EEMBC MACHINE LEARNING BENCHMARKS: A PROGRESS REPORT
August 30, 2018

By Peter Torelli
President and CTO
EEMBC

Introduction

At the end of May 2018, EEMBC announced that the organization was seeking participants for a new working group tasked with developing EEMBC's Machine Learning Benchmark Suite, which will identify the performance potential and power efficiency of processor cores used for accelerating machine learning jobs on clients such as virtual assistants, smartphones, and IoT devices.

This progress report is intended to provide the EEMBC community and its followers with an overview of how the benchmark suite is being developed, including the motivation for this benchmark suite and the choices that have been made so far about how and what it will measure.

Why A Machine Learning Benchmark is Needed

The primary "audience" members for EEMBC's Machine Learning Benchmark Suite are the companies making virtual assistants, smartphones, and IoT devices. Doubtless they are already richly provisioned with their own proprietary benchmarks, so how would they benefit from a new machine learning benchmark from EEMBC?

For one thing, running proprietary benchmarks is not an uncomplicated task. It involves purchasing hardware, setting it up, and then running all the measurements. Perhaps for the Googles, Amazons, and Facebooks of the world this is an expected cost of doing business. But even for them, it's desirable to shift the cost of processor benchmarking back to the processor manufacturers—provided that the manufacturers are all using the same standard of

measurement. This standardization—a rigorous standardization—is what EEMBC benchmarks in general and its machine learning benchmarks in particular are intended to provide.

Besides the household names, there is an enormous ecosystem of developers, integrators, and smaller competitors that may or may not have access to proprietary benchmarks of their own or someone else. Everyone benefits if the players can agree through a consensual process what the standard metrics will be. As with previous EEMBC benchmarks, this process eventually leads to a database of benchmark scores that everyone in the ecosystem can use to help them with their research, or to compare their own benchmark results against, without having to go out and buy a hundred different pieces of hardware and run tests on them. This common baseline serves everyone by improving transparency and making it that much easier for OEMs to talk to processor vendors and vice versa.

The Choice of Machine Learning Workloads

In nearly any processor benchmark development, one of the early decisions that must be made is which workloads will be included in the benchmark suite. For its first-generation benchmarks, the Machine Learning Work Group is evaluating the use cases, neural net models, and training datasets shown in the table below.

Use case/category	Neural Net Models Under Consideration	Training dataset
Image classification	ResNet 50 MobileNet V2	ImageNet
Object detection	SSD (MobileNet) YOLO v3	MS Coco
Language translation	Google NMT	Stanford NMT
Speech recognition	Baidu DeepSpeech 2 Listen, Attend and Spell (LAS)	LibriSpeech

These choices are based on where we think the market is moving right now, and thus what will be most useful from the point of view of the benchmarks' ultimate "customers." This approach is somewhat different from what EEMBC did in its early years, when the consortium's principal

task was to develop benchmarks for general-purpose processors that could be deployed in any number of different applications. Twenty years later, the focus on innovation is in very application-specific processors being created for a very dynamic market. Therefore, we fully expect the workloads we specify for the first-generation Machine Learning benchmarks to have a relatively short shelf-life, and we plan to do periodic refreshes that will be driven entirely by what the market is demanding.

How the Benchmarks Will Work

Unlike the MLPerf benchmark, the EEMBC Machine Learning benchmarks focus on inference rather than training. The benchmark aims to measure raw inference performance as well as the time required to generate a neural-net model.

The benchmark suite will use behavioral models that will be delivered “pre-trained,” allowing users to adopt whatever frameworks or runtimes they choose in hopes of maximizing optimization. This means the numerical “weights” that define the strength of feedback paths will be pre-programmed by EEMBC.

A convolutional neural net would be used for object classification, while for natural language processing the benchmarks would use a recursive neural net or some other kind of architecture that has different feedback loops and different weights in it. The choices shown in the table above are tentative as of this writing.

Within this scenario there are multiple opportunities for measurement. For example, the inference dataset will yield information on how fast the DUT can go from getting a dataset, to responding with an answer, to what the throughput is when you batch it large numbers of images.

Let's say you have a trained neural net and you want to feed it 500 images. One metric is how long before it's ready to respond from when you present it the first image, to when it gives a response, to then how fast it can process the next 499 images. So there's a latency associated there and it's very important for time-critical systems.

Especially if you're a processor trying to make a decision in a car, you want to make the decision as quickly as possible and for your latency to be as small as possible. But then, once that latency's done, when you're in steady state driving mode, you want to be able to process these images as they're coming in, in real time.

The inference data set and the training data sets will be the same for all DUTs, but the benchmarks will measure the throughput, the latency, and the accuracy of the object classification performed by the DUT, which will always be expressed as a classification of probabilities—such as a 99.9% certainty that the object being classified is a car.

Open Questions

Part of the rationale for this progress report is to recruit participants for the EEMBC Machine Learning Working Group. Although the development of the new benchmark suite is well underway, new working group members are still welcome, and they will have the opportunity to influence decisions on the following open issues:

Score reporting – Up for discussion is whether the benchmark test results can or should be reported as a single-number composite score, which tends to be more easily communicated, or whether results will need to be expressed as a variety of scores

Model format – It isn’t yet clear if providing a single Caffe model or multiple models such as PyTorch, TensorFlow and MXnet would be desirable: the choice impacts the balance of portability versus development overhead and cumulative upkeep

Model Optimization – What kind of optimizations should be allowed, such as quantization, pruning and fusion

Test dataset – In addition to negotiating consensus the test datasets’ content, we also need to consider licensing and copyright; whether the test sets come from the internet or are donated by EEMBC members, what are the distribution rights surrounding them?

Accuracy validation set – Do we need additional test datasets that exhibit characteristics that lend themselves to validation?

Target accuracy – What are the units and how much deviation is allowed between inferences between different architectures, and how is this deviation accounted for in a fair and equitable manner?

The Machine Learning Working Group will be meeting throughout the remainder of 2018 to resolve these issues. EEMBC welcomes your participation. To find out more, please E-mail me at peter.torelli@eembc.org.