MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance

---

Google Brain, NVIDIA, Harvard Univ., Intel, Microsoft, Facebook

IEEE Micro

Presenter: Jemin Lee

https://leejaymin.github.io/index.html

May 12, 2020

Neural Acceleration Study Season #2

Introduction

- The increasing use of DNN in industry is driving the rapid development of specialized hardware architecture and software frameworks.
- Need a performance benchmark to evaluate these competing ML systems.
  - SEPC benchmark (1988)
  - DeepBench
  - Fathom
  - DAWNBench
- MLPerf was founded in 2018 to combine the best of prior efforts: a broad benchmark set with a time-to-convergence metric and the support of an academic/industry consortium.

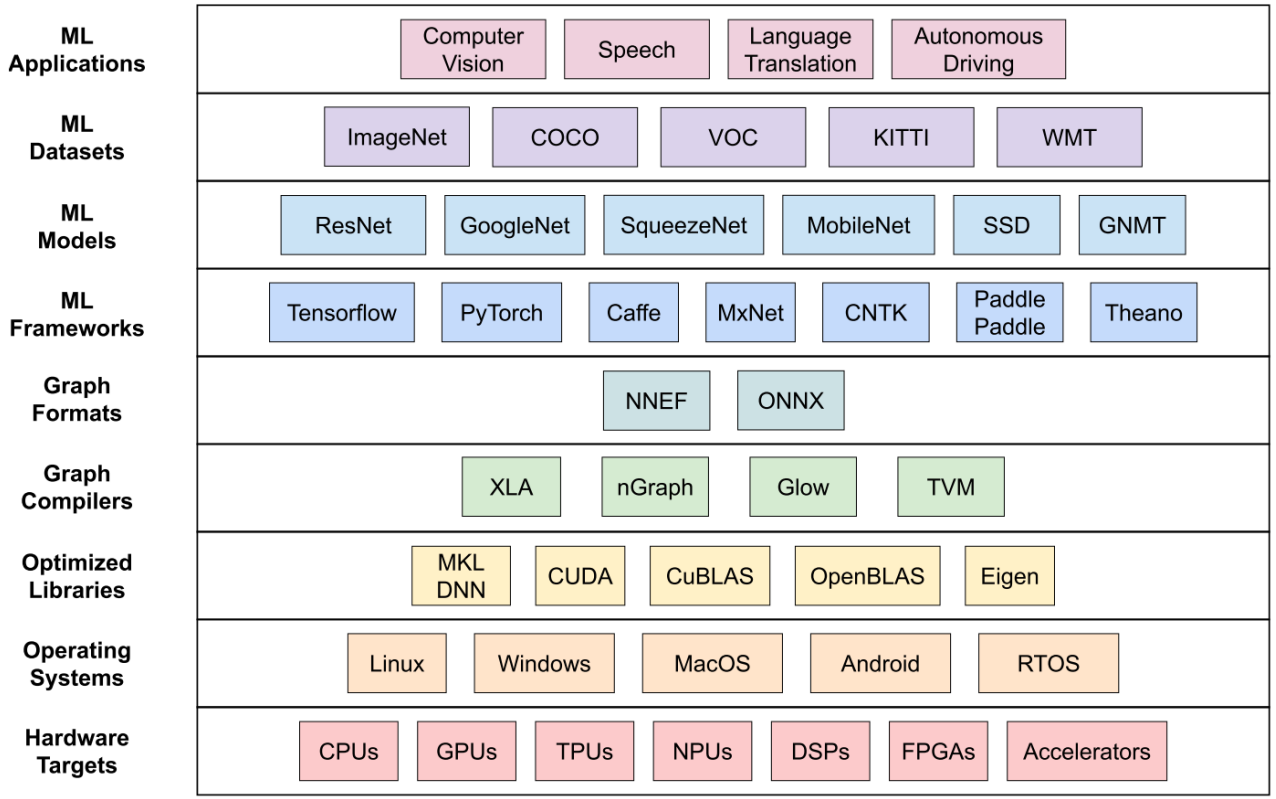
https://youtu.be/hQRBLW6giRc

Introduction (cont.)

MLPerf

Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services.

