tinyML Perf: Expanding the MLPerf Inference benchmark to microcontrollers and tiny devices

Colby Banbury, Max Lam, and Vijay Janapa Reddi
Harvard University, Cambridge, MA

Problem Statement
ML inference on the tiny devices is an attractive prospect due to its potential for increasing energy efficiency, privacy, responsiveness, and autonomy of edge applications, however, applications are limited by the tiny power envelope of these platforms. To overcome this constraint, significant strides have been made in developing novel architectures that are optimized for inference. These optimizations have decreased the relevance of traditional, general-purpose hardware benchmarks which can no longer accurately reflect ML performance. The benchmarking suites that do target edge inference, like MLPerf and MLMark, lack support for devices smaller than a mobile processor. Without a reliable benchmark, tinyML system performance is difficult to evaluate and compare, which limits the information available for product design decisions.

Approach
To address this need, we have created a working group within the MLPerf organization with the goal of extending the existing inference benchmark suite to support the evaluation of tinyML systems. By utilizing collaboration between industry and academia, and embracing an iterative approach, we ensure the relevance of the resulting benchmark suite. Once complete, the benchmarks can be integrated into MLPerf's existing inference benchmark suite which will have the additional benefit of increasing the visibility and recognition of tinyML.

Results
In under a month, the working group has grown to 66 members representing 21 companies and 5 universities. This breadth of perspective will ensure the development of a fair benchmark and increase the probability of widespread adoption.
Our first task was to compile a list tinyML specific use cases, from which we have selected three to target for our preliminary set of benchmarks: audio wake words, visual wake words, and anomaly detection. We believe these use cases are sufficiently representative of the space to comprise the working version of the tinyMLPerf benchmark suite.

Future Work
The working group is still very early in the process of developing a tinyML benchmark. Significant work remains to be done in developing the individual benchmarks and the overall framework, including tackling difficult challenges, such as reliable power measurement. To ensure continued progress and eventual adoption, our primary goal is increasing participation and engagement. The tinyMLSummit is the ideal venue to reach the relevant community and foster additional collaborations.