The Twelfth International Conference on Cloud Computing, GRIDs, and Virtualization

CLOUD COMPUTING 2021

Comparison of Benchmarks for Machine Learning Cloud Infrastructures

Authors:

Prof. Dr. Christoph Reich (IDACUS, Furtwangen University of Applied Science**)**

Manav Madan (IDACUS, Furtwangen University of Applied Science**)**

Presenter

Manav Madan (Manav.Madan@hs-furtwangen.de)

Research associate under Prof. Dr. Christoph Reich at Hochschule Furtwangen Universirty (HFU).

Field of Interest: : Machine and deep learning.

Currently, he is enrolled in the project Q-AMeLiA (Quality Assurance of Machine Learning Applications) which is funded by the Ministry of Science, Research and Arts of Baden-Württemberg (MWK). Q-AMeLiA is a consortium of three universities of applied sciences and five companies. The project aims to support these companies in the implementation of the special machine learning software development lifecycle (ML-SDLC) and the incorporation of important quality indicators.

Contents

1. What is a Benchmark?
2. Standard Benchmark Vs ML Benchmark
3. Properties of an ML Benchmark
4. Metrics for an ML Benchmark
5. Categorization of Benchmarks
6. Different Benchmarks
7. Conclusion
8. Bibliography

1) What is a Benchmark?

Definition :

In computing, a **benchmark** is the act of running a computer program, a set of programs, or other operations, in order to assess the relative **performance** of an object, normally by running a number of standard tests and trials against it [1].

In Machine learning the benchmarking is related to two different tasks that are training and inference. These two tasks do not overlap with each other. In training, the weights of a model have to be learned. To do so samples in mini-batches are shown to the model and the intermediate results are stored in the memory. **Training is usually a resource-intensive process as these intermediate results acquire a lot of memory and also training usually involves lots of complex mathematics calculations**. On the other hand, the inference is about evaluating a single data sample on the trained model at once.

Training benchmarks aim to know how well/fast a particular hardware trains a domain specific model.

2) Standard Benchmark Vs ML Benchmark

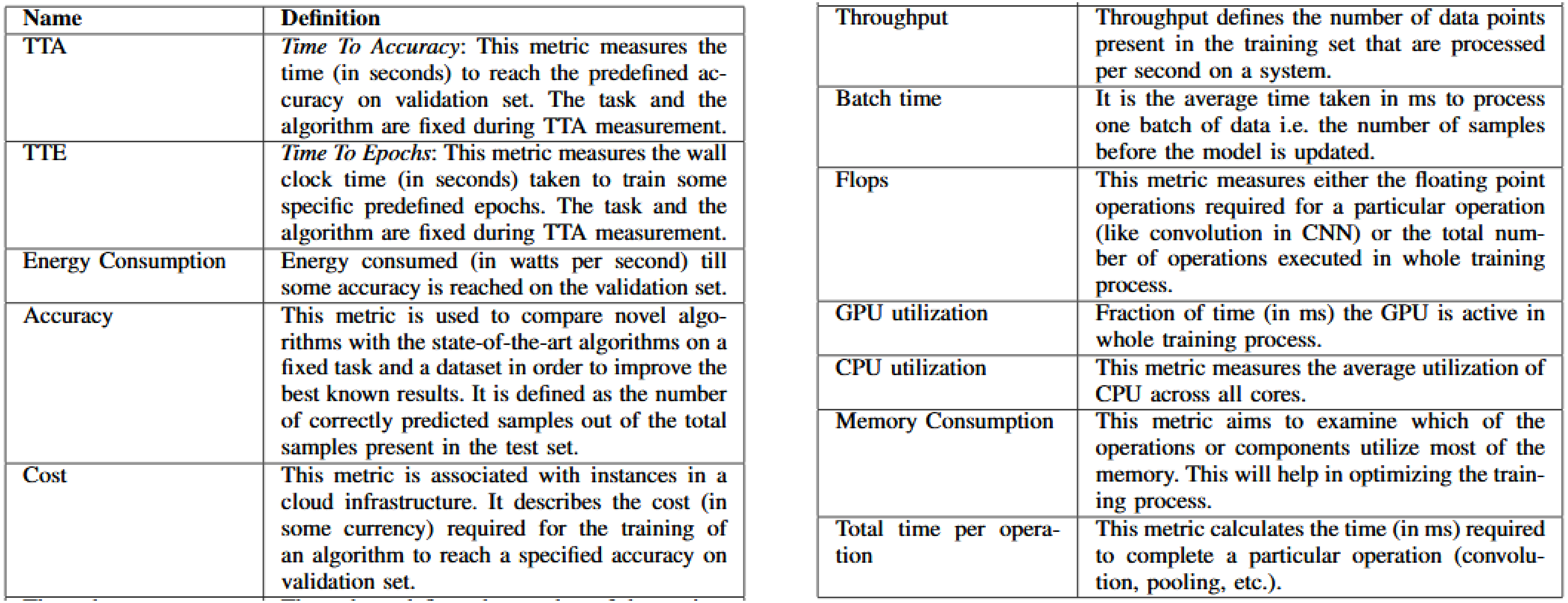
Standard Computing field:	Machine and deep learning:
Fixed set of tasks and metric.	The immense diversity of applications (domains), frameworks, metrics, and libraries.
Common standards exist, for e.g. SPEC [2].	No common standard set (still in development for e.g. MLPerf [3]).
End to end testing with a fixed set of programs.	No fixed process, different workflow for different domains.
Mostly, a deterministic metric exists in the standard computing field which is used to measure software and hardware capability. Additionally, multiple runs of the same benchmark often have low variability in between the results observed.	Stochasticity is involved in the whole process (lower precision training, etc.). The variation in results obtained with multiple runs of the algorithm is quite high.

