TSG-AI TS.47 – AI Mobile Devices - Change Request for Sect 3.1 – AI Hardware Requirements – Measuring Meaningful AI Performance

James Goel
Qualcomm Inc.

November 25th, 2019

TS.47 v0.9 Line 49

AI mobile device hardware is required to support AI software applications efficiently.

https://www.gsma.com/aboutus/workinggroups/wp-content/uploads/2019/09/TS.47-CR1001-1.docx

Current Section 3.1 – AI Hardware Proposal
Modified VGG TOPS Baseline
Unrelated to Efficient AI software applications

https://www.gsma.com/aboutus/workinggroups/wp-content/uploads/2019/09/TS.47-CR1001-1.docx

AI computational resources evolve to match efficient AI Neural Network models

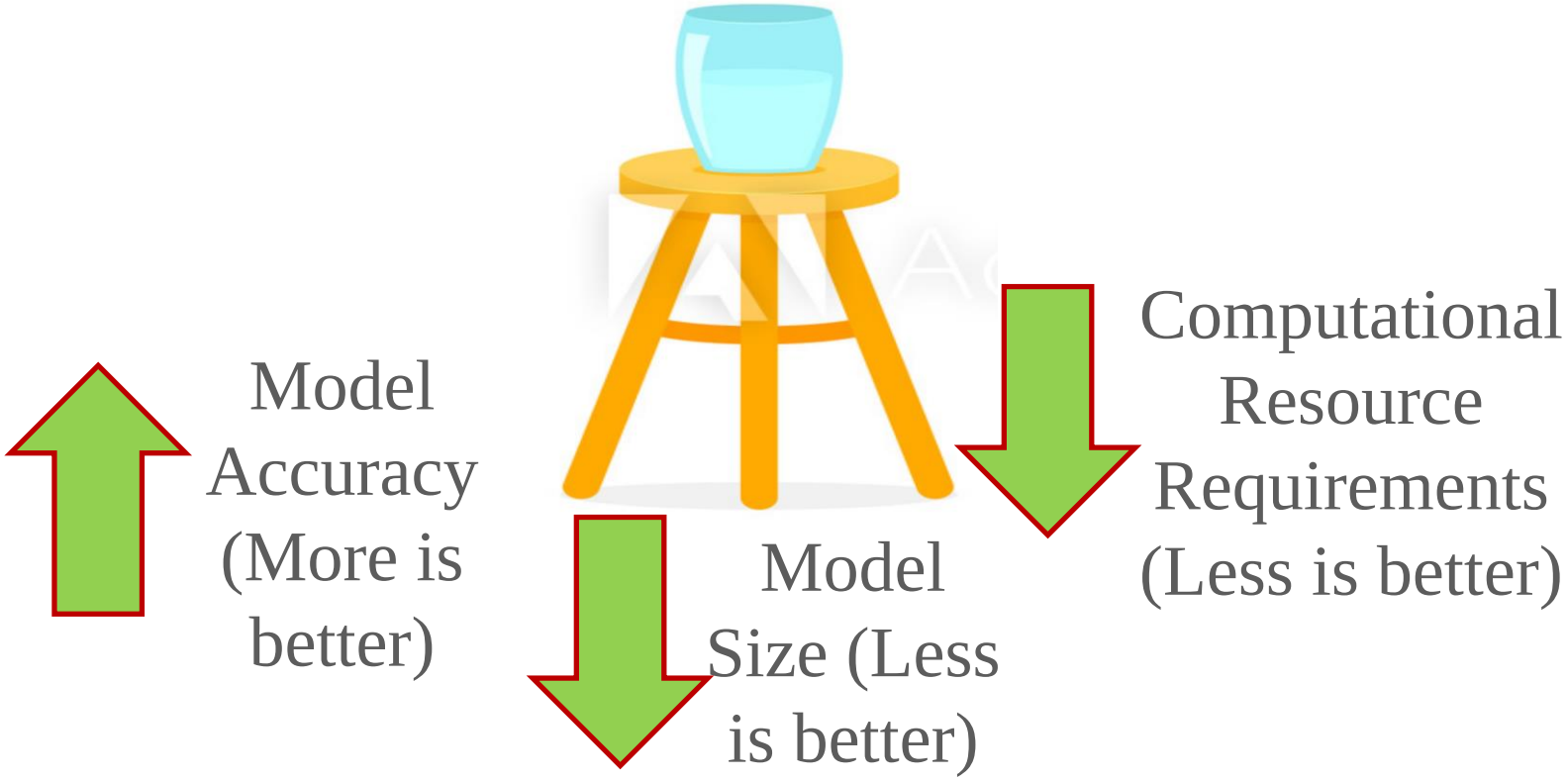


Figure 1: https://arxiv.org/abs/1810.00736