Inference / Training



Design Choices

- Implementation equivalence:
  - Reimplement the benchmark for a certain hardware
- Training hyperparameter equivalence
  - Which hyperparameters are tunable?
- Training convergence variance

- Inference weight equivalence
  - Pruning or quantization

MLPerf Training: Benchmark Definition

- Specify an MLPerf Training benchmark as training a model on a specific data set to reach a target quality
- Closed division: using a specific model for direct comparisons
- Open division: using any model to support model innovation

MLPerf Training: Metric Definition

- Throughput: the number of data processed per second
  - Do not train the model to completion
  - Relatively low variance because the compute cost per datum is const ant in most models
  - Increased at the cost of time-to-train by using optimizations like low er precision numeric or larger batch sizes
- Time to train: the wall clock time
  - It takes for the model to reach a target quality
- All metrics could be normalized by cost or power

MLPerf Training: Benchmark Selection(v0.5)

Table 1. MLPerf Training v0.5 Benchmarks.

Area	Problem	Data set	Model
Vision	Image recognition	ImageNet[7]	ResNet[7]
	Object detection	COCO[7]	SSD[7]
	Object segmentation	COCO[7]	Mask R CNN[7]
Language	Translation	WMT Eng.-German[7]	GNMT[7]
	Translation	WMT Eng.-German[7]	Transformer[7]
Commerce	Recommendation	Movielens-20M[7]	NCF[7]
Research	RL	Go, 9×9 board	MiniGo[7]

MLPerf Training: Implementation Equivalence

- No single ML frameworks supported by all architectures

- Require performing the same set of mathematical operations as the reference implementation to produce each output, usi ng the same optimizer to update the weights, and using the same preprocessing and evaluation methods

MLPerf Training: Hyperparameter Tuning

- Batch sizes
- Learning rate schedule
- Other optimizer hyperparameters
- The hyperparameter playing fields
  - Search limits: MLPerf limits which hyperparameters can be tuned
  - Hyperparameter borrowing:

MLPerf Training: Variance

- The time to train a model to a target quality has relatively hig h variance
- The time to train is roughly proportional to the number of epochs (passes over the training data) required
- The number of epochs required varies:
  - The starting weights
  - Nondeterministic floating-point ordering
- MLPerf balances variance and computation cost by averaging over a number of runs but still accepting relatively high variance

MLPerf Inference: Benchmark Definition

- MLPerf inference benchmark as processing a series of inputs to a trained model to produce outputs that meet a quality target

