HighLight:
Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity

Yannan Nellie Wu\textsuperscript{1}, \underline{Po-An Tsai}\textsuperscript{2}, Saurav Muralidharan\textsuperscript{2}, Angshuman Parashar\textsuperscript{2}, Vivienne Sze\textsuperscript{1}, Joel S. Emer\textsuperscript{1,2}

\textsuperscript{1}MIT, \textsuperscript{2}NVIDIA

http://emze.csail.mit.edu/highlight
Many Applications Involve DNNs

- Natural Language Processing
- Autonomous Navigation
- Medical Imaging & Diagnostics

data and computation intensive
subject to prediction accuracy & latency requirements

In great need of optimizations and accelerations
Different DNN Optimizations Introduce Different Sparsity

Optimizations to Reduce Model Size

- **Pruning Techniques**
  - *Han, NeurIPS15*
  - Introduces Sparse Weights

- **Depth-wise Separable Layers**
  - *Howard, CVPR17*
  - Introduces Dense Weights
Different DNN Optimizations Introduce Different Sparsity

Optimizations to Reduce Model Size
- **Pruning Techniques**
  [Han, NeurIPS15]
- **Depth-wise Separable Layers**
  [Howard, CVPR17]

Optimizations to Improve Accuracy
- **Activation Functions**
  [Apicella, NN21]
- **Attention-based Modules**
  [Vaswani, NeurIPS17]

Modern DNNs can weights and activations that are either dense or sparse with various sparsity degrees.
High-Level Opportunities for Sparse DNNs

Zero Values Can be Compressed Away

Ineffectual Operations Can be Eliminated

Important to design sparse DNN accelerators to exploit such opportunities

\[ x \times 0 = 0 \]
\[ x + 0 = x \]
Requirements for an Ideal Sparse DNN Accelerator

Flexible

exploit many sparsity degrees
Requirements for an Ideal Sparse DNN Accelerator

Flexible

exploit many sparsity degrees

Efficient

low sparsity tax for eliminating ineffectual operations
Existing Works Do Not Meet Such Requirements

Unstructured Sparse Accelerators

Dual-Side Sparse Tensor Core (DSTC)
[Wang, ISCA21]

 STRUCTURED SPARSE ACCELERATORS

Sparsity Degree Spectrum

- High Sparsity Tax
  + Flexible

Structured Sparse Accelerators

NVIDIA Sparse Tensor Core (STC)
[NVIDIA, TechReport20]

Sparsity Degree Spectrum

0% sparse (dense)
+ Low Sparsity Tax
- Inflexible
Naïve Way to Increase Flexibility Structured Sparse Designs

Extend the Number of $G:H$ Ratios Supported

Sparsity Degree Spectrum

- 0% sparse (dense)
- 50% sparse
- 67% sparse
- 75% sparse

Not Scalable
Sparsity tax increases approximately in proportion to the number of sparsity degrees
Our Proposal

Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity
Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

N-Rank HSS: G:H $\rightarrow$ G:H $\rightarrow$ G:H

$\text{Rank N-1}$ $\rightarrow$ $\text{Rank N-2}$ $\rightarrow$ $\text{Rank 0}$

What does a $3:4 \rightarrow 2:4$ pattern look like?

Dense Vector
Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

What does a \( 3:4 \to 2:4 \) pattern look like?

Rank1: 3 nonempty blocks out of the 4 blocks

Vector with Rank1 Sparsity Applied
Hierarchical Structured Sparsity (HSS)

Compose G:H sparsity patterns in a hierarchical fashion

What does a 3:4→2:4 pattern look like?

Rank1: 3 nonempty blocks out of the 4 blocks
Rank0: 2 nonzero values out of 4 values within the block

Vector with Both Ranks’ Sparsity Applied
 Hierarchical Structured Sparsity (HSS)

DNN Workloads Often Have Tensors with Multiple Dimensions

<p>| | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Per-Row 3:4-2:4 Tensor

HSS can be applied to an arbitrary dimension in a multi-dimensional tensor
HSS Introduces A Flexible Way to Express Sparsity Degrees

4 sparsity degrees

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>4:4</th>
<th>4:5</th>
<th>4:6</th>
<th>4:7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0%)</td>
<td>(20%)</td>
<td>(33%)</td>
<td>(43%)</td>
</tr>
</tbody>
</table>

