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

Cheng Li ★, Abdul Dakkak ★, , Jinjun Xiong†, Wei Wei *, Ling Jie Xu *, Wei Zhang * Wen-Mei Hwu ★
{cli99,dakkak,w-hwu}@illinois.edu, jinjun@us.ibm.com, {w.wei, lingjie.xu, wz.ww}@alibaba-inc.com
★University of Illinois Urbana-Champaign, †IBM Research, *Alibaba Group

C3SR center for cognitive computing systems research

IBM | ILLINOIS

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-sc19.netlify.com**

Across-Stack Profiling

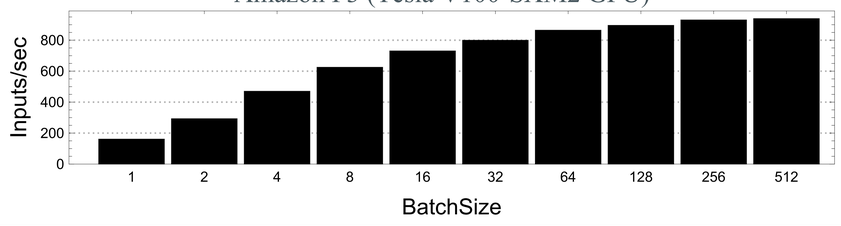
- Model profile: the time spent running the inference for C API (TF_SessionRun for TensorFlow)
- Layer profile: captured by the framework's profiling capability (RunOptions.TraceLevel for TensorFlow)
- GPU kernel profile: captured by NVIDIA CUDA Profiling Tools Interface (CUPTI)
- MLModelScope processes 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

ID	Name	Peak Throughput (inputs/s)	Batch Size
1	MobileNet-v1	2585.5	128
2	ResNet50-v1.5	996.3	256
3	SSD-MobileNet-v1-300x300	35.5	64
4	SSD-ResNet34-1200x1200	11.34	1
5	Densenet-121	944.8	128
6	ResNet152-v1	468.5	256
7	Faster-RCNN-ResNet50	16.8	4
8	Mask-RCNN-ResNet50-v2	4.4	1