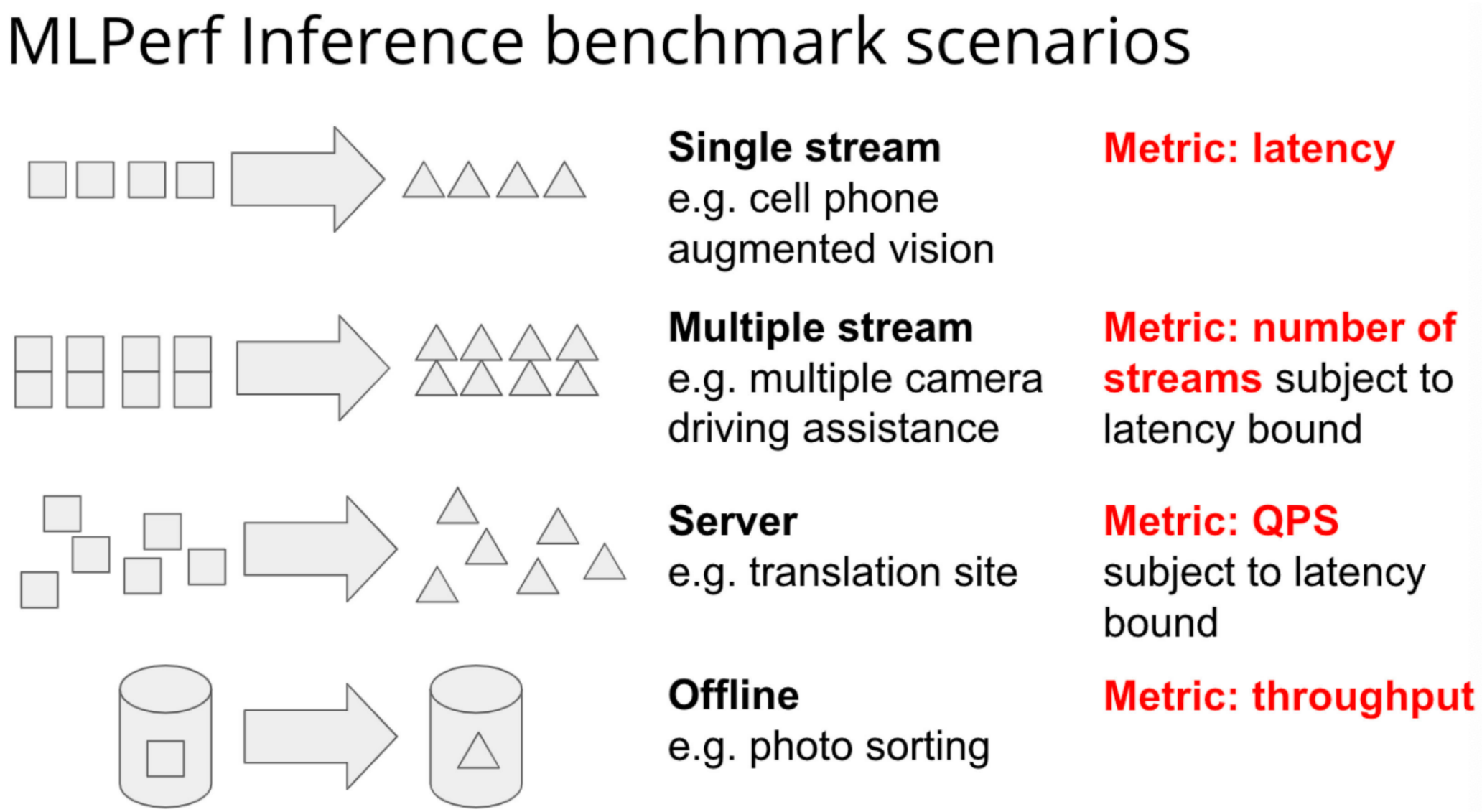


Figure 1. MLPerf inference scenarios and metrics.

Table 2. MLPerf Inference v0.5 Benchmarks.

Area	Task	Data set	Model
Vision	Image classification	ImageNet (224×224)[7]	Resnet50-v1.5[7]
Vision	Image classification	ImageNet (224×224)[7]	MobileNets-v1 224p[7]
Vision	Object detection	COCO (1200×1200)[7]	SSD-ResNet34[7]
Vision	Object detection	COCO (300×300)[7]	SSD-MobileNets-v1[7]
Language	Machine translation	WMT Eng.-German[7]	GNMT[7]

Area	Task	Model	Dataset	Quality	Server latency constraint	Multi-Stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	99% of FP32 (76.46%)	15 ms	50 ms
Vision	Image classification	MobileNets-v1 224	ImageNet (224x224)	98% of FP32 (71.68%)	10 ms	50 ms
Vision	Object detection	SSD-ResNet34	COCO (1200x1200)	99% of FP32 (0.20 mAP)	100 ms	66 ms
Vision	Object detection	SSD-MobileNets-v1	COCO (300x300)	99% of FP32 (0.22 mAP)	10 ms	50 ms
Language	Machine translation	GNMT	WMT16	99% of FP32 (23.9 BLEU)	250 ms	100 ms

Quantization, Retraining, and Sparsity

- Not allowing retraining or sparsification
  - 32-bit floating point weights

- MLPerf Inference achieves a quality target within 1% of the FP32 reference model's accuracy.
  - The closed division: INT4, INT8, INT16, UINT8, UNIT16, FP11 (sign, 5 bit mantissa, and 5bit exponent), FP16, bfloat16, and FP32