Mlperf.org

https://www.mlperf.org – Scroll to middle screen

Mlperf.org

Harvard University

Stanford University

University of Arkansas, Littlerock

University of California, Berkeley

University of Illinois, Urbana Champaign

University of Minnesota

University of Texas, Austin

University of Toronto

https://www.mlperf.org – Scroll to middle screen

Efficient region

NG
Not Good

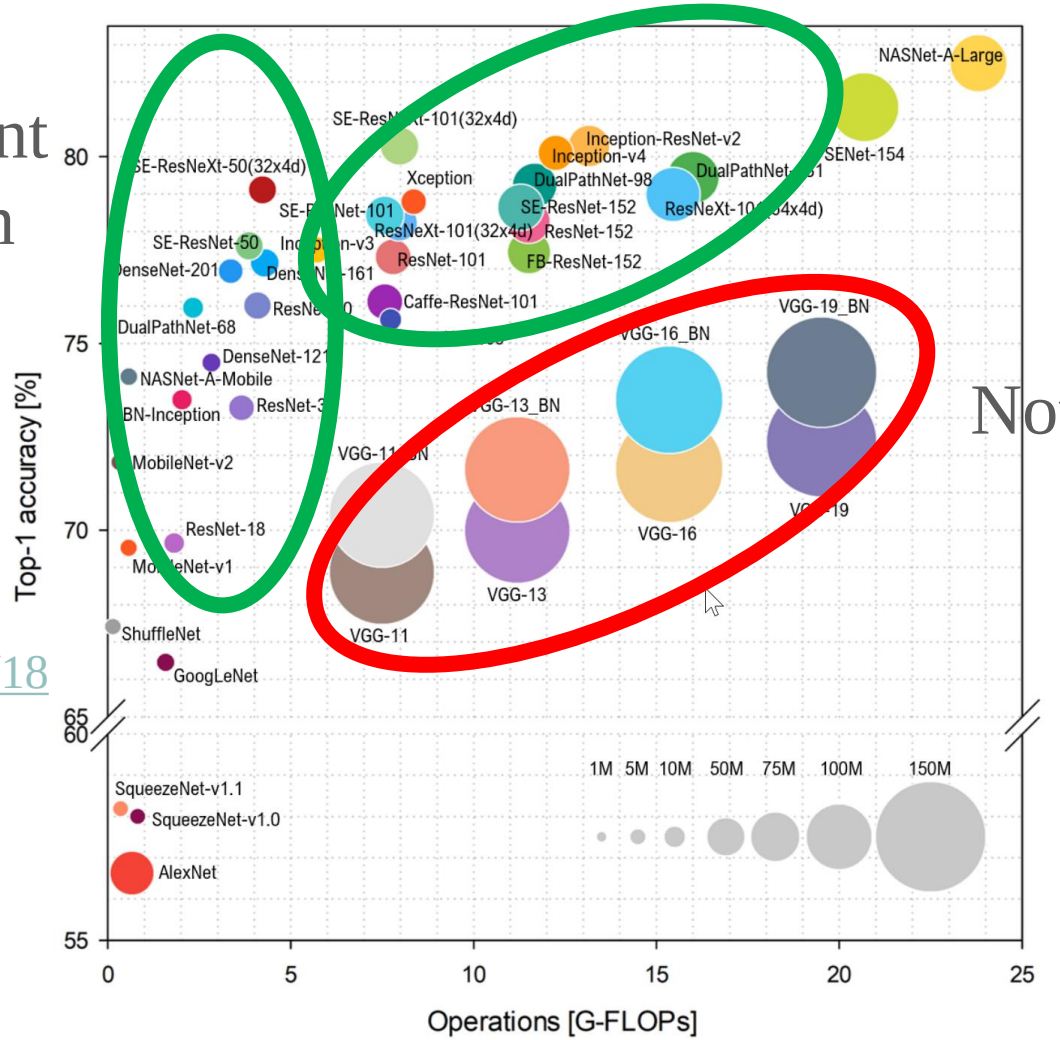


Figure 1: https://arxiv.org/abs/1810.00736

Solution: GSMA TSG-AI to use mlperf.org Networks and Inferencing Benchmark

Benchmarks

Each MLPerf Inference benchmark is defined by a model, a dataset, a quality target, and a latency constraint. The following five benchmarks are in version v0.5 of the suite:

