Achieving Lossless Accuracy with Lossy Programming for Efficient Neural-Network Training on NVM-Based Systems

Wei-Chen Wang, Yuan-Hao Chang, Tei-Wei Kuo, Chien-Chung Ho, Yu-Ming Chang, and Hung-Sheng Chang

1 Macronix International Co., Ltd., Taiwan
2 Academia Sinica, Taiwan
3 City University of Hong Kong, Hong Kong
4 National Taiwan University, Taiwan
5 National Chung Cheng University, Taiwan

Why NVM-Based Systems for NN Training?

- Dilemma of Big Data and Neural-Network Training on DRAM:
  - Size of Intermediate Data
  - Size of Weights
  - Size of Biases

<table>
<thead>
<tr>
<th></th>
<th>VGGNet-D</th>
<th>ResNet</th>
<th>GoogLeNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Data</td>
<td>ImageNet (224x224x3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mini-batch Size</td>
<td>256</td>
<td>256</td>
<td>1024</td>
</tr>
<tr>
<td>Size of Intermediate Data</td>
<td>15 GB</td>
<td>54 GB</td>
<td>24 GB</td>
</tr>
<tr>
<td>Size of Weights</td>
<td>528 MB</td>
<td>230 MB</td>
<td>51 MB</td>
</tr>
<tr>
<td>Size of Biases</td>
<td>52 KB</td>
<td>71 KB</td>
<td>8 KB</td>
</tr>
</tbody>
</table>

- NVM-Based (PCM-Based) Training Solutions:
  - Density, Unit Cost, Endurance, Read Latency, Nearby-Zero Leakage Power
  - Write Latency (Asymmetric Write), Wear Out Problem

- Lossy Programming (Lossy-SET)

- How to Explore a Great Programming Scheme with the Considerations of Performance, Retention Time, Endurance of NVM and Accuracy of NN

Viewpoint of Data Content

- Weights and Biases
- Intermediate Data

Viewpoint of Data Flow

- High-Performance Computing Platform (e.g., GPU)
- Low-Performance Computing Platform (e.g., CPU)

Data-Aware Programming Design (DAP)

- Layer-Aware SET Policy:
  - Viewpoint of Data Flow
  - Intermediate Data

- Bit-Aware Dual-SET Policy:
  - Viewpoint of Data Content - Weights and Biases

Buffered Marching-Based Wear Leveling:

- To Solve the Asymmetric Damage of Different Data

Experiment Results:

- Remarkably improve the average memory access latency up to 4.3x and enhance the lifetime of PCM to 3.4x