The Vision Behind MLPerf (mlperf.org):
An ML Benchmark Suite for ML Software Frameworks and ML Hardware Accelerators in ML Cloud and Edge Computing Platforms

2018 International Symposium on Benchmarking, Measuring and Optimizing

December 11th, 2018
Prof. Vijay Janapa Reddi

(representing the viewpoints of many, many, people in MLPerf)

John Hennessy and David Patterson

Jeff Dean, David Patterson, Cliff Young
The (Rapid) Rise of ML

- The number of ML papers published Arxiv each year is growing exponentially.
- The pace of growth is on par and if not exceeding the rate of Moore’s Law scaling.

Source: https://blog.openai.com/ai-and-compute/
AI to Compute: 300,000x Increase in Compute

“... since 2012 the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time (by comparison, Moore’s Law had an 18-month doubling period). Since 2012, this metric has grown by more than 300,000x (an 18-month doubling period would yield only a 12x increase). Improvements in compute have been a key component of AI progress, so as long as this trend continues, it’s worth preparing for the implications of systems far outside today’s capabilities.”

Source: https://blog.openai.com/ai-and-compute/
Deep Learning has Reinvigorated Hardware

GPUs ⇒ AlexNet, Speech.

TPUs ⇒ Many Google applications: AlphaGo and Translate, WaveNet speech.

→ Rapidly fueling the renaissance of the hardware industry, including startups
Today, at least 45 start-ups are working on chips that can power tasks like speech and self-driving cars, and at least five of them have raised more than $100 million from investors. Venture capitalists invested more than $1.5 billion in chip start-ups last year, nearly doubling the investments made two years ago, according to the research firm CB Insights.
How do we compare the hardware?
How do we compare the hardware, today?

Answer is “surprisingly badly.”

- Example: single-benchmark measurement of throughput
  - Synthetic training data
  - Measure performance, ignoring accuracy

- Poor reproducibility
  - No means to effectively reproduce the same results
  - Hard to compare numbers across different models, inputs and datasets

- “ResNet-50” is not a precise specification, but it’s what everyone reports.
How do we design better hardware?
How do we design better hardware? More MACS?!

- Model performance cannot be evaluated using raw hardware performance (MACs)
- Model latency varies across different levels of MAC capability
- Latency ultimately impacts or dictates the experience

The **Three Cornerstones** for ML Performance

- **ML Model**
  - (Inception, MobileNets etc)

- **Inference Engine**
  - (TFLite, CoreML, vendor-specific SW toolchains)

- **ML Hardware Accelerators**
  - (TPU, Huawei NPU, etc)
Agenda

Why ML needs a benchmark suite?

Are there lessons we can borrow?

What is MLPerf?
  ○ How does MLPerf curate a benchmark?
  ○ What is the “science” behind the curation?
  ○ Where are we heading now?

What comes next for MLPerf?
Are there lessons we can borrow?
Are there lessons we can borrow? Yes!

A1: Look to successful history in benchmark suites: SPEC and TPC.

A2: Draw on experiences of those who have done ML benchmarking.
SPEC Impact

- Settled arguments in the marketplace (grow the pie)
- Resolved internal engineering debates (better investments)
- Cooperative ⇒ nonprofit Corporation with 22 members
- Universities join at modest cost and help drive innovation
- Became standard in marketplace, papers, and textbooks
- Needed to revise regularly to maintain usefulness:
  SPEC89, SPEC92, SPEC95, SPEC2000, SPEC2006, SPEC2017

Coincides with (caused?) the Golden Age of microprocessors...
Can we start a new Golden Age for ML Systems?
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?
Growing Number of Supporting Organizations

- 500+ discussion group members
- Researchers from 7 institutions
- 28+ Companies

