Li, Wei Mao, and Hao Yu

Southern University of Science and Technology
Shenzhen, China
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
Deep learning for edge computing: 
convolutional neural network (CNN) models are becoming more complex with larger parameters

Neural Architecture Search (NAS) can search for optimized multi-precision neural network models:
• negligible loss of accuracy
• energy efficiency

Fig. 1 Illustration of energy aware neural architecture search framework
Existing multi-precision multiply-accumulate (MAC) disadvantages:

- Bottom-up low-precision-combination (LPC) large hardware cost huge power consumption

- Top-down high-precision-split (HPS) poor throughput performance

Proposed Work:

- Bit-split-and-combination (BSC) method tradeoff cost and throughput

- Multi-precision systolic dataflow data reuse and energy efficient

Fig. 2 Typical NAS flow with proposed multi-precision systolic architecture
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
Bit-Split-and-Combination MAC

\[
A = \sum_{i=0}^{n-1} 4^i \times a_i 
\]  \hspace{1cm} (1)

\[
B = \sum_{j=0}^{m-1} 4^j \times b_j 
\]  \hspace{1cm} (2)

\[
A \times B = \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} 4^{i+j} \times a_i b_j 
\]  \hspace{1cm} (3)

Fig. 3 The mathematically characteristics of multi-precision multiplication operation for BSC method
Fig. 4 BSC MACs for multi-precision operations including 8-bit, 4-bit and 2-bit modes
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
BitBrick(BB) and Multi-Precision BSC PE

<table>
<thead>
<tr>
<th>Parameter</th>
<th>3×3</th>
<th>5×5</th>
<th>7×7</th>
</tr>
</thead>
<tbody>
<tr>
<td>size of the convolution (K)</td>
<td>3</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>number of updated input activations (N)</td>
<td>15</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>calculation cycles (J)</td>
<td>45</td>
<td>50</td>
<td>49</td>
</tr>
<tr>
<td>number of convolution kernel groups (L)</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1 Parameters of the computation in BitBrick

- BB contains a 4-bit width MAC, corresponding storage registers and a shuffler.
- 4 BBs form a PE which performs multi-precision computations.
Systolic PE Array

- Weights are stored in PE array fetched from weight buffer.
- Input activations flow from PE$_1$ to PE$_{256}$ sequentially.
- Outputs from PE array are transmitted to psum buffer.

Fig. 7 Multi-precision systolic dataflow of PE array
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
**systolic dataflow**

<table>
<thead>
<tr>
<th>Shape Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>numbers of 3D filters</td>
</tr>
<tr>
<td>C</td>
<td>numbers of ifmap/filter channels</td>
</tr>
<tr>
<td>H</td>
<td>ifmap width/height</td>
</tr>
<tr>
<td>R</td>
<td>filter width/height</td>
</tr>
<tr>
<td>E</td>
<td>ofmap width/height</td>
</tr>
</tbody>
</table>

Table 2 Parameters of a conv layer

Fig. 8 Computation of a CONV layer

Fig. 9 Processing of the systolic dataflow

Liuyao Dai / Southern University of Science and Technology
Data reuse in the systolic dataflow

Fig. 10 Data reuse in the processing of the systolic dataflow

- Input activations from different input channels are input to the same PE in different cycles
- Different filters are stored in different PEs

Fig. 11 Data reuse in BSC BB

- Input activation is reused $K$ times in BSC BB
- Weight is reused $E^2$ times in BSC BB

Liuyao Dai / Southern University of Science and Technology
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
MAC Performance Comparison

<table>
<thead>
<tr>
<th>Precision</th>
<th>HPS(TOPS/W)</th>
<th>LPC(TOPS/W)</th>
<th>BSC(TOPS/W)</th>
<th>BSC/HPS</th>
<th>BSC/LPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 bit</td>
<td>12.6</td>
<td>32.7</td>
<td>27.36</td>
<td>2.4</td>
<td>0.82</td>
</tr>
<tr>
<td>4 bit</td>
<td>6.29</td>
<td>8.16</td>
<td>13.68</td>
<td>2.4</td>
<td>1.64</td>
</tr>
<tr>
<td>8 bit</td>
<td>3.14</td>
<td>2.04</td>
<td>3.42</td>
<td>1.2</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Table 3 Energy efficient comparison of multi-precision mac units

Fig. 12 Performance metrics comparison between traditional HPS, LPC and the BSC MAC units

- the proposed BSC MAC unit in this work has the characteristics of low power consumption, low hardware cost and fast calculation speed.