Presentation

- Results or Single Summary Score
- Scale Information and Normalization

Results

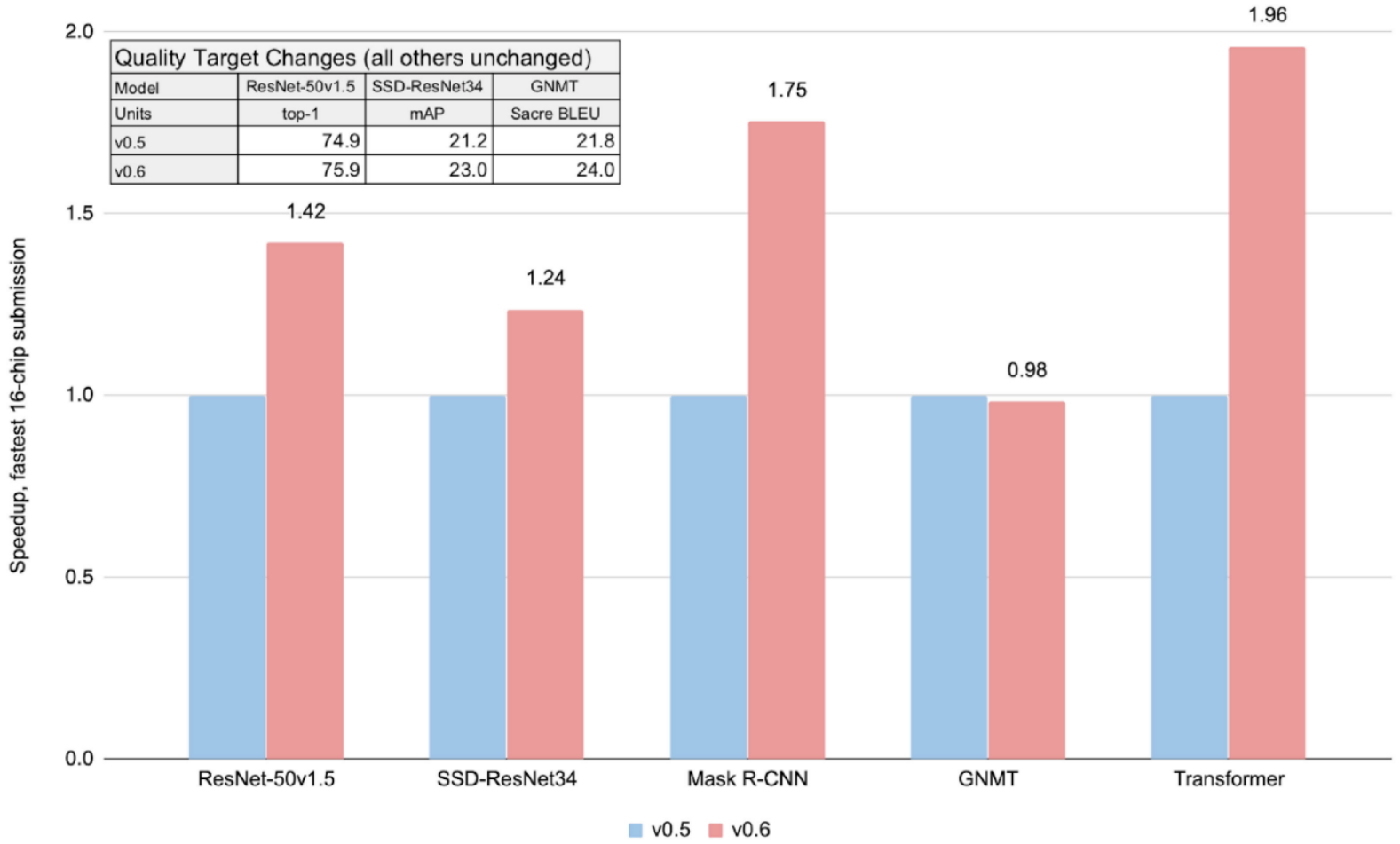


Figure 2. Speedup in the fastest 16-chip entry from MLPerf Training version v0.5 to v0.6, despite more timed work due to increased quality targets as shown.

