A Benchmark for Machine Learning from an Academic/Industry Cooperative

Researchers from:
Baidu, Google, Harvard, Stanford, and UC Berkeley
Contributors

Baidu: Siddharth Goyal, Sharan Narang, Greg Diamos

Google: Karmel Allison, Victor Bittorf, Kathy Wu, Cliff Young, Peter Mattson

Harvard: Udit Gupta, Lillian Pentecost, Brandon Reagen, Gu-Yeon Wei

Stanford: Cody Coleman, Daniel Kang, Deepak Narayanan, Peter Bailis, Matei Zaharia

University of California, Berkeley: Ion Stoica, David Patterson
<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Metric</th>
<th>When</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibson Instruction Mix</td>
<td>MIPS: Million Instructions Per Second</td>
<td>1970</td>
</tr>
<tr>
<td>(Frequency of instructions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whetstone, Dhrystone</td>
<td>Whetstones, Dhrystones per second</td>
<td>1976, 1984</td>
</tr>
<tr>
<td>(Synthetic programs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puzzle, Quicksort</td>
<td>MIPS</td>
<td>1981</td>
</tr>
<tr>
<td>(Toy programs)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Kernels)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SPEC: System Performance Evaluation Cooperative

- Application level benchmarking (enable via high-level languages and portability of UNIX OS)
- Cross-platform benchmarking and evaluation
- Industry and academia to join at reasonable cost
- Standard in marketplace, papers, and textbooks

**Fig 1.17, Computer Architecture: A Quantitative Approach, 6th Edition, 2018**
Companies:

- AMD
- Baidu
- Google
- Intel
- SambaNova
- Wave Computing

Researchers from these educational institutions:

- Harvard University
- Stanford University
- University of California, Berkeley
- University of Minnesota
- University of Toronto
Goals for MLPerf

1. Accelerate progress in ML via fair and useful measurement
2. Encourage innovation across state-of-the-art ML systems
3. Serve both industrial and research communities
4. Enforce replicability to ensure reliable results
5. Keep benchmark effort affordable so all can play
Difficulties of ML Benchmarking

1. Diversity in deep learning models used
   a. Problem domain
   b. Models
   c. Datasets

2. Pace of field
   a. State-of-the-art models evolve every few months

3. Lack of evaluation metric
   a. Accuracy
   b. Time to train, latency of inference

4. Multi-disciplinary field
   a. Algorithms, Systems, Hardware
Outline

- Model diversity
- Agile benchmark development
- Evaluation metrics
- Open and closed divisions
- Contributing to MLPerf
Fathom suite showed breadth in ML benchmarking

- Collection of 8 diverse learning models
- Clear, tested implementations in TensorFlow
- Training and inference modes provided
- Provided broad view and coverage
- Models have drastically changed and greatly advanced since 2015
## Benchmarks Considered for MLPerf

<table>
<thead>
<tr>
<th>Area</th>
<th>Vision</th>
<th>Language</th>
<th>Audio</th>
<th>Commerce</th>
<th>Action / RL</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
<td>Image Classification</td>
<td>Translation Language Model</td>
<td>Speech Recognition</td>
<td>Rating</td>
<td>Games</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Object Detection /</td>
<td>Model Word Embedding</td>
<td>Text-to-Speech</td>
<td>Recommendations</td>
<td>Go</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Segmentation</td>
<td></td>
<td>Question Answering</td>
<td>Sentiment Analysis</td>
<td>Robotics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Face ID</td>
<td></td>
<td>Keyword Answering</td>
<td>Next-action</td>
<td>Health Care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HealthCare (Radiology)</td>
<td></td>
<td>Language Spotting</td>
<td>Healthcare (EHR)</td>
<td>Bioinformatics</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Video Detection</td>
<td></td>
<td>Modeling</td>
<td>Fraud detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-Driving</td>
<td></td>
<td>Chatbots</td>
<td>Anomaly detection</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Speaker ID</td>
<td>Time series prediction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Graph embeddings</td>
<td>Large scale regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Content ID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datasets</td>
<td>ImageNet</td>
<td>WMT</td>
<td>LibriSpeech</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>COCO</td>
<td>English-German</td>
<td>SQuAD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LM-Benchmark</td>
<td></td>
<td>Atari</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MovieLens-20M</td>
<td></td>
<td>Go</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amazon</td>
<td></td>
<td>Chess</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IMDB</td>
<td></td>
<td>Grasping</td>
<td></td>
</tr>
<tr>
<td>Models</td>
<td>ResNet-50</td>
<td>Transformer</td>
<td>Deep Speech 2</td>
<td>Neural Collaborative</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TF Object Detection</td>
<td>OpenNMT</td>
<td>SQuAD Explorer</td>
<td>Filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Detectron</td>
<td></td>
<td></td>
<td>CNNs</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>COCO mAP</td>
<td>BLEU</td>
<td>WER</td>
<td>Prediction accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metrics</td>
<td>Prediction accuracy</td>
<td></td>
<td>Perplexity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Datasets**
- ImageNet
- COCO
- WMT
- English-German
- LibriSpeech
- SQuAD
- LM-Benchmark
- MovieLens-20M
- Amazon
- IMDB
- Atari
- Go
- Chess
- Grasping

**Models**
- ResNet-50
- TF Object Detection
- Detectron
- Transformer
- OpenNMT
- Deep Speech 2
- SQuAD Explorer
- Neural Collaborative Filtering
- CNNs
- DQN
- PPO