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	GMNT	WMT16	99% of FP32 (23.9 BLEU)	250 ms	100 ms

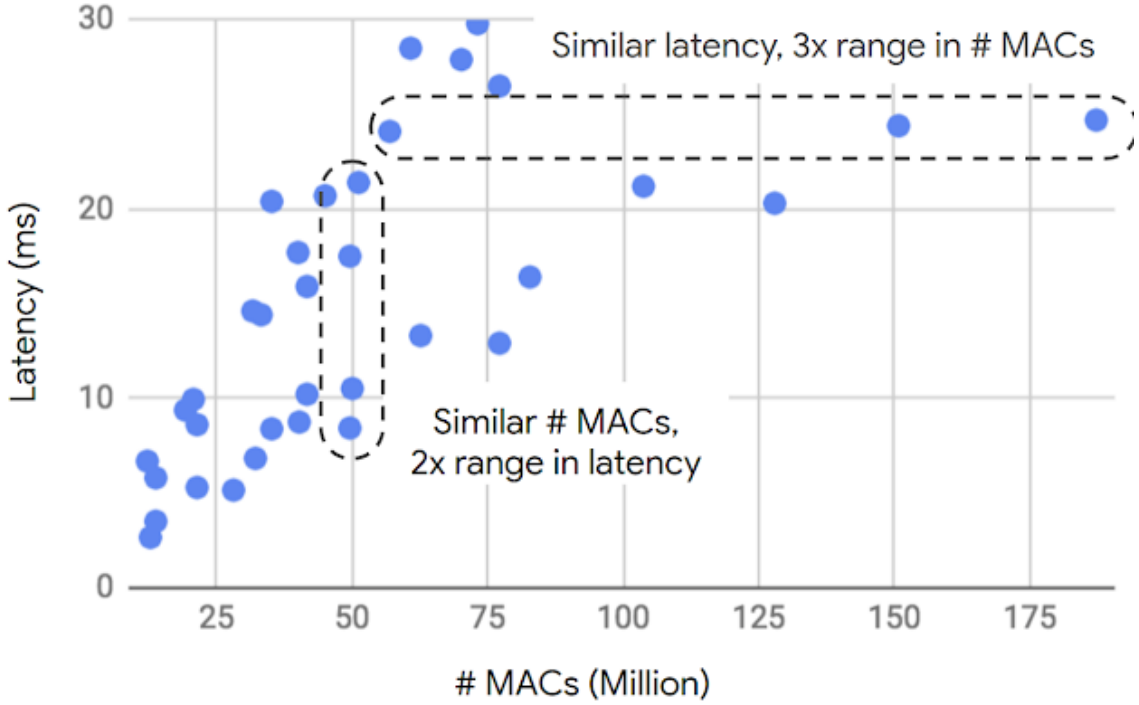
https://mlperf.org/inference-overview

GSMA and mlperf.org Liaison Relationship

- Provide GSMA TSG-AI feedback to mlperf.org
- Work with standards org to develop Section 3.1 Hardware Requirements

https://www.mlperf.org – Scroll to middle screen

TOPS/MACs are unrelated to AI Performance



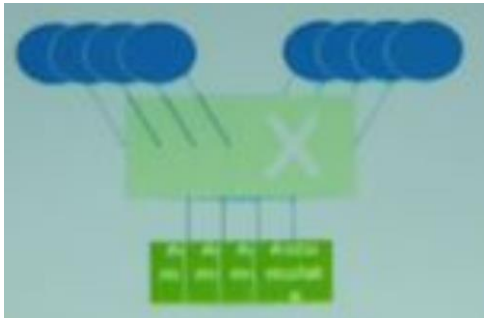
https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html

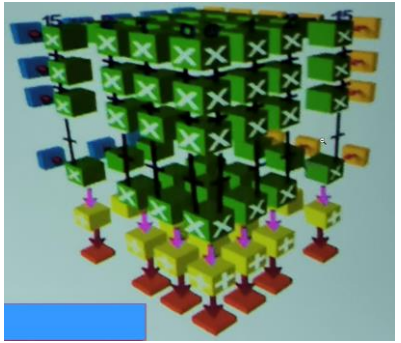
TOPS/MACs are unrelated to AI Performance
Why?

https://www.anandtech.com/show/14756/hot-chips-live-blogs-huawei-da-vinci-architecture