Results

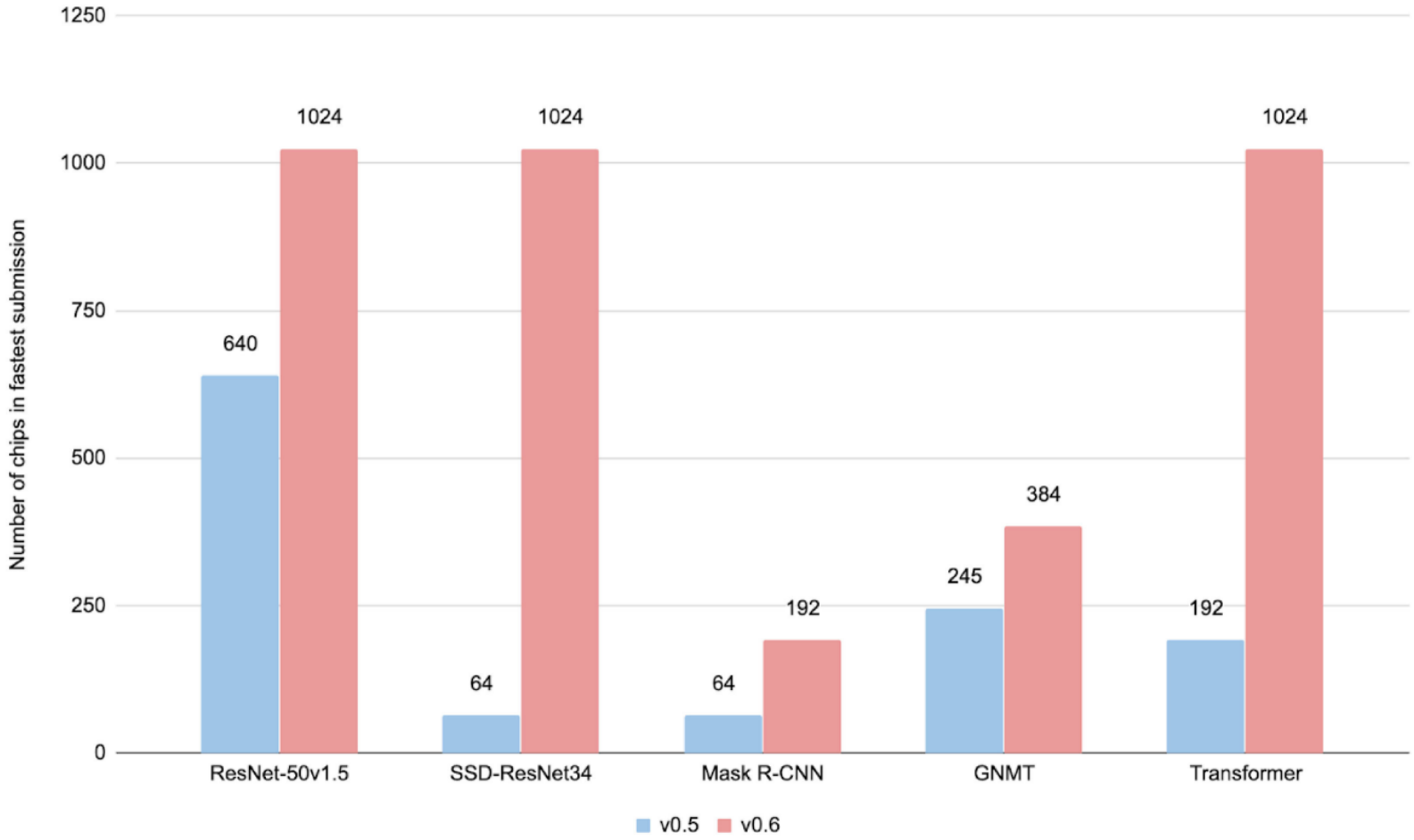


Figure 3. Increase in the number of chips used in the system that produced the fastest overall score from MLPerf Training version v0.5 to v0.6.

