Across-stack Profiling and Characterization of State-of-the-art Machine Learning Models on GPU

Motivation

- ML performance is impacted by the interplay between frameworks, system libraries, compilers, and hardware platforms.
- There is lack of tools that allow inspection of ML model performance across the HW/SW stack and researchers have to switch between tools and manually stitch the outputs.
- We propose an across-stack profiling design and integrated it with MLModelScope -- a hardware/software agnostic platform for evaluating and benchmarking ML models at scale.
- We coupled the profiling capabilities with automatic analyses that reveal insights which can not be obtained easily through other tools or method.
- We characterized the model/layer/GPU kernel performance of several state-of-the-art models.
- Results for all models are available at mlmodelscope-scl9.netlify.com

Cross-Stack Profiling

- Model profile: the time spent running the inference for C API (TF_SessionRun for TensorFlow).
- Layer profile: captured by the framework’s profiling capability (RunOptions.TraceLevel for TensorFlow).
- GPU kernel profile: captured by NVIDIA CUDA Profiling Tools Interface (CUPi).
- MLModelScope processes and places all the profiles into a single timeline, and sends the “trace” to a database.
- Analyses are done automatically at three levels.
- ResNet v1.5 with batch size 256 is shown as an example.

Table 1: Eight models from MLPerf, AI-Matrix, and TensorFlow-Flow model zoo were selected for evaluation. We measured the peak throughput achieved on Amazon P3 and the corresponding batch size.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Peak Throughput (inputs/s)</th>
<th>Batch Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MobileNet-v1</td>
<td>2585.5</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>ResNet50-v1.5</td>
<td>996.3</td>
<td>256</td>
</tr>
<tr>
<td>3</td>
<td>SSD-MobileNet-v1-1000x1000</td>
<td>35.5</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>SSD-ResNet50-1200x1200</td>
<td>11.34</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>DenseNet-121</td>
<td>944.8</td>
<td>128</td>
</tr>
<tr>
<td>6</td>
<td>ResNet52-v1</td>
<td>468.5</td>
<td>256</td>
</tr>
<tr>
<td>7</td>
<td>Faster-RCNN-ResNet50</td>
<td>16.8</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>Mask-RCNN-ResNet50-v2</td>
<td>4.4</td>
<td>1</td>
</tr>
</tbody>
</table>

Across-Stack Profiling

- Model Level Analysis
- GPU Kernel Level Analysis

Layer Level Analysis

- Layer occurrence
- Layer aggregated latency
- Layer latency
- Layer allocated memory
- Layer GPU vs CPU latency

Conclusion

- More details are described in our paper (QR code →).
- We are currently working on using the data captured from MLModelScope to give suggestions on the model/system to use for a dataset.