REQUIREMENTS FOR AN ENTERPRISE AI BENCHMARK

31 August 2018
10th TPCTC (Technology Conference on Performance Evaluation & Benchmarking), co-located with VLDB 2018 in Rio de Janeiro, Brazil
http://www.tpc.org/tpctc/tpctc2018/default.asp
Atos : Cedric Bourrasset, France Boillod-Cerneux, Ludovic Sauge, Myrtille Deldossi, François Weillenreiter
IBM : Rajesh Bordawekar, Susan Malaika, Jean-Armand Broyelle, Marc West, Brian Belgodere

MOTIVATION

- IBM & Atos are facing similar challenges regarding AI performance evaluation for enterprise customer engagement:
  - Which part of AI model development process is relevant for enterprise customers (data preparation, model training, inference performance) ?
  - Which part of the platform solution should be benchmarked (Hardware, Software, ML framework, all, …) ?
  - Which relevant performance metrics should be used for ensuring fair comparison between competitive solutions ?
  - Which existing benchmarks can be mapped to these needs ?

- Launch this joint IBM & Atos initiative: Assemble team, interview existing customers, share our knowledge, map metrics to benchmarks, and then we decided to write a paper and share these information with the community

SUMMARY

- This talk reviews the challenges and metrics for enterprise workloads, the benchmark tests that are available, and the gaps which need to be filled.
- The paper, that this talk is based on, identifies the following areas as important to enterprises concerned about performance:
  - **1. Model training performance**
    - data labeling / preparation
    - time-to-accuracy
    - computational time / cycles
    - throughput-to-accuracy
  - **2. Hyper-parameter optimization performance**
  - **3. Inference runtime performance**
- The talk offers a summary table of the main three AI areas important to enterprises, alongside:
  - Workload profile
  - Important performance indicators to assess the task's efficiency
  - Potential technical bottlenecks to look out for that could limit the AI tasks performance delivered by a given solution.

AI TASKS : MODEL TRAINING

AI Task	Workload Profile	Important Performance Indicators	Potential Technical Bottlenecks in Standalone Scenarios	More Potential Technical Bottlenecks in Concurrent Scenarios
Model Training	Batch task	Trained models, duration of the training process, scalability of the training mechanism if any	GPU memory capacity and latency/bandwidth, GPU compute capabilities and capacities	Platform ability to efficiently manage systems resources and schedule AI training workload (similar to HPC workload management tool benchmark)
	GPU intensive workload	Price-Performance metrics: in regard to TCA (Total Cost of Acquisition) and or TCO (Total Cost of Ownership)	GPU-CPU and CPURAM communication characteristics could matter for large dataset training and/or Out-of-GPU-memory training	GPU-CPU communication characteristics could matter for large datasets
	Minutes to days	For concurrent scenarios, the level of model training concurrency (similar to batch concurrency benchmarks)	Server-server communication characteristics could matter for intra-parallelism model training (Training a single model across multiple servers)	CPU-RAM communication characteristics could matter for Out-of-Core training

AI TASKS : HYPER-PARAMETER OPTIMIZATION

AI Task	Workload Profile	Important Performance Indicators	Potential Technical Bottlenecks in Standalone Scenarios	More Potential Technical Bottlenecks in Concurrent Scenarios
Hyper-parameter Optimization	Batches tasks managed by a workload orchestrator and hyper-parameters solver	Hyper-parameter combinatorial values to cover	Solver Algorithm limitations	All the Model Training potential bottlenecks apply here as well
	GPU-tasks	Optimum value found for the model Accuracy	All the Model Training potential bottlenecks apply here	
	Minutes to days	Overall duration to find the model with the best hyper- parameters		

AI TASKS : INFERENCE RUN-TIME PERFORMANCE

AI Task	Workload Profile	Important Performance Indicators	Potential Technical Bottlenecks in Standalone Scenarios	More Potential Technical Bottlenecks in Concurrent Scenarios
Deployed Model Inference Run-time	Online Service or library API	Latency of inference	GPU latency, GPU compute capabilities and capacities	Platform ability to efficiently manage the systems
	Aiming for real-time request response-time in most cases. Milli-seconds to seconds.	Price performance metrics: in regard to TCA (Total Cost of Acquisition) and or TCO (Total Cost of Ownership)	GPU-CPU and CPU-RAM latency and throughput	
	Mostly hardware AI accelerator intensive workload (GPU, FPGA, neuromorphic chip, Embedded Solutions)	For embedded and/or autonomous systems: Energy consumption/performance metric	Infrastructure network communication characteristics	
		For concurrent scenario, the level of model inference concurrency (similar to OLTP metrics)		