Liuyao Dai / Southern University of Science and Technology
### System Performance Comparison

<table>
<thead>
<tr>
<th>CNN</th>
<th>Dataset</th>
<th>Model Weights</th>
<th>2bit/4bit/8bit proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16</td>
<td>CIFAR-10</td>
<td>138.0 MBytes</td>
<td>0%/89.8%/10.2%</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>ImageNet</td>
<td>13.0 MBytes</td>
<td>0%/94.5%/5.5%</td>
</tr>
<tr>
<td>LeNet-5</td>
<td>MNIST</td>
<td>0.5 MBytes</td>
<td>45.0%/55.0%/0%</td>
</tr>
</tbody>
</table>

Table 4 Evaluated CNN benchmarks

<table>
<thead>
<tr>
<th>Technology</th>
<th>Gemmini</th>
<th>Bit-serial</th>
<th>Bit-fusion</th>
<th>BSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cores</td>
<td>256 PEs</td>
<td>4096 SIPs</td>
<td>512 fusion units</td>
<td>256 PEs</td>
</tr>
<tr>
<td>On-chip (Memory)</td>
<td>64 KB</td>
<td>2 MB eDRAM</td>
<td>181.5 KB</td>
<td>180 KB</td>
</tr>
<tr>
<td></td>
<td>16 KB SRAM</td>
<td>/</td>
<td>/</td>
<td>1.43</td>
</tr>
<tr>
<td>Chip area (mm(^2))</td>
<td>0.467</td>
<td>1.40</td>
<td>/</td>
<td>1.43</td>
</tr>
<tr>
<td>Frequency</td>
<td>500 MHZ</td>
<td>980 MHZ</td>
<td>500 MHZ</td>
<td>500 MHZ</td>
</tr>
</tbody>
</table>

Table 5 Evaluated accelerators
Fig. 13 Improvement ratios of the BSC systolic accelerator to Gemmini, Bit-fusion and Bit-serial on multi-precision CNN benchmarks: (a) Area efficiency, (b) Energy efficiency.

- **Area efficiency performance:** Compared with Bit-fusion, at most $1.43 \times$ ratios are achieved owing to the heavy additional logics of Bit-fusion.
- **Energy efficiency performance:** Compared with Gemmini and Bit-serial, at most $1.85 \times$ and $6.38 \times$ ratios are achieved owing to the proposed work supporting both input activations and weights for multi-precision operations.
the proposed accelerator provides flexibility in both inputs and weights bit-width, leading to the highest reduction ratio by $6.38 \times$, $1.85 \times$ and $1.65 \times$ compared with Bit-serial, Gemmini and Bit-fusion.

Fig. 14 Power breakdown of Gemmini, Bit-serial, Bit-fusion and the proposed BSC accelerator under different benchmarks: (a) VGG-16, (b) ResNet-18, (c) LeNet-5.
Outline

• Introduction
• Bit-Split-and-Combination MAC
• Multi-precision systolic accelerator
• Systolic dataflow
• Experiment
• Conclusion
In this paper, an energy-efficient multi-precision systolic accelerator is designed

• The reconfigurable architecture supports NAS-based CNNs with 2-8 various bit-widths.

• Compared with the state-of-the-art accelerators Gemmini, Bit-serial and Bit-fusion on the multi-precision CNN benchmarks, the proposed BSC accelerator achieves at least $1.18 \times$, $5.37 \times$ and $1.46 \times$ energy efficiency.

• For area efficiency, the improvement ratio of $1.2 \times$, $2.7 \times$ and $1.14 \times$ are achieved at least.

• The results show the proposed work is of great potential for multi-precision edge-computing
Thanks for your attention!