**Accuracy Metrics**
- COCO mAP
- Prediction accuracy
### MLPerf benchmarks (version 0.5)

<table>
<thead>
<tr>
<th>Area</th>
<th>Benchmark</th>
<th>Dataset</th>
<th>Model</th>
<th>Reference Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision</td>
<td>Image classification</td>
<td>ImageNet</td>
<td>ResNet</td>
<td>TensorFlow</td>
</tr>
<tr>
<td></td>
<td>Object detection</td>
<td>COCO</td>
<td>Mask R-CNN</td>
<td>Caffe 2</td>
</tr>
<tr>
<td>Language/Audio</td>
<td>Translation</td>
<td>WMT Eng-Germ</td>
<td>Transformer</td>
<td>TensorFlow</td>
</tr>
<tr>
<td></td>
<td>Speech recognition</td>
<td>LibriSpeech</td>
<td>Deep Speech 2</td>
<td>PyTorch</td>
</tr>
<tr>
<td>Commerce</td>
<td>Recommendation</td>
<td>MovieLens-20M</td>
<td>NCF</td>
<td>PyTorch</td>
</tr>
<tr>
<td></td>
<td>Sentiment Analysis</td>
<td>IMDB</td>
<td>Seq-CNN</td>
<td>PaddlePaddle</td>
</tr>
<tr>
<td>Action</td>
<td>Reinforcement Learning</td>
<td>Go</td>
<td>Mini-go</td>
<td>TensorFlow</td>
</tr>
</tbody>
</table>

- Balance benchmarks that represent
  - Industry workloads
  - Coverage of different areas and characteristics
Outline

- Model diversity
- **Agile benchmark development**
- Evaluation metrics
- Open and closed divisions
- Contributing to MLPerf
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field

AlexNet (2012)

VGG16 (2014)

Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field

From Samy Bengio’s opening remarks at NIPS 2017
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field

NIPS 2017 had **3240 submissions**
NIPS 2018 had **~4900 submissions**

From Samy Bengio’s opening remarks at NIPS 2017
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Scale problems to match faster hardware
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Scale problems to match faster hardware

A **300,000x Increase in Compute** since 2012

From OpenAI Blog “AI and Compute”
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Scale problems to match faster hardware
  - Correct inevitable mistakes in the formulation
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Scale problems to match faster hardware
  - Correct inevitable mistakes in the formulation
- At least initially, revise annually? MLPerf18, MLPerf19, ...
Agile Benchmark Development

- Rapidly iterate the benchmark suite:
  - Remain relevant in the very fast moving ML field
  - Scale problems to match faster hardware
  - Correct inevitable mistakes in the formulation
- At least initially, revise annually? MLPerf18, MLPerf19, ...
- Like SPEC, have quarterly deadlines and then publish results for that quarter via searchable database
Outline

● Model diversity
● Agile benchmark development
● **Evaluation metrics**
● Open and closed divisions
● Contributing to MLPerf
Metrics Should Capture Performance and Quality

- **Performance**: how fast is a model for training, inference?
  - Focus of benchmarks like DeepBench, Fathom
- **Quality**: how good are a model’s predictions?
  - Focus of benchmarks like ImageNet, MS COCO
Performance and Quality aren’t always correlated

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

Measures Performance (Time, Cost) to Fixed Quality Target

DAWNBench
An End-to-End Deep Learning Benchmark and Competition

---

**Training Time**

**Objective:** Time taken to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Time to 93% Accuracy</th>
<th>Model</th>
<th>Hardware</th>
<th>Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0:30:43</td>
<td>ResNet50 Google source</td>
<td>Half of a TPUv2 Pod</td>
<td>TensorFlow 1.8.0-rc1</td>
</tr>
</tbody>
</table>

---

MLPerf
MLPerf metric: **Training time** to reach quality target + cost or power

- 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
  - Median of 5 runs
- Time includes preprocessing and validation
- Reference implementations that achieve quality target

**In addition, either:**
- *Cost* of public cloud resources *(no spot/preemptible instances)*
- *Power utilization* for on-premise hardware
Summary result combines benchmark metrics

Why?

- Provide a concise indicator of “general purpose ML” performance
- Encourage the field to move in a common direction, ultimately leading to greater performance across the board

How? For participants that submit to each benchmark category:

- For each benchmark task, normalize the time result to the reference implementation on baseline hardware
- Summary score computed via geometric mean of results
Outline

- Model diversity
- Agile benchmark development
- Evaluation metrics
- **Open and closed divisions**
- Contributing to MLPerf
Goal: Encourage Innovation and fair comparison
Goal: **Encourage Innovation** and fair comparison

- ML algorithms are under active development

低精度

稀疏性
Goal: **Encourage Innovation** and fair comparison

- ML algorithms are under active development
- Many models with different trade-offs

![Diagrams of ResNet, ResNeXt, and Wide ResNet](image-url)
Goal: Encourage Innovation and **fair comparison**

- ML algorithms are under active development
- Many models with different trade-offs
Goal: Encourage Innovation and **fair comparison**

Innovative algorithm or **overfitting**
Open/Closed Divisions + Replication

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

- **Open** division allows using any model
  - Encourages innovation
  - Ensures closed division does not stagnate
Outline

● Model diversity
● Agile benchmark development
● Evaluation metrics
● Open and closed divisions
● Contributing to MLPerf
Plan: move fast, become independent standard

- Start as small cooperative to quickly publish good benchmark suite soon

- Invite every like-minded group who shares the goals of MLPerf:
  - Big companies
  - Startups
  - Universities

- Current version “0.5”. For 1.0, transfer to independent org.
Ways to support and be involved

- **Github**: reference code
- **Submissions**: data points
- **Google group**: discussion of the benchmark and changes
- **Meetings**: community building and focused discussion towards action
- **Working groups**: targeted groups to flesh out specific areas
  - Inference
  - Reinforcement learning
  - Summary score
  - Measuring power and cost
More at MLPerf.org, or contact info@mlperf.org