Deep Compression and EIE: Efficient Inference Engine on Compressed Deep Neural Network

Song Han*, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark Horowitz, Bill Dally

Stanford University
Our Prior Work: Deep Compression

- Memory reference is expensive.
- Small DNN models are critical.

[1]. Han et al. NIPS 2015

Our network parameters are stored using compressed sparse row (CSR) format, which requires

\[ \text{Compression Ratio} = \frac{\text{Original Size}}{\text{Compressed Size}} \]

Table: Compression Results

<table>
<thead>
<tr>
<th>Network</th>
<th>Original Size</th>
<th>Compressed Size</th>
<th>Compression Ratio</th>
<th>Original Accuracy</th>
<th>Compressed Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet</td>
<td>240MB</td>
<td>6.9MB</td>
<td>35x</td>
<td>80.27%</td>
<td>80.30%</td>
</tr>
<tr>
<td>VGGNet</td>
<td>550MB</td>
<td>11.3MB</td>
<td>49x</td>
<td>88.68%</td>
<td>89.09%</td>
</tr>
<tr>
<td>GoogleNet</td>
<td>28MB</td>
<td>2.8MB</td>
<td>10x</td>
<td>88.90%</td>
<td>88.92%</td>
</tr>
<tr>
<td>SqueezeNet</td>
<td>4.8MB</td>
<td>0.47MB</td>
<td>10x</td>
<td>80.32%</td>
<td>80.35%</td>
</tr>
</tbody>
</table>
EIE: First Accelerator for Sparse DNN

- Deep Compression solves the model size problem.
- But it creates another problem: irregular computation pattern.
- CPU/GPU are only good at dense linear algebra.
- So we create EIE that supports: static-sparse M, dynamic-sparse V, indirect indexing, weight sharing.

- **Sparse Matrix**
  - 90% *static* sparsity in the weights,
  - 10x less computation,
  - 5x less memory footprint

- **Sparse Vector**
  - 70% *dynamic* sparsity in the activation
  - 3x less computation

- **Weight Sharing**
  - 4bits weights
  - 8x less memory footprint

Fully fits in SRAM
- 120x less energy than DRAM

Savings are multiplicative: 5x3x8x120=14,400 theoretical energy improvement.

Dally. NIPS tutorial 2015; Han et al. ISCA 2016
EIE: First Accelerator for Sparse DNN

Dally. NIPS tutorial 2015; Han et al. ISCA 2016
Compared to CPU and GPU:
189x and 13x faster
24,000x and 3,400x more energy efficient
Beyond EIE: a Multi-Dimension Sparse Recipe for Deep Learning

Faster Speed: EIE accelerator

Higher Accuracy: DSD regularization

Smaller Size: Deep Compression, SqueezeNet++

[1]. Han et al. “Learning both Weights and Connections for Efficient Neural Networks”, NIPS 2015
[5]. Iandola, Han, et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size”, arXiv 16
[6]. Yao, Han, et.al, “Hardware-friendly convolutional neural network with even-number filter size”, ICLR workshop 2016