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

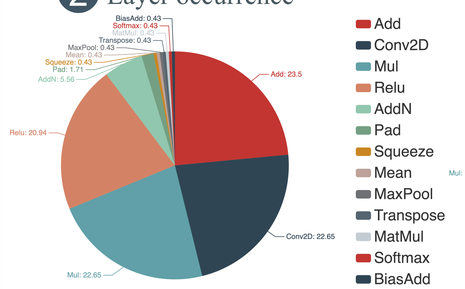
Model Level Analysis

1 MLPerf ResNet v1.5 throughput across different batch sizes on Amazon P3 (Tesla V100-SXM2 GPU)

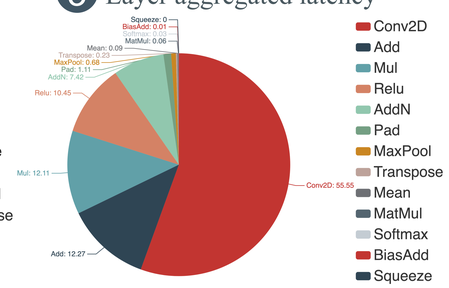


Layer Level Analysis

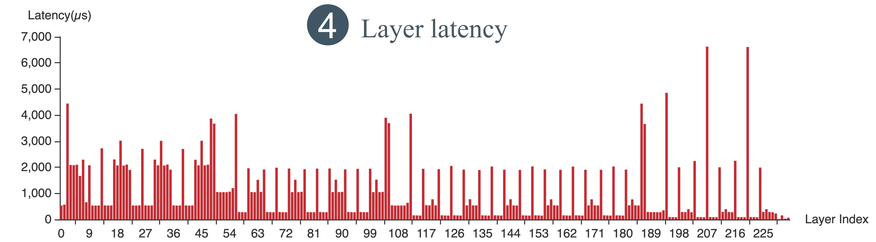
2 Layer occurrence



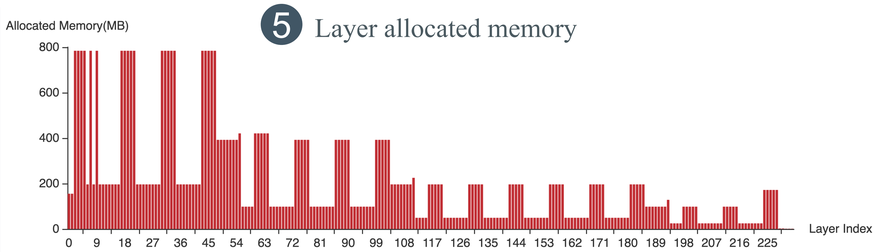
3 Layer aggregated latency



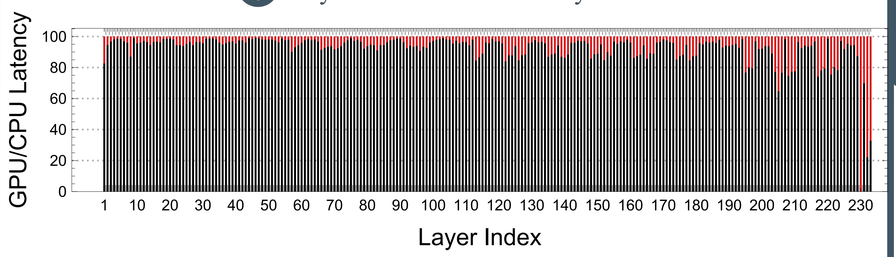
4 Layer latency



5 Layer allocated memory

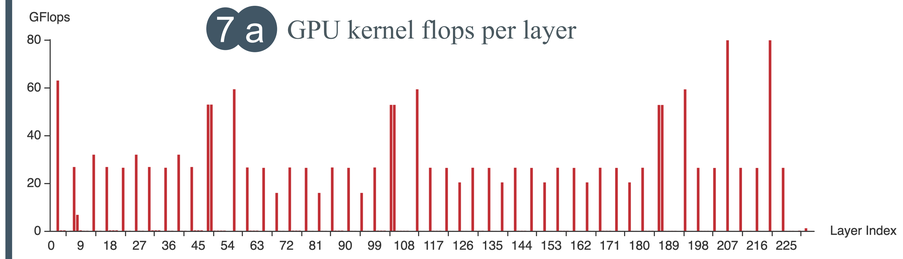


6 Layer GPU vs CPU latency

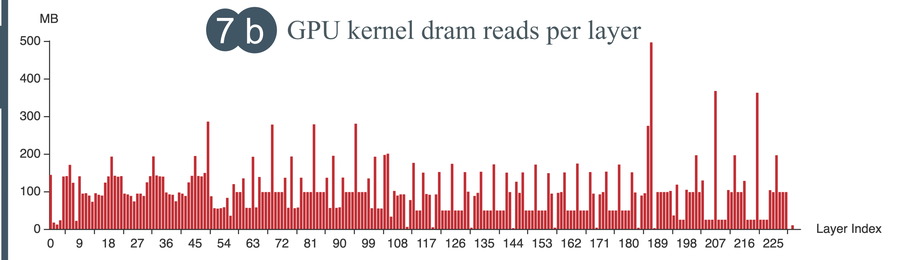


GPU Kernel Level Analysis

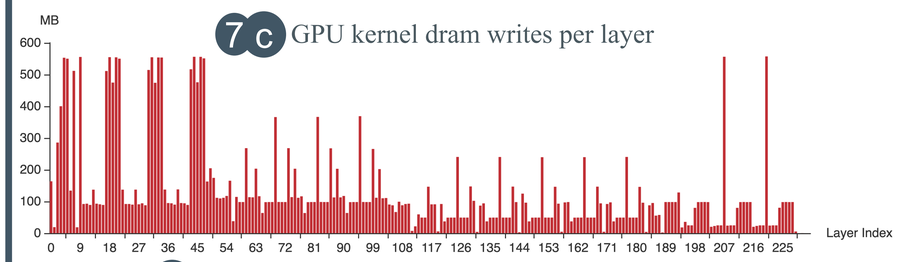
7a GPU kernel flops per layer



7b GPU kernel dram reads per layer



7c GPU kernel dram writes per layer



8 Model, layer, GPU kernel compute or memory bound

Item		Latency Percentage (%)	Memory Bound
Model (ResNet v1.5)		100	✕
Top 3 layers	resnet_model/conv2d_48/Conv2D	2.6	✕
	resnet_model/conv2d_51/Conv2D	2.6	✕
	resnet_model/conv2d_45/Conv2D	1.9	✕
Top 3 GPU kernels	volta_scudnn_128x64_relu_interior_nn_v1	28.7	✕
	scalar_prodcut_op (Eigen)	11.1	✓
	scalar_sum_op (Eigen)	10.3	✓

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

MLModelScope.org