Building Blocks and their Computational Intensity

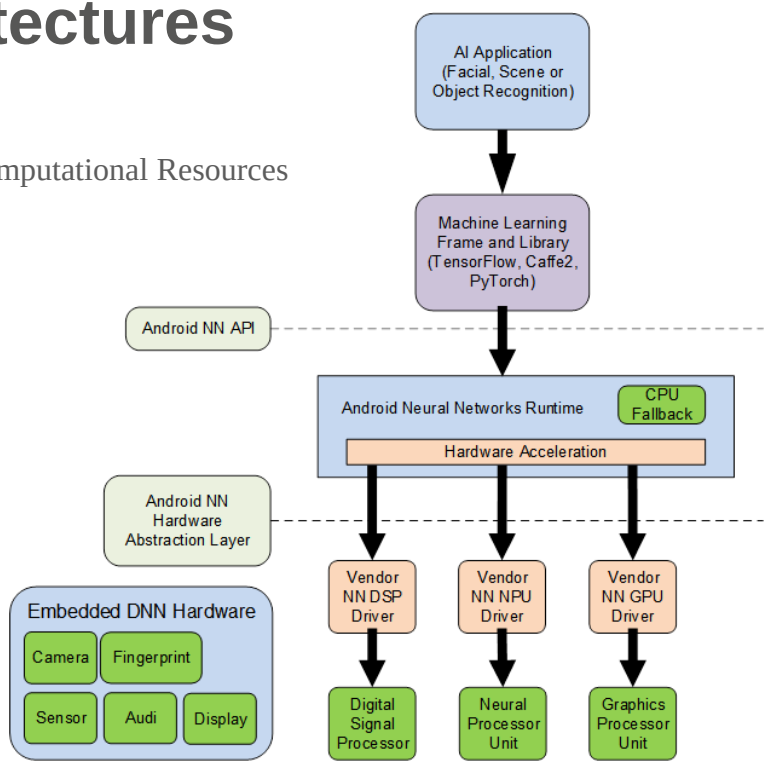
1D Scalar Unit + 2D Vector Unit + 3D Matrix Unit





VGG TOPS analysis is not accurate in modern architectures

Figure 1 - Multiple AI Computational Resources



https://developer.android.com/ndk/guides/neuralnetworks

Modified VGG TOPS Baseline
Unrelated to Efficient AI software applications

Modified VGG Network is non-standard

This definition?

[Link] REMODEL: Rethinking Deep CNN Models to Detect and Count on a NeuroSynaptic System

Or this definition?

[Link] Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size

Or something else?

Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size

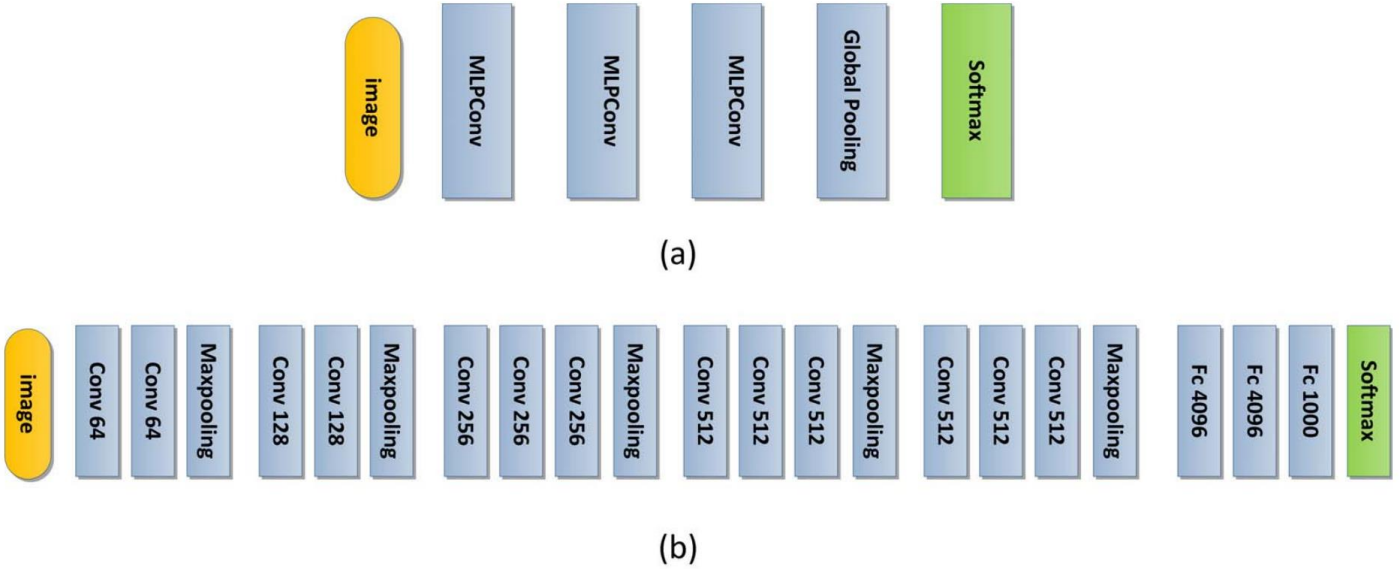
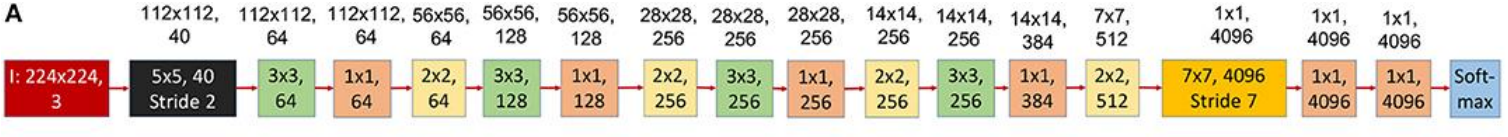