Table 1: Comparison between a ML benchmark and a standard benchmark.

3) Properties of an ML Benchmark

- Should provide a fair comaprsion between hardware system.
- Should provide a fair comparison between frameworks.
- Should standardize a set of rules.
- Provide a quantitative analysis between system level operations.
- Should measure systems on the basis of scalability.
- Should be able to especially handle stochasticity involved in machine learning workloads.
- Should be representative of industrial needs as many Benchmarks.
- Should be transparent that providers of hardware or infrastructure accepting the benchmark
- Should be open source that everyone can validate the correctness of the implementation.

4) Metrics for an ML Benchmark



Name	Definition
TTA	*Time To Accuracy*: This metric measures the time (in seconds) to reach the predefined accuracy on validation set. The task and the algorithm are fixed during TTA measurement.
TTE	*Time To Epochs*: This metric measures the wall clock time (in seconds) taken to train some specific predefined epochs. The task and the algorithm are fixed during TTA measurement.
Energy Consumption	Energy consumed (in watts per second) till some accuracy is reached on the validation set.
Accuracy	This metric is used to compare novel algorithms with the state-of-the-art algorithms on a fixed task and a dataset in order to improve the best known results. It is defined as the number of correctly predicted samples out of the total samples present in the test set.
Cost	This metric is associated with instances in a cloud infrastructure. It describes the cost (in some currency) required for the training of an algorithm to reach a specified accuracy on validation set.
Throughput	Throughput defines the number of data points present in the training set that are processed per second on a system.
Batch time	It is the average time taken in ms to process one batch of data i.e. the number of samples before the model is updated.
Flops	This metric measures either the floating point operations required for a particular operation (like convolution in CNN) or the total number of operations executed in whole training process.
GPU utilization	Fraction of time (in ms) the GPU is active in whole training process.
CPU utilization	This metric measures the average utilization of CPU across all cores.
Memory Consumption	This metric aims to examine which of the operations or components utilize most of the memory. This will help in optimizing the training process.
Total time per operation	This metric calculates the time (in ms) required to complete a particular operation (convolution, pooling, etc.).

Table 2: The list of metrics which could be used by a ML benchmark.