Sparsity Degree Spectrum

0% 20% 33% 43%
HSS Introduces A Flexible Way to Express Sparsity Degrees

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sparsity degrees</strong></td>
<td><strong>Sparsity degrees</strong></td>
</tr>
<tr>
<td>4:4  (0%)</td>
<td>4:4  (0%)</td>
</tr>
<tr>
<td>4:5  (20%)</td>
<td>2:4  (50%)</td>
</tr>
<tr>
<td>4:6  (33%)</td>
<td>1:4  (75%)</td>
</tr>
<tr>
<td>4:7  (43%)</td>
<td></td>
</tr>
</tbody>
</table>

Sparsity Degree Spectrum
HSS Introduces A Flexible Way to Express Sparsity Degrees

**4 sparsity degrees**

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:4 (0%)</td>
<td>4:4 (0%)</td>
</tr>
<tr>
<td>4:5 (20%)</td>
<td>2:4 (50%)</td>
</tr>
<tr>
<td>4:6 (33%)</td>
<td>4:6 (33%)</td>
</tr>
<tr>
<td>4:7 (43%)</td>
<td>1:4 (75%)</td>
</tr>
</tbody>
</table>

**Sparsity Degree Spectrum**

- **4:5-2:4 (60%)**

**Multiplication of Fractions**
HSS Introduces A Flexible Way to Express Sparsity Degrees

Sparsity Degree Spectrum

Multiplication of Fractions

Rank 1

4:4 (0%) 4:5 (20%) 4:6 (33%) 4:7 (43%)

Rank 0

4:4 (0%) 2:4 (50%) 1:4 (75%)

4:5-2:4 (60%)

4:6-1:4 (83%)
HSS Introduces A Flexible Way to Express Sparsity Degrees

<table>
<thead>
<tr>
<th>sparsity degrees</th>
<th>Rank 1</th>
<th>Rank 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>4:4</td>
<td>4:5 (20%)</td>
<td>4:4 (0%)</td>
</tr>
<tr>
<td>4:5</td>
<td>4:6 (33%)</td>
<td>2:4 (50%)</td>
</tr>
<tr>
<td>4:6</td>
<td>4:7 (43%)</td>
<td>1:4 (75%)</td>
</tr>
<tr>
<td>4:7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fraction multiplication allows flexible representation of many sparsity degrees in a wide range

Sparsity Degree Spectrum

12 sparsity degrees

0% 20% 33% 43% 50% 60% 67% 71% 75% 80% 83% 86%
HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

Example Accelerator Architecture Organization

Architecture Level 0

Architecture Level 1
HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

Example Accelerator Architecture Organization
HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks

Example Accelerator Architecture Organization

Each level only needs to accelerate for a few sparsity degrees
HSS Enables Modularized Acceleration

Modularity of HSS allows different architecture levels to accelerate for different HSS ranks.

Simple Acceleration at Each Architecture Level Leads to Low Hardware Overhead

Efficient Processing with Low Sparsity Tax

Each level only needs to accelerate for a few sparsity degrees.
HighLight: Flexible and Efficient Sparse DNN Accelerator

Hierarchical Skipping
Reduce Latency and Energy Consumption

HSS-based Compressed Representation Format
Reduce Storage Requirement and Energy Consumption

HighLight translates two-rank HSS into reduction in latency and energy
Experimental Results
We Compare HighLight with Representative Designs

dense  structured sparse  unstructured sparse  structured sparse

TC  STC  DSTC  S2TA  HighLight
We evaluate the designs with synthetic workloads with different sparsity degrees ranging from 0%-75%.
We evaluate the designs with synthetic workloads with different sparsity degrees ranging from 0%-75%

HighLight is efficient across evaluated metrics
We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any).
We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any).
We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any).
We evaluate the designs with representative DNNs pruned to different sparsity degrees, each with its respective sparsity structure (if any).

HighLight sits on the accuracy-energy delay product pareto frontier.
More Details in Paper!

• How to systematically represent the diverse sparsity patterns in DNNs?
  – Short answer: sparsity specification via fibertree abstraction.

• What does HighLight’s energy and area sparsity tax breakdowns look like?
  – Short answer: low sparsity tax as HighLight independently accelerates simple sparsity patterns at different architecture levels.

• ...
Summary

Hierarchical Structured Sparsity (HSS)

- Composed of multiple levels of simple sparsity patterns
- Allows flexible expression of diverse sparsity degrees

HighLight Accelerator

- Supports two-rank HSS for a few degrees at each level
- Implements low-overhead support for each rank at different architecture levels
- Ensures both efficiency and flexibility

Acknowledgement: NVIDIA Research, MIT AI Hardware Program

http://emze.csail.mit.edu/highlight