tinyMLPerf

Benchmarking Resource-constrained Machine Learning Systems

Colby Banbury, Max Lam, Vijay Janapa Reddi, David Kanter, Amin Fazel, Xinyuan Huang, Danilo Pietro Pau and the tinyMLPerf working group

Harvard University, MLPerf, Samsung Semiconductor, Inc., Cisco Systems, STMicroelectronics

Abstract

Advancements in ultra-low-power machine learning (tinyML) hardware promises to unlock an entirely new class of intelligent applications. However, the complexity and dynamicity of the field obscure the measurement of progress and make application design decisions intractable. In order to enable the continued innovation, a fair, replicable and robust method of comparison is needed. Since progress is often the result of increased hardware capability, a reliable tinyML hardware benchmark is required.

To fulfill the need, we have created a community effort to extend the scope of the existing MLPerf benchmarking suite to include tinyML devices. With the help of over 75 member organizations, this group, dubbed tinyMLPerf, has begun the process of developing a benchmarking suite.

Existing Benchmarks

Existing benchmarks do not represent ML workloads or they are too large to fit on tinyML constrained processors.

Benchmark	ML?	Power?	Tiny?
CoreMark	×	√	√
MLMark	√	×	×
MLPerf Inference	√	√	×
TinyML Requirements	√	√	√

Survey of tinyML Use Cases, Models, and Datasets

The landscape of tinyML use cases is large and wide. We surveyed many state of the art use cases to determine the scope of a representative tinyMLPerf benchmark.

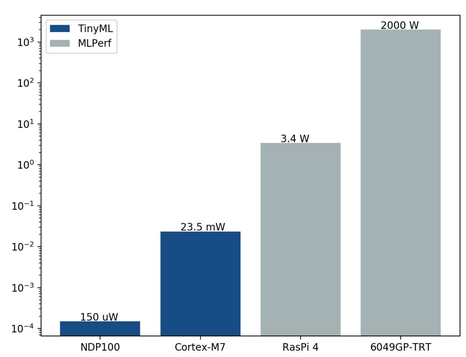
Input Type	Use Case	Model Type	Dataset
Audio	Audio Wake Words Context Recognition Control Words Keyword Detection	DNN CNN RNN LSTM	Speech Commands Audioset ExtraSensory Freesound DCASE
Image	Visual Wake Words Object Detection Gesture Recognition Object Counting Text Recognition	DNN CNN SVM Decision Tree KNN Linear	Visual Wake Words CIFAR10 MNIST ImageNet DVS128 Gesture
Physiological / Behavioral Metrics	Segmentation Anomaly Detection Forecasting Activity Detection	DNN Decision Tree SVM Linear	Physionet HAR DSA Opportunity
Industry Telemetry	Sensing Predictive Maintenance Motor Control	DNN Decision Tree SVM Linear Naive Bayes	UCI Air Quality UCI Gas UCI EMG NASA's PCoE

Challenges: Energy

An ideal tinyML benchmark would profile the energy efficiency of each system. Unfortunately, there are many challenges in fairly measuring energy usage:

- Maintaining the accuracy of energy measurement across the diverse range of processors, silicon technologies and memory architectures.
- Determining the scope of the measurement.
  - Memories (RAM, FLASH)? Peripherals? PLL?
  - Pre/Post processing? Interfaces ?
- Measuring energy consumption without significant work or alterations to the SUT.
- Preventing energy measurements from impacting the other metrics

Scope of tinyMLPerf vs. MLPerf Inference: Power Envelope



tinyML Systems consume drastically less power than traditional ML systems, yet still cover a large scope.

Challenges: Model Infancy

Despite the nascency of the field, tinyML systems are already diverse. This poses a number of challenges:

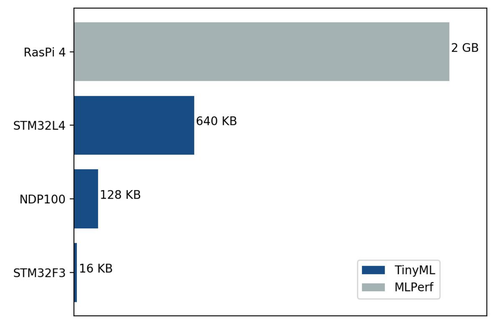
- Lack of standardization makes collecting metrics harder to formalize.
- Novel architectures have drastically different constraints and topologies.
- System requirements vary significantly across use cases.
- Performances are difficult to normalize.
- Different manufacturing technologies jeopardize comparisons with a fairly acceptable methodology.

Challenges: Memory

Memory constraints are one of the primary motivating factors for the creation of a tinyML specific benchmark. However, memory constraints add additional developmental challenges:

- Traditional benchmarks use NN models that are far too large in weights and activations.
- The overhead of the benchmark is more significant factor, pushing the need for non-intrusive inspection of key metric indicators.
- The System Under Test cannot hold the entire testing set, w/out involving host communication.
- Software (e.g. RTOS, drivers, built-in libraries) will require further discrimination.

Scope of tinyMLPerf vs. MLPerf Inference: Memory Envelope



Limited memory is a significant constraint for tinyML systems and the degree of which can vary widely.

Challenges: Processors Heterogeneity

tinyML is still a new field. It creates an opportunity to foster growth through community efforts (with industry support) but also poses a number of challenges for developing a robust benchmark that features industry acceptance and consensus:

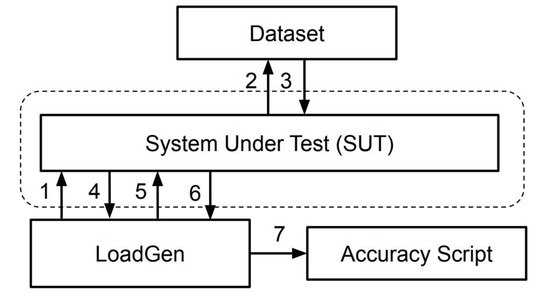
- Widely accepted tinyML neural network models.
- Large open source tinyML datasets.
- Frameworks are still evolving and few de-facto industry standards have become popular therefore model portability/interoperability is evolving:
  - e.g. tensorflow, keras, pytorch, mxnet, caffe etc, associated interoperable file formats (e.g. tflite, keras, onnx, nnef).

Use Cases Selected for v0.1

The criteria for the preliminary selection was to select three use cases that represented the scope of tinyML in terms of input type, size, neural network model type, and maturity. Model selection is still in progress.

Use Case	Dataset
Audio Wake Words	Speech Commands
Visual Wake Words	Google's VWW dataset
Anomaly Detection	Physionet, HAR, DSA, Opportunity

LoadGen and System Under Test



(workflow is subject to change)

Working Group Member Organizations

Join Us!

tinyMLPerf

Help us create a tinyML Benchmark!

Join https://groups.google.com/forum/#!members/mlperf-tiny