Results

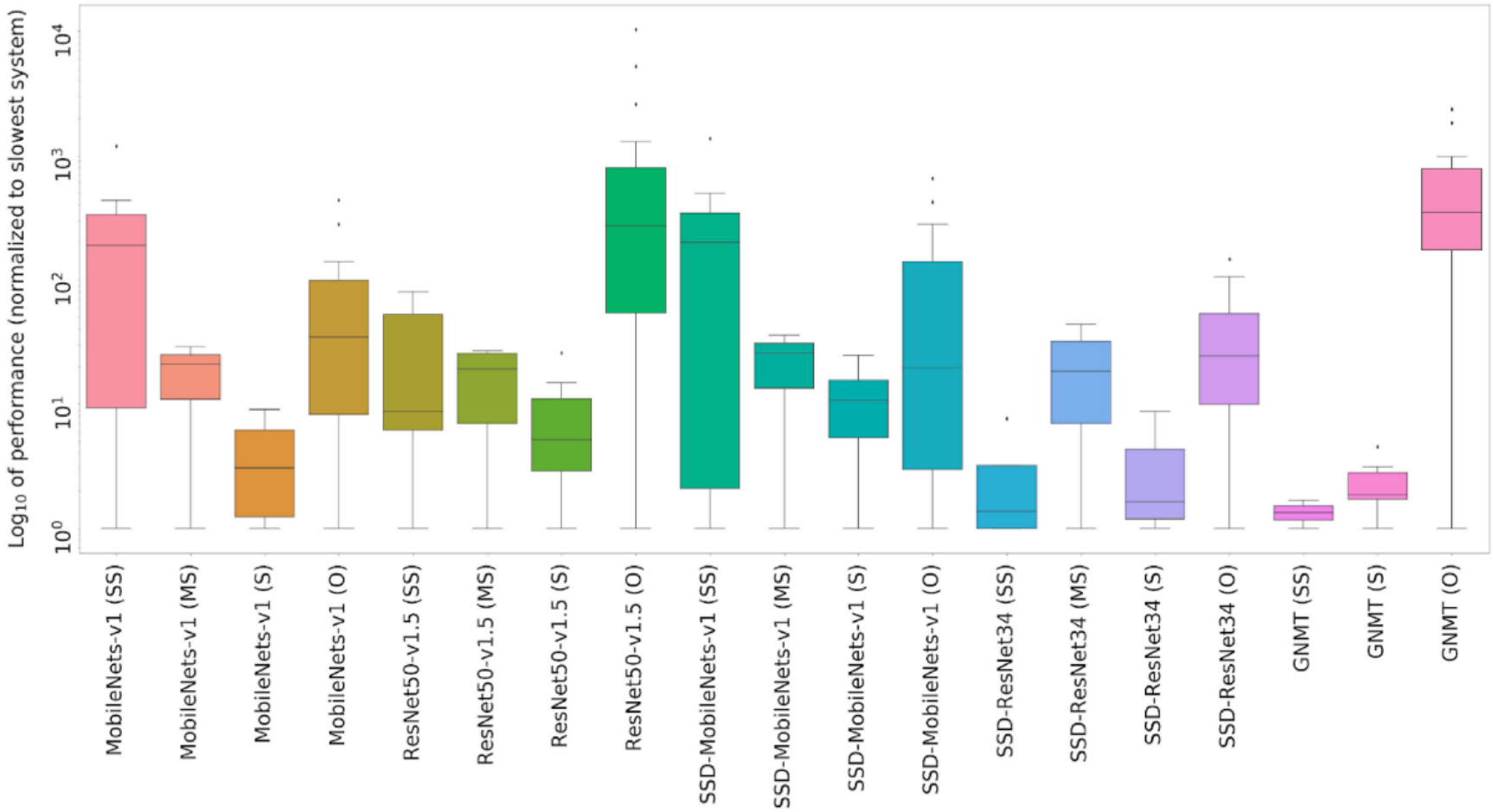


Figure 4. Normalized performance distribution in log scale from results in the closed division.

MLPerf-reference implementation

- Reference models using 32-bit floating-point weights
  - 실행 엔진은 TensorFlow, PyTorch, ONNXRuntime 제공

README.md

MLPerf Inference Benchmarks for Image Classification and Object Detection Tasks

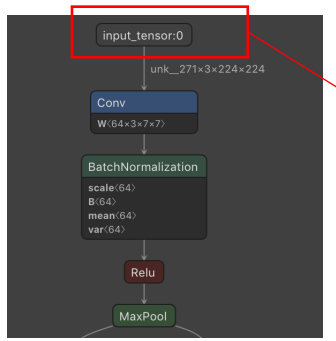
This is the reference implementation for MLPerf Inference benchmarks.

You can find a short tutorial how to use this benchmark here.

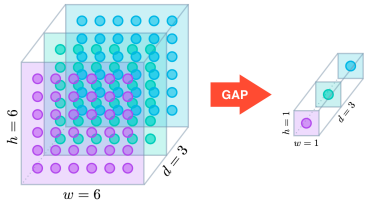
Supported Models

model	framework	accuracy	dataset	model link	model source	precision	notes
resnet50-v1.5	tensorflow	76.456%	imagenet2012 validation	from zenodo	mlperf, tensorflow	fp32	NHWC. More information on resnet50 v1.5 can be found here.
resnet50-v1.5	onnx, pytorch	76.456%	imagenet2012 validation	from zenodo	from zenodo converted with this script	fp32	NCHW, tested on pytorch and onnxruntime
mobilenet-v1	tensorflow	71.676%	imagenet2012 validation	from zenodo	from tensorflow	fp32	NHWC
mobilenet-v1 quantized	tensorflow	70.694%	imagenet2012 validation	from zenodo	from tensorflow	int8	NHWC
mobilenet-v1	tflite	71.676%	imagenet2012 validation	from zenodo	from tensorflow	fp32	NHWC