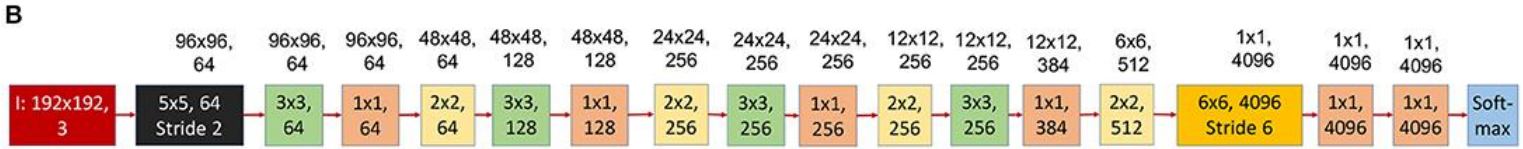
Figure 1. Two famous CNNs used to fit small datasets and big datasets, respectively. (a) Network in Network. Used to fit CIFAR-10 and CIFAR-100 (b)VGG-net. Used to fit ImageNet. People always think that very deep network should only used to fit giant dataset like ImageNet.

[Link] Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size

REMODEL: Rethinking Deep CNN Models to Detect and Count on a NeuroSynaptic System



Modified VGG-16 neural network model (1)



Modified VGG-16 neural network model (2)

Input image size

3x3 Image convolution layer with padding of 1 and stride 1. This layer is followed by TN defined activation function.

Pre-processing layer for input image. 5x5 convolution with padding of 2 and stride 2. This layer is followed by TN defined activation function.

1x1 convolution with no padding and stride of 1. This layer is followed by TN defined activation function.

Pooling Layer (2x2 convolution with stride of 2). This layer is followed by TN defined activation function.

Softmax classification output

7x7 (6x6) convolution with no padding and stride of 7 (or 6). This layer is followed by TN defined activation function.

[Link] REMODEL: Rethinking Deep CNN Models to Detect and Count on a NeuroSynaptic System

Inferences/sec is better that TOPS

Mlperf.org Inferences per Second is a **direct measurement of AI Hardware Performance**

However, mlperf.org has over 80 companies and universities involved in the selection of useful networks

https://www.mlperf.org – Scroll to middle screen

Background AI Benchmark References

- **Benchmark Analysis of Representative Deep Neural Network Architectures**
  - https://arxiv.org/abs/1810.00736
- Google Neural Networks API
  - https://developer.android.com/ndk/guides/neuralnetworks
- **A Modular Benchmarking Infrastructure for High-Performance and Reproducible Deep Learning**
  - Benchmark Analysis - https://arxiv.org/abs/1901.10183
- https://mlperf.org/inference-results
- https://mlperf.org/press#mlperf-inference-v0.5-results
- https://mlperf.org/inference-overview

Problem with Current VGG and TOPS

- TOPS and TOPS/Watt are not good assessments of on-device AI Application Performance (see next slides)
- VGG is not a good network model for on-device computational resource measurement and TOPS (see next slides)
- A better approach is to use a Standardized Benchmark
  - Mlperf.org has broad industry support