EXISTING AI BENCHMARKS

- **DeepBench** : This benchmark targets low-level operations that are fundamental to deep learning, such as matrix-multiplication, convolutions, and communications, and aims to identify the most appropriate hardware but the benchmark does not consider time-to-accuracy.
- **TensorFlow** : The TensorFlow performance benchmarks are similar to DeepBench, in that they identify the most appropriate hardware, but not time-to-accuracy currently. They are also tied to the TensorFlow Framework.
- **DAWNBench** : DAWNBench allows different deep learning methods to be compared by running a number of competitions. It was the first major benchmark suite to examine end-to-end deep learning training and inference. It does not address data preparation and hyper-parameter optimization work.
- **MLPerf** : MLPerf defines the primary metric as the wall clock time to train a model to a target quality, often hours or days. The target quality is based on the current state of the art publication results, less a small delta to allow for run-to-run variance. MLPerf does not address hyper-parameter optimization nor data preparation.
  - The **MLPerf Closed Model Division** specifies the model to be used and restricts the values of the hyper parameters (batch size, learning rate, etc.) which can be tuned in an attempt to create a fair and balance comparison of the hardware and software systems.
  - The **MLPerf Open Model Division**, only requires that same task must be achieved using the same data, but provides fewer restrictions

CONCLUSION

- Leverage knowledge from existing AI benchmarks, but must enhance or expand
- Community effort needed to advance benchmarks
  - Academic
  - Industry
- Benchmarks need to be domain-specific
  - Prioritize domains to focus on

If you are interested in collaborating on Enterprise AI benchmarks, please connect with one of the paper co-authors e.g., malaika@us.ibm.com

RELATED MATERIALS

- Full Paper: Requirements for an Enterprise AI Benchmark, *Cedric Bourrasset, France Boillod-Cerneux, Ludovic Sauge, Myrtille Deldossi, François Weillenreiter, Rajesh Bordawekar, Susan Malaika, Jean-Armand Broyelle, Marc West and Brian Belgodere*at Tenth TPC Technology Conference on Performance Evaluation & Benchmarking (TPCTC 2018)
  - Conference http://www.tpc.org/tpctc/tpctc2018/default.asp
  - Conference Proceedings http://www.tpc.org/tpctc/default.asp
  - Blog https://developer.ibm.com/opentech/2018/08/13/requirements-enterprise-ai-benchmark/
- TPC: Active TPC Benchmarks http://www.tpc.org/information/benchmarks.asp
- SPEC: https://www.spec.org/cpu2017/Docs/overview.html
- DeepBench: https://github.com/baidu-research/DeepBench
- TensorFlow: https://www.tensorflow.org/performance/benchmarks
- DAWNBench: https://dawn.cs.stanford.edu//benchmark/
- MLPerf: https://mlperf.org/

BACKUP

CSIG (Cognitive Systems Institute Group) Talk – Aug 23, 2018 - 10:30-11am US Eastern

Title: Requirements for an Enterprise AI Benchmark
Speakers: Cedric Bourrasset, Atos Bull; Rajesh Bordawekar, IBM

Description: At present, AI benchmarks either focus on evaluating deep learning approaches or infrastructure capabilities. These approaches don't capture end-to-end performance behavior of enterprise AI workloads. It is also clear that there is not one reference metric that will be suitable for all AI applications nor all existing platforms. We first present the state of the art regarding the current basic and most popular AI benchmarks. We then present the main characteristics of AI workloads from various industrial domains. Finally, we focus on the needs for ongoing and future industry AI benchmarks and conclude on the gaps to improve AI benchmarks for enterprise workloads.

Bios:

Cedric Bourasset : After receiving a Ph.D. in Electronics and computer vision in 2016 from the Blaise Pascal University of Clermont-Ferrand defending the dataflow model of computation for FPGA High Level Synthesis problematic in embedded machine learning application, Cedric is now working as AI Product Manager at Atos Bull with the mission to develop Atos AI product line. One product is a software solution for developing AI enterprise solutions and the other one is computer vision solution for people detection, tracking and reidentification into multi-camera environments.

Rajesh Bordawekar: Rajesh is a member of the Systems Acceleration department at the IBM T. J. Watson Research Center. Prior to joining IBM Research in September 1998, he was a post-doctoral fellow at the Center for Advanced Computing Research, California Institute of Technology. He received his PhD in Computer Engineering from Syracuse University. Rajesh studies interactions between applications, programming languages/runtime systems, and computer architectures. He is interested in understanding how modern hardware, multi-core processors, GPUs, and SSDs impact design of optimal algorithms for main-memory and out-of-core problems.

Zoom meeting Link: https://zoom.us/j/7371462221
Zoom Callin: (415) 762-9988 or (646) 568-7788 Meeting id 7371462221
Zoom International Numbers: https://zoom.us/zoomconference
(Check the website in case the date or time changes: http://cognitive-science.info/community/weekly-update/)

Thu, Aug 23, 10:30am US Eastern ❖ https://zoom.us/j/7371462221

More Details Here : http://cognitive-science.info/community/weekly-update/

@sumalaika