<table>
<thead>
<tr>
<th>Alibaba</th>
<th>AMD</th>
<th>arm</th>
<th>Baidu</th>
<th>Cadence</th>
<th>Cerebras</th>
<th>Cisco</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cray</td>
<td>dôôô</td>
<td>Enflame</td>
<td>Esperanto</td>
<td>Google</td>
<td>Groq</td>
<td>Huawei</td>
</tr>
<tr>
<td>Intel</td>
<td>MediaTek</td>
<td>Mentor Graphics</td>
<td>Mythic</td>
<td>NetApp</td>
<td>NVIDIA</td>
<td>One Convergence</td>
</tr>
<tr>
<td>Rpa2ai</td>
<td>SambaNova</td>
<td>Samsung S.LSI</td>
<td>Sigopt</td>
<td>Synopsys</td>
<td>Tensyr</td>
<td>Wave Computing</td>
</tr>
</tbody>
</table>
Supporting Research Institutions

Harvard University  Stanford University  University of Arkansas, Littlerock  University of California, Berkeley  University of Minnesota  University of Texas, University of Toronto, Austin
MLPerf Goals

- Accelerate progress in ML via **fair and useful measurement**
- Serve both the **commercial and research communities**
- **Encourage innovation** to improve the state-of-the-art of ML
- **Enforce replicability** to ensure reliable results
- Use **representative workloads**, reflecting production use-cases
- Keep **benchmarking affordable** (so that all can play)
MLPerf Philosophy: **Agile** Benchmark Development

- **Extremely fast growth...**

- Rapidly iterate the benchmark suite
  - Remain relevant in the very fast moving machine learning field
  - Correct inevitable mistakes during the fast-paced benchmark formulation
  - Scale problems to match faster hardware, and better systems

- At least initially, revise annually? MLPerf18, MLPerf19, ...

- Like SPEC, have quarterly deadlines and then publish searchable results
Large public datasets

Cloud Training
Benchmarks
Standard logging
TF, pyTorch, ...

Edge Training
References
Standard logging
TF, pyTorch, ...

Cloud Inference
Benchmarks
Std. test harness, logging
TF saved model, ONNX, ...

Edge Inference
Benchmarks
Std. test harness, logging
TF saved model, ONNX, ...
Agile Benchmarking (Training) Timeline (in 2018)

May
- First general meeting

June
- Added benchmarks (volunteers!)

July
- Chartered working groups:
  - On-premise, Cloud, Submitters, special topics

August
- WGs report solid progress; inference WG chartered

September
- More WG progress

October
- First v0.5 submissions, with review period

November
- First results submissions!

December
- MLPerf results discussion (December 13th)
<table>
<thead>
<tr>
<th>June</th>
<th>Proposed an inference benchmark suite</th>
</tr>
</thead>
<tbody>
<tr>
<td>July</td>
<td>Gathered the important ML tasks to consider for inclusion</td>
</tr>
<tr>
<td>August</td>
<td>WG sanctions the tasks to generate implementations</td>
</tr>
<tr>
<td>September</td>
<td>Discussions on the models, datasets and metrics</td>
</tr>
<tr>
<td>October</td>
<td>Discussions on power and performance measurement</td>
</tr>
<tr>
<td>November</td>
<td>Code development and specification refinement</td>
</tr>
<tr>
<td>December</td>
<td>Code drops coming in (as we “speak”)</td>
</tr>
</tbody>
</table>
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?
Bootstrapping MLPerf 0.5v **Training** Effort

- Gathered researchers
  - Baidu (DeepBench)
  - Google (TF benchmarks)
  - Harvard (Fathom)
  - Stanford (DAWNBench)

- Combined the best parts from all of our experiences

- Planned to cover both training and inference; initial focus on **training**
## MLPerf Training Benchmarks 0.5v

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image Classification</td>
<td>ResNet-50</td>
<td>ImageNet</td>
</tr>
<tr>
<td>Object Detection</td>
<td>Mask-RCNN, SSD</td>
<td>MS-COCO 2017</td>
</tr>
<tr>
<td>Translation</td>
<td>Google NMT, Transformer</td>
<td>WMT16, WMT17</td>
</tr>
<tr>
<td>Recommendation</td>
<td>Neural Collaborative Filtering</td>
<td>MovieLens ml-20m</td>
</tr>
<tr>
<td>Reinforcement Learning</td>
<td>Minigo</td>
<td>NA</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>DeepSpeech2*</td>
<td>Librispeech</td>
</tr>
</tbody>
</table>
Toward the Definition of a ML **Inference** Task