TOPS and TOPS/Watt Problems

- It is not a good idea to have our specification force device makers to have ***hardware that can run any type of model*****.** It is critical that device makers build hardware that runs important, practical, relevant and **useful** models. Hardware that runs ***any type of model*** will lead to inefficient, large, expensive and power-hungry implementations that mobile carriers will not accept. TS.47 should specify a requirement that pushes the industry to meet the requirements of GSMA members based on the AI Mobile Device ability to execute AI Applications effectively. ***Computational resources and accuracy are two different things*** and because these are different, GSMA TS.47 shall require the AI mobile device makers to ensure that they only build computational resources that return accuracy for ***relevant and important models*** stated above. We shall discourage all computational resources that demonstrate accuracy on **irrelevant networks**. As stated above, TOPS and TOPS/Watt **are not related** to accurate, practical, relevant and useful AI Application models. The most accurate way to assess on-device AI application performance is to use a benchmark system like mlperf.org that performs a comprehensive evaluation of many parameters using relevant networks.
- Consider Figure below. This figure illustrates a system that can execute multiple AI DNN models in parallel by using either dedicated AI hardware units and/or programmable CPU, DSP, NSP and GPU resources. Executing a single VGG network on this type of system architecture would not yield valid results.

Copied from -- GSMA TSG-AI e-mail reflector

VGG Network Problems

- The VGG network is very large and does not stress the maximum capabilities of the on-device computational resources. The following figure 1 analyzes and compares a wide variety of DNN models. Please note the large number of smaller mobile networks (SENet-154 and NASNet-A-Large) that require more computational resources (Gflops/sec) than VGG. The large 150 Megabyte size of VGG emphasizes memory transaction throughput over on-device computational resources and does not represent a realistic network used by AI Applications.
-

