Benchmarking AI compiler for the TinyML market

Peter Chang
TinyML Asia 2023, 2023/11/16
Outline

1. Introduction to the MLPerf Tiny benchmark
2. Optimization of AI models from compiler’s perspectives
3. Possibility of the future benchmark designs for TinyML market
4. Epilogue
1. Introduction to the MLPerf Tiny benchmark
What is **MLPerf** benchmark?

The foundation for **MLCommons®** started with the **MLPerf™** benchmarks in 2018, which established industry-standard metrics to measure **machine learning** performance and...

https://mlcommons.org/en/history/
What is **MLPerf Tiny** benchmark?

- MLPerf Tiny serves as a Machine Learning Inference benchmark collection tailored for TinyML systems.
- MLPerf enables the assessment of energy consumption and inference speed for AI models focused on visual and audio tasks.
- All submitters must fit the quality targets for each use case for close division.

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Use Case</th>
<th>Model</th>
<th>Quality Target</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Anomaly Detection</td>
<td>Deep AutoEncoder</td>
<td>0.85 (AUC)</td>
<td>270 kPar</td>
</tr>
<tr>
<td>KWS</td>
<td>Keyword Spotting</td>
<td>DS-CNN</td>
<td>90% (Top 1)</td>
<td>52 kPar</td>
</tr>
<tr>
<td>IC</td>
<td>Image Classification</td>
<td>ResNet</td>
<td>85% (Top 1)</td>
<td>96 kPar</td>
</tr>
<tr>
<td>VWW</td>
<td>Visual Wake Words</td>
<td>MobileNet</td>
<td>80% (Top 1)</td>
<td>325 kPar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Typical Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Memory</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td>AI Model Size</td>
</tr>
</tbody>
</table>

https://www.youtube.com/watch?v=i4wCqoVcdUJ
Submission Requirements

Notice:
Energy Number is optional.

System Category:
1. Available System
2. Preview Systems
3. Research, Development, or Internal (RDI)

Available System Category comprise solely of components that can be obtained for purchase or leased from cloud services.
Submission Process

**Submission**
- Sign CLA
- Provide POCs with Github handles and email addresses

**Review**
- All submitters are peer-reviewers
- Reviewers fill objection opinions
- Peer review objections
- Submitters revise based on objections
- Vote for accept or not

**Publication**
- Write Supplemental materials to describe your work
- Join press conference meeting before publish
- Release the results on the MLCommons website
2. Optimization of AI models from compiler’s perspectives
**ONNC & Our Submissions**

- ONNC is an AI compilation suite developed by Skymizer for various markets, from cloud to tiny devices.

- For tiny devices, ONNC compiles AI models to C codes which call Neural Network Library for the target board.

- We use ONNC to compile AI models to C codes for Nuvoton NuMaker M467HJ Cortex-M4 board and the benchmark’s reference board.
MLPerf Submissions on MLPerf v1.1

Skymizer submit two numbers, one for Skymizer and another as the agent for Nuvoton which is Skymizer’s collaboration partner.

MLPerf Inference – Tiny

“The outstanding results achieved in the MLPerf Tiny Benchmark’s Cortex-M4 MCU segment highlight Nuvoton’s and Skymizer’s dedication to pushing the boundaries of machine learning performance in resource-constrained environments.”


ONNC Optimization in Latency and Energy

Reduce Latency
Our latency is 35% less in the best-case scenario.

Reduce Energy
Our energy is 32% less in the best-case scenario.

Software Stack Comparison between ONNC and TVM

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Quality Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>Deep AutoEncoder 0.85 (AUC)</td>
</tr>
<tr>
<td>IC</td>
<td>ResNet 85% (Top 1)</td>
</tr>
<tr>
<td>KWS</td>
<td>DS-CNN 90% (Top 1)</td>
</tr>
<tr>
<td>VWW</td>
<td>MobileNet 80% (Top 1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Os</th>
<th>ONNC-NUCLEO_14-R8ZI-Zephyr</th>
<th>TVM-NUCLEO_14-R8ZI-Zephyr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (ms/inf)</td>
<td>AD</td>
<td>IC</td>
</tr>
<tr>
<td>ONNC</td>
<td>8</td>
<td>296.6</td>
</tr>
<tr>
<td>TVM</td>
<td>8.6</td>
<td>389.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Os</th>
<th>ONNC-NUCLEO_14-R8ZI-Zephyr</th>
<th>TVM-NUCLEO_14-R8ZI-Zephyr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy (uJ/inf)</td>
<td>AD</td>
<td>IC</td>
</tr>
<tr>
<td>ONNC</td>
<td>409.666</td>
<td>14927.33</td>
</tr>
<tr>
<td>TVM</td>
<td>443.2</td>
<td>20236.3</td>
</tr>
</tbody>
</table>

MLPerf Benchmark
TensorFlow Lite
ONNC VS TVM
Arm CMSIS-NN
ST Nucleo M4 L4R5ZI

ONNC-NUCLEO_L4R5ZI-Zephyr
TVM-NUCLEO_L4R5ZI-Zephyr
Key components in an AI model compiler

Quantization
- Convert higher-bit computation into lower-bit computation.
- Hardware-friendly.
- Speed up.

Graph-Level Optimization
- Hardware-friendly.
- Speed up.

Operator-Level Optimization
- Speed up.
- Hardware instructions.
3. Possibility of the future benchmark designs for TinyML market
More Energy-Centric

For tiny device developers, energy consumption usually will be the first key factor to decide whether they should adapt AI or not.

Designing a more energy-centric/energy-priority benchmark would fit developers’ needs more.

More Whole-System’s View

The whole system benchmark can show the performance and energy numbers not only from AI inferencing but also from pre-/post-processing and OS.

The whole system performance and energy analysis will also facilitate their decisions for those who try to decide which evaluation board to buy.

Figure 5: Energy and power consumption on Pixel 4.

More Comprehensive

AI benchmarks for tiny devices would be better if it can cover not only audio and visual but also other more sensing modalities, like environmental sensing modalities.

Also, for data sets, an open and comprehensive data set suite would be more realistic, more comprehensive and be more non-discrimination.
4. Epilogue
ONNC RISC-V Support & Booth
- ONNC also support RISC-V as MCU.
- We have a demo using Tinker V board with Andes IP in the booth area.

MLPerf Tiny Next Round Submission
- Submission: February 23th, 2024 (Planning)
- Publication: March 27th, 2024 (Planning)
- Would add 2 streaming benchmarks (Planning)
  - Streaming Denosing LSTM & Streaming KWS

LLM on tiny devices
- Model compression with accuracy ensurance will be the key in landing LLM or Transformer-based models to tiny devices.
- Skymizer also have set this agenda within our roadmap.
Copyright Notice

This presentation in this publication was presented as a tinyML® Asia Technical Forum. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org