- **Task description**
  - An overview of the ML task

- **Dataset**
  - A set of inputs and the corresponding ground-truth outputs. The dataset associated with a task also prescribes the input/output data format for the task

- **Quality metric**
  - A measure of the model’s quality/accuracy that is calculated using the ML task’s output(s), the ground-truth output(s) from the dataset and a loss function

<table>
<thead>
<tr>
<th>Task</th>
<th>Task Description</th>
<th>Dataset</th>
<th>Quality metric</th>
<th>Sample Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition</td>
<td>Classify an input into one of many categories. Alternatively, generate a high dimensional embedding that can be used for recognition</td>
<td>Imagenet/COCO</td>
<td>Top-1 error rate</td>
<td>Face authentication, Music recognition</td>
</tr>
<tr>
<td></td>
<td>Input: RGB image of size XX x YY</td>
<td>Output: label</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Output: label index</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML Tasks</td>
<td>Owner</td>
<td>Framework</td>
<td>Model</td>
<td>Dataset</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------</td>
<td>------------------</td>
<td>------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>Image Classification</strong></td>
<td>Guenther</td>
<td>TF and ONNX</td>
<td>Resnet50 1.5v</td>
<td>ImageNet</td>
</tr>
<tr>
<td><strong>Object Detection</strong></td>
<td>Itay Hubara</td>
<td>PyTorch</td>
<td>(1) VGG16</td>
<td>MS-COCO</td>
</tr>
<tr>
<td></td>
<td><a href="mailto:ihubara@habana.ai">ihubara@habana.ai</a>/</td>
<td></td>
<td>(2) SSD-MobileNet</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="mailto:christine.cheng@intel.com">christine.cheng@intel.com</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Speech Recognition</strong></td>
<td>Gennady/Anton</td>
<td>PyTorch</td>
<td>DeepSpeech2</td>
<td>Librispeech</td>
</tr>
<tr>
<td><strong>Machine Translation</strong></td>
<td><a href="mailto:rohit.kalidindi@intel.com">rohit.kalidindi@intel.com</a></td>
<td>Tensorflow</td>
<td>(1) GNMT</td>
<td>WMT16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="http://download.tensorflow.org/models/nmt/10122017/deen_gnmt_model_4_layer.zip">http://download.tensorflow.org/models/nmt/10122017/deen_gnmt_model_4_layer.zip</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(2) transformer</td>
<td></td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
<td><a href="mailto:adselvar@cisco.com">adselvar@cisco.com</a>,</td>
<td>PyTorch</td>
<td>Neural Collaborative Filtering</td>
<td>MovieLens 20M</td>
</tr>
<tr>
<td></td>
<td><a href="mailto:manasa.kankanala@intel.com">manasa.kankanala@intel.com</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Text (e.g. Sentiment) Classification</strong></td>
<td>Itay Hubara</td>
<td>PyTorch</td>
<td>seq2-CNN</td>
<td>IMDB</td>
</tr>
<tr>
<td></td>
<td><a href="mailto:ihubara@habana.ai">ihubara@habana.ai</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Language Modeling</strong></td>
<td><a href="mailto:gregdiamos@baidu.com">gregdiamos@baidu.com</a></td>
<td>TF</td>
<td><a href="https://github.com/tensorflow/models/tree/master/research/lm_1b">https://github.com/tensorflow/models/tree/master/research/lm_1b</a></td>
<td>(1) 1 billion words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) Amazon reviews</td>
</tr>
<tr>