Background AI Benchmark References

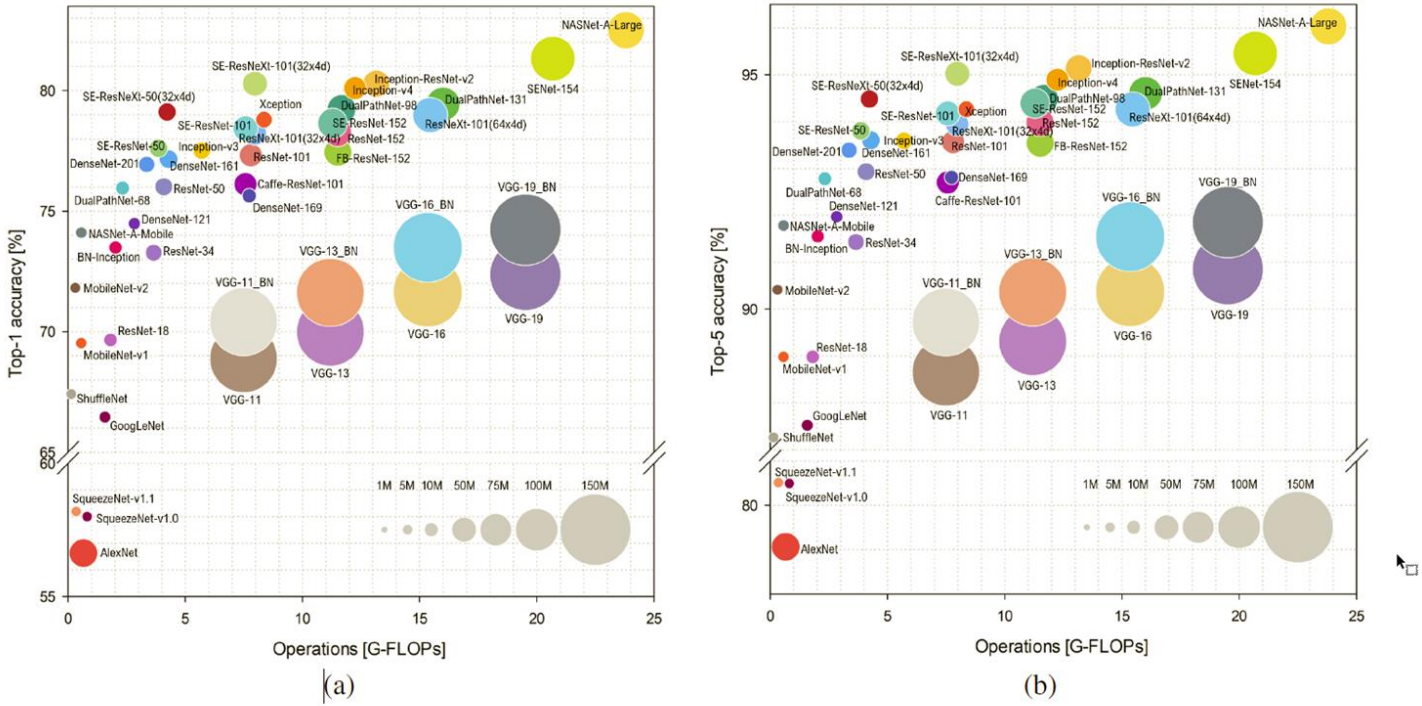


FIGURE 1: Ball chart reporting the Top-1 and Top-5 accuracy *vs.* computational complexity. Top-1 and Top-5 accuracy using only the center crop versus floating-point operations (FLOPs) required for a single forward pass are reported. The size of each ball corresponds to the model complexity. (a) Top-1; (b) Top-5.

Figure 1: https://arxiv.org/abs/1810.00736

Mlperf.org – AI Benchmark

Open Source With Broad Industry Support

Benchmarking Performance

Starting to move towards standardized benchmark

- Popular closed, non-standard AI benchmarks
- AI-Benchmark.com (ETH Switzerland, Huawei)
- AITUTU Benchmark (Antutu related mobile phone app)
- Ludashi AI Benchmark (https://www.ludashi.com/page/pc.php)
- China Telecom internal benchmarks
  - Uploaded to GSMA TSG-AI documents
- https://openai.com/progress/