5) Categorization of Benchmarks

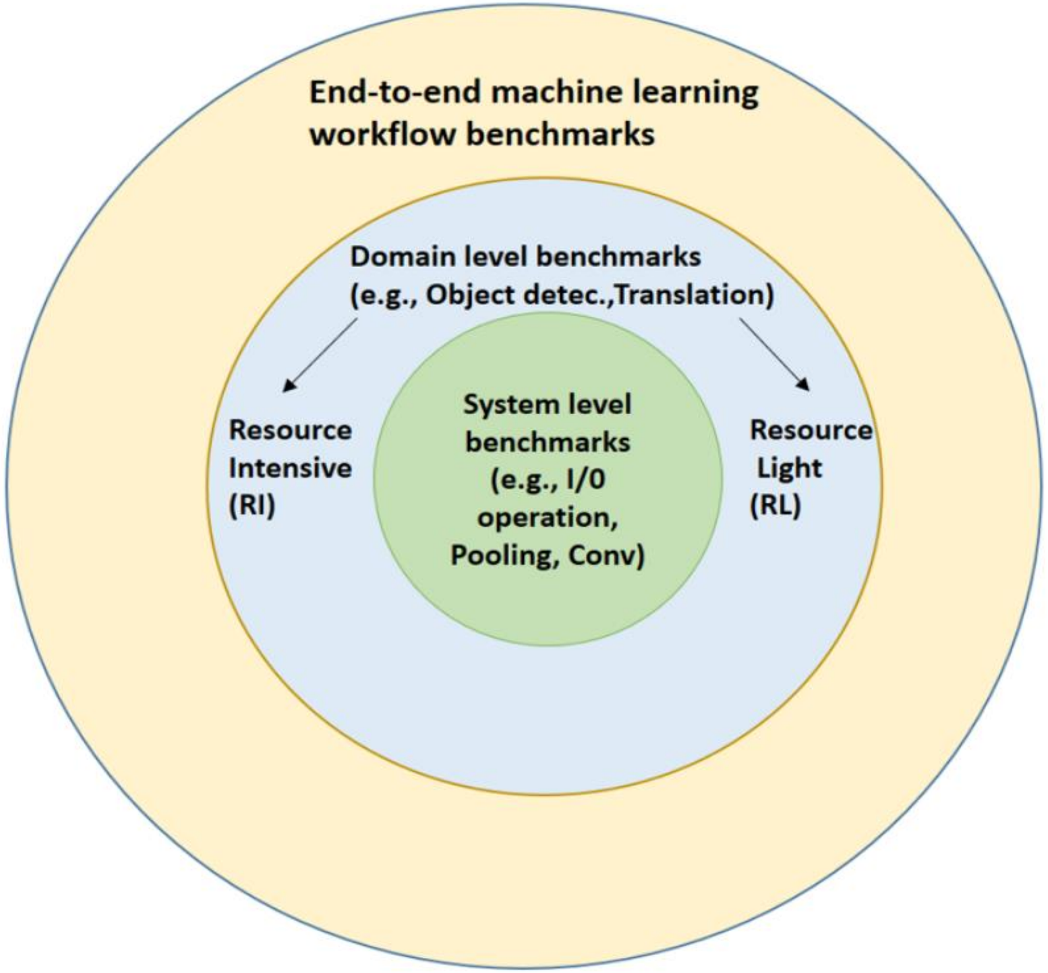

End-to-end machine learning workflow benchmarks
Domain level benchmarks (e.g., Object detec.,Translation)
Resource Intensive (RI)
System level benchmarks (e.g., I/0 operation, Pooling, Conv)
Resource Light (RL)

Fig 1: Category of ML Benchmarks.

6) Different Benchmarks

	DeepBench [6]	AIBench [5]	DLBS [4]	MLPerf (v0.7) [3]
Metrics	• Flops • Total time per operation (ms)	• TTA • TTE • Energy Consumption	• Throughput • Batch time (ms)	• TTA
Datasets	No real data is used for the benchmarks i.e. random numbers are generated in a fitting format.	17 Fixed datasets	Synthetic and real datasets.	6 fixed datasets
Domain	• GEMM operation • Convolutional layers • Recurrent layers • All Reduce	17 domains (Image Classification, Image Generation, Language translation ...)	• Language translation • Image classification	6 domains
Type	System Level	End-to-end	Domain level	Domain level

Table 4: A summary of four out of seven Benchmarks.

6) Conclusion

In this paper, we have summarized **seven prominent benchmarking suites** that help in making an informed decision about which hardware or software is the best for a specific application.

Some of the benchmarking suites are still in their development phases and in the future, they can accelerate further progress in their respective fields.

Different benchmark suites employ different metrics but there seems to be no agreement on a standardized set.

None of the benchmarks provide any implementation for domains like predictive maintenance which is highly relevant for the manufacturing industry

The benchmarks mentioned in this paper other than DAWNBench do not target the cloud platforms directly.

With the addition of more domains, the inclusion of cost as a metric, improvement on documentation, and support for cloud platforms; the **MLPerf** and **AIBench** benchmark suites have the potential to become the goto benchmarking suite for all ML applications.

7) Bibliography

[1]	https://en.wikipedia.org/wiki/Benchmark_(computing)#cite_note-1
[2]	https://www.spec.org/
[3]	https://arxiv.org/pdf/1910.01500.pdf
[4]	https://hewlettpackard.github.io/dlcookbook-dlbs/#/
[5]	https://www.benchcouncil.org/AIBench/
[6]	https://svail.github.io/DeepBench/

Vielen Dank für Ihre Aufmerksamkeit