<td><strong>Text To Speech</strong></td>
<td>Amit Bleiweiss</td>
<td>Caffe2</td>
<td>WaveNet</td>
<td>LJSpeech</td>
</tr>
<tr>
<td></td>
<td><a href="mailto:amit.bleiweiss@intel.com">amit.bleiweiss@intel.com</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Image Segmentation</strong></td>
<td>N/A</td>
<td></td>
<td>MaskRCNN</td>
<td>COCO</td>
</tr>
<tr>
<td>ML Tasks</td>
<td>Owner</td>
<td>Framework</td>
<td>Model</td>
<td>Dataset</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------</td>
<td>---------------</td>
<td>--------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Image Classification</td>
<td>(1) Anton</td>
<td>(1) TF-Lite</td>
<td>(1) MobileNets-v1.0 224??</td>
<td>ImageNet</td>
</tr>
<tr>
<td></td>
<td>(2) Fei and Mejia, Andres</td>
<td>(2) Caffe2/ONNX</td>
<td>(2) ShuffleNet</td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="mailto:andres.mejia@intel.com">andres.mejia@intel.com</a></td>
<td></td>
<td>(<a href="https://s3.amazonaws.com/download.onnx/models/opset_6/shufflenet.tar.gz">https://s3.amazonaws.com/download.onnx/models/opset_6/shufflenet.tar.gz</a>)</td>
<td></td>
</tr>
<tr>
<td>object Detection</td>
<td>(1) Yuchen (<a href="mailto:yuchen.zhou@gm.com">yuchen.zhou@gm.com</a>)</td>
<td>(1) TF</td>
<td>(1) SSD-ResNet50</td>
<td>(1) VOC</td>
</tr>
<tr>
<td></td>
<td>(2) Scott Gardner (MN)/christine.cheng@intel.com</td>
<td>(2) TF-Lite</td>
<td>(2) SSD-MobileNetsV1</td>
<td>(2) COCO</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Scott Gardner</td>
<td>TF</td>
<td>DeepSpeech1 (Mozilla)</td>
<td>(1) Librispeech</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2) &quot;noisy&quot; validation</td>
</tr>
<tr>
<td>Machine Translation</td>
<td><a href="mailto:rohit.kalidindi@intel.com">rohit.kalidindi@intel.com</a></td>
<td>Tensorflow</td>
<td>GNMT</td>
<td>WMT16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(<a href="http://download.tensorflow.org/models/nmt/10122017/deen_gnmt_model_4_layer.zip">http://download.tensorflow.org/models/nmt/10122017/deen_gnmt_model_4_layer.zip</a>)</td>
<td></td>
</tr>
<tr>
<td>Text To Speech</td>
<td></td>
<td>Tensorlayer</td>
<td>WaveNet</td>
<td></td>
</tr>
<tr>
<td>Face Identification</td>
<td>David Lee <a href="mailto:david.lee@mediatek.com">david.lee@mediatek.com</a></td>
<td>TF-Lite</td>
<td>SphereFace</td>
<td>LFW</td>
</tr>
<tr>
<td>Image Segmentation</td>
<td>Carole Wu/Fei Sun</td>
<td>Caffe2/ONNX</td>
<td>MaskRCNN2Go</td>
<td>COCO</td>
</tr>
<tr>
<td>Image Enhancement</td>
<td><a href="mailto:christine.cheng@intel.com">christine.cheng@intel.com</a></td>
<td>Tensorflow based on <a href="https://github.com/tensorlayer/srgan">https://github.com/tensorlayer/srgan</a></td>
<td>SRGAN</td>
<td>DIV2K</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(<a href="https://github.com/tensorlayer/srgan/releases/tag/1.2.0">https://github.com/tensorlayer/srgan/releases/tag/1.2.0</a>)</td>
<td></td>
</tr>
</tbody>
</table>
MLPerf Cloud/Edge Inference Matrix
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
    - What is the “science” behind the curation?
    - Where are we heading now?
- What comes next for MLPerf?
“Science”

Metrics    Method
“Science”

Metrics

Method
Toward a Unified Metric: Performance *and* Quality

- **Performance:** how fast is a model for training, inference?
- **Quality:** how good are a model’s predictions?