Mlperf.org

https://www.mlperf.org – Scroll to middle screen

Mlperf.org

Harvard University

Stanford University

University of Arkansas, Littlerock

University of California, Berkeley

University of Illinois, Urbana Champaign

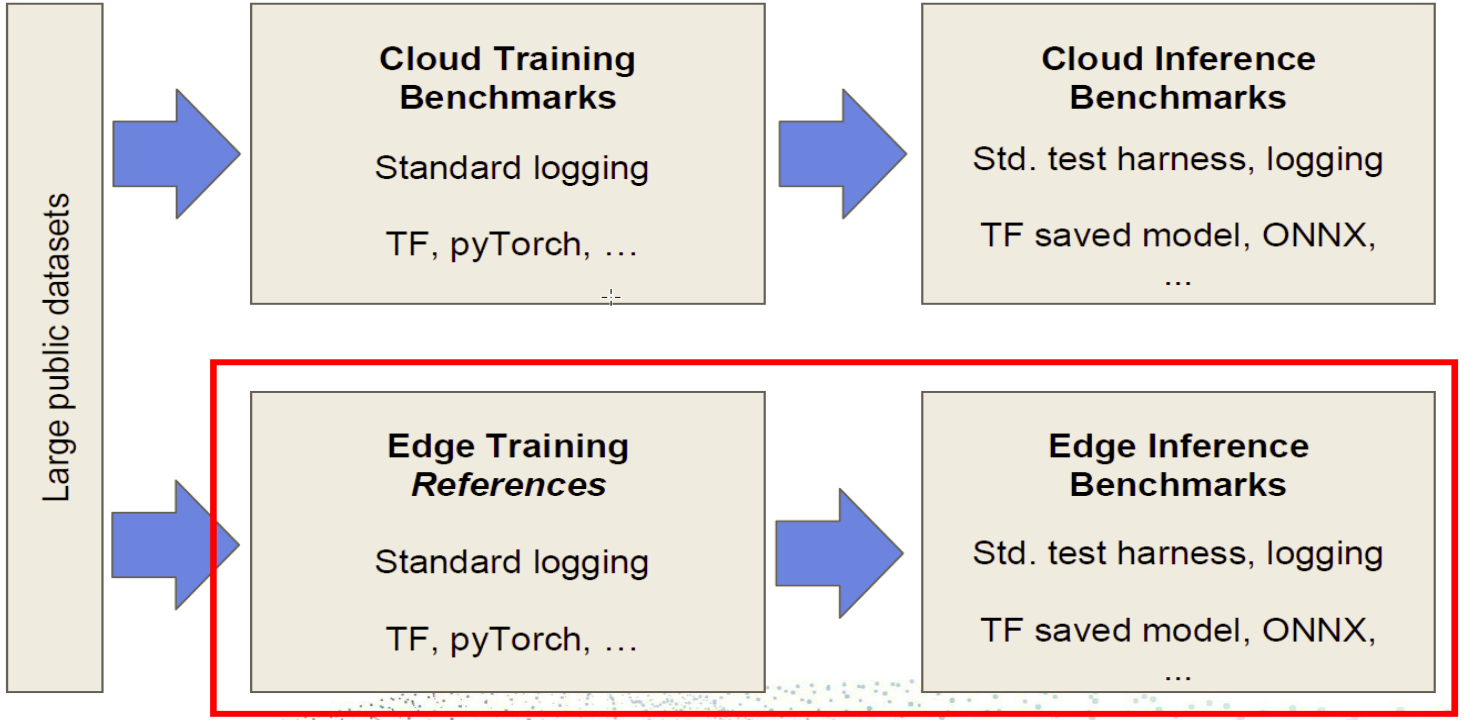
University of Minnesota

University of Texas, Austin

University of Toronto

https://www.mlperf.org – Scroll to middle screen

MLPerf
(as of Dec 10)



MLPerf

http://prof.ict.ac.cn/Bench18/mlperf.pdf

Solution: GSMA TSG-AI to use mlperf.org Networks and Inferencing Benchmark

Benchmarks

Each MLPerf Inference benchmark is defined by a model, a dataset, a quality target, and a latency constraint. The following five benchmarks are in version v0.5 of the suite:

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	GMNT	WMT16	99% of FP32 (23.9 BLEU)	250 ms	100 ms

https://mlperf.org/inference-overview

Strict Rules for Results Reporting

- **2.1. Strive to be fair**
- **2.2. System and framework must be consistent**
- **2.3. System and framework must be available**
- **2.4. Benchmark implementations must be shared**
- **2.5. Non-determinism is restricted**
- **2.6. Benchmark detection is not allowed**
- **2.7. Input-based optimization is not allowed**
- **2.8. Replicability is mandatory**
- Results that cannot be replicated are not valid results.

https://github.com/mlperf/inference_policies/blob/master/inference_rules.adoc

Public Results Reporting

AI Training and Inferencing Benchmarks

Closed Division Times																	
ID	Submitter	System	Benchmark results (Single Stream		Processor	#	Accelerator	#	Software	Form Factor (Mobile/ Handheld,				Details	Code	Notes
			Image classification													
			ImageNet MobileNet-v1 Stream	ImageNet ResNet-50 v1.5 Stream						m	d	s	e			
CATEGORY: Available																
Inf-0.5-7	dividiti	Linaro HiKey960 (hikey960)	121.11	518.07	HiSilicon Kirin960	1			TFLite v1.15.0-rc2	x			x	details	code	Mobile chip in embedded form factor (development board).
Inf-0.5-10	dividiti	Huawei Mate 10 Pro (mate10pro)	74.2	354.13	HiSilicon Kirin970	1	Arm Mali-G72 MP12	1	ArmNN v19.08 (OpenCL)	x				details	code	
Inf-0.5-11	dividiti	Huawei Mate 10 Pro (mate10pro)	111.6	494.92	HiSilicon Kirin970	1			ArmNN v19.08 (Neon)	x				details	code	
Inf-0.5-28	NVIDIA	NVIDIA Jetson AGX Xavier (Xavier)	0.58	2.04	NVIDIA Carmel (ARMv8.2)	1	NVIDIA Xavier	1	TensorRT 6.0, Jetpack 4.3-DP, CUDA 10.0, cuDNN 7.6.3				x	details	code	GPU and both DLAs are used in Offline and MultiStream scenarios
Inf-0.5-29	Qualcomm	SDM855 QRD	3.02	8.95	Qualcomm Kryo485	1	Qualcomm Hexagon 690 Processor: Hexagon Vector Extensions (HVX), Hexagon	1	Snapdragon Neural Processing Engine (SNPE) V1.30	x				details	code	Hexagon Vector Extensions (HVX) being used

https://mlperf.org/inference-results/

Open Source Tools and Code

AI Training and Inferencing Benchmarks

- Extensive Github Repository
- Google backed mlperf APK
- https://github.com/mlperf/mobile_app/tree/master/prebuilt

TS.47 Section 3.1
AI Hardware Requirements (Normative)

- Justification for this important MNO use-case:
  - **Evaluate** AI mobile device hardware performance to distinguish ***AI*** devices from ***traditional*** non-AI mobile terminals
  - **Identify good, better and best tiers** of AI mobile devices based on hardware performance
  - Evaluate support for key hardware support of ***AI software applications.***
- Requires a Compliance Test Specification
  - AI benchmark to accurately measure AI Hardware performance