ONNX-MLPerf-ResNet1.5-50 검증 및 실행 결과



MLPerf-Resnet1.5-50



GlobalAveragePool

Opset11: ReduceMean
-axes: 2,3
-keepdims: 1

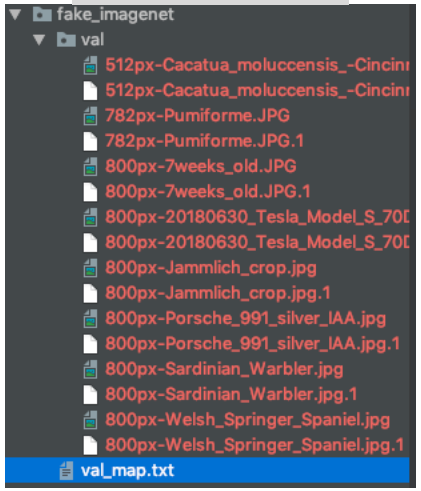
GEMM

ArgMax:0, [818]
Softmax_tensor:0, [[5.4259047e-10 8.7503516e-10 3.5769931e-10 ... 3.7001036e-10
1.3894656e-08 1.7087004e-08]]

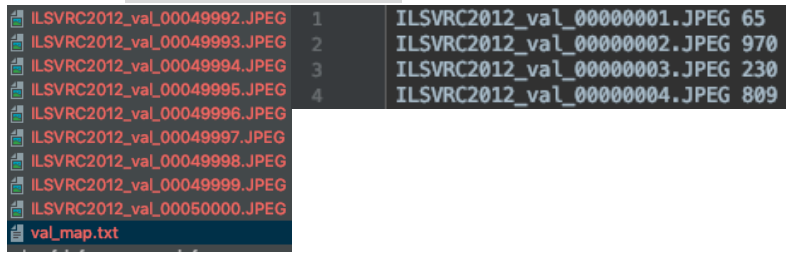
Top-1: class=n04285008 sports car, sport car ; probability=0.838233
Top-2: class=n03100240 convertible ; probability=0.097503
Top-3: class=n03459775 grille, radiator grille ; probability=0.021509
Top-4: class=n02814533 beach wagon, station wagon, wagon, estate car, beach waggon, station waggo
n, waggon ; probability=0.021157
Top-5: class=n02974003 car wheel ; probability=0.017151

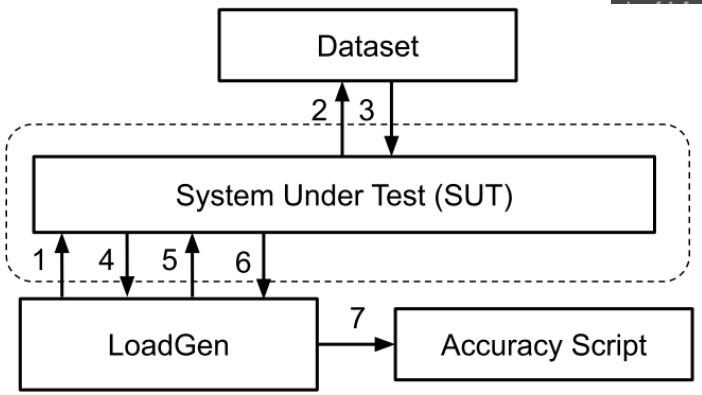
Imagenet2012, validation set 5만장으로 top#1 acc=76.456% 재현됨

Fakeimagenet



Imagenet





https://mlperf.org/inference-results

INFO:imagenet:loaded 8 images, cache=0, took=0.0sec INFO:main:starting TestScenario.SingleStream TestScenario.SingleStream qps=25.31, mean=0.0356, time=0.316, acc=75.000%, queries=8,
tiles=50.0:0.0355,
80.0:0.0360,
90.0:0.0365,
95.0:0.0370

INFO:imagenet:loaded 50000 images, cache=0, took=889.9sec
INFO:main:starting TestScenario.SingleStream TestScenario.SingleStream qps=1601.22, mean=0.0370, time=31.226, acc=76.456%, queries=50000,
tiles=50.0:0.0355,
80.0:0.0373,
90.0:0.0393,
95.0:0.0427,
99.0:0.0746,
99.9:0.1191

Conclusion

- MLPerf inference and training are both driven by active working groups (WGs): a submitter's WG that maintains the rules, a special topics WG that explores deep technical issues, and a results WG that handles submission review and results presentation.
- Developing a long-term benchmark roadmap
- The MLPerf effort is now supported by more than 65 companies and researchers from eight educational institutions

Thank you