Important for benchmark to capture both performance and quality
Performance and Quality Aren’t Always Correlated

**Training**

- End-to-end training of a ResNet56 CIFAR10 model
- Nvidia P100 machine with 512 GB of memory and 28 CPU cores
- TensorFlow 1.2 compiled from source with CUDA 8.0 and CuDNN 5.1

![Graphs showing throughput and accuracy with varying batch sizes](image)
Training Metric: **Time to reach quality target**

- Quality target is *specific for each benchmark and close to state-of-the-art*
  - Updated w/ each release to keep up with the state-of-the-art

- Time includes preprocessing, validation over median of 5 runs

- Available: reference implementations that achieve quality target
Performance and Quality Aren’t Always Correlated

Inference

- For a given latency target, you can achieve different levels of model quality
- Possible to trade-off model accuracy with complexity
- Model performance (inference/s) is insufficient

Performance and Quality Aren’t Always Correlated

Inference

- For a given latency target, you can achieve different levels of model quality
- Possible to trade-off model accuracy with complexity
- Model performance (inference/s) is insufficient
“Science”

Metrics

Method
What start/ends do we measure and why?

Options:
1. Option 1
2. Option 2
3. Option 3
On-Device OCR: A case study
PhotoOCR Normalized Performance (CPU only)

- Sparse

- Dense
Do we account for pre- and post-processing times in the inference run test?
MLPerf Challenges

Cloud (Training E.g.)
- Hyperparameters
- Scale
- Power
- Cost
- Variance
- On-premise vs. cloud
- ...

Edge (Inference E.g.)
- Quantizations
- Sparsity
- Pruning
- Scores
- Variance
- Power
- ...

48
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
    - Where are we heading now?
- What comes next for MLPerf?
Where are we heading now?

- First version: reference code, in two frameworks, of each benchmark.
- Resolving or controlling the variance issues.
- Working on the inference suite.
- Getting to governance, and an umbrella organization.
Reference Implementations → Call for Submissions

Closed division submissions
- Requires using the specified model
- Limits overfitting
- Enables apples-to-apples comparison
- Simplifies work for HW groups

Open division submissions
- Open division allows using any model
- Encourages innovation
- Ensures Closed division does not stagnate
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?
Training

Inference

Cloud
On-Prem.

Cloud
Edge

Cellular
Automotive
Robotics

???
Large public datasets

Cloud Training
Benchmarks
Standard logging
TF, pyTorch, ...

Edge Training
References
Standard logging
TF, pyTorch, ...

Cloud Inference
Benchmarks
Std. test harness, logging
TF saved model, ONNX, ...

Edge Inference
Benchmarks
Std. test harness, logging
TF saved model, ONNX, ...

Create industry driven public datasets
Policy

Large public datasets

Benchmarks and Standardization (MLPerf)

(nothing is set in stone yet, we are looking for ideas)
Agenda

- Why ML needs a benchmark suite?
- Are there lessons we can borrow?
- What is MLPerf?
  - How does MLPerf curate a benchmark?
  - What is the “science” behind the curation?
  - Where are we heading now?
- What comes next for MLPerf?
Concluding thoughts...
Recap of “The Vision Behind MLPerf”

- Machine Learning needs benchmarks!
- Goals: agility, both research and development, replicability, affordability
- MLPerf Training: v0.5 deadline was in November
- MLPerf Inference is under construction
  - Inference workload suite under development
  - Q1 reference implementations finalized
  - Q2/3 solicit inference result submissions”

(for rapid iteration to work, we need good input!)
MLPerf needs your help!

- Join the discussion community at MLPerf.org

- Help us by joining a working group:
  Cloud scale, on-premises scale, submitters, special topics, inference.
  Help us design submission criteria, to include the data you want

- Propose new benchmarks and data sets

- Submit your benchmark results!
More at MLPerf.org, or contact info@mlperf.org
Acknowledgements

Peter Mattson

Cliff Young

David Patterson

Carole-Jean Wu

Greg Diamos

... and countless other working group members!
Thank You