Supplemental Discussion

MLPerf™ Tiny v1.0 Results Discussion

The submitting organizations provided the following descriptions as a supplement to help the public understand the submissions and results. The statements **do not reflect the opinions or views of the MLCommons® Association.**

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

GreenWaves Technologies

GreenWaves designs highly efficient and easy to program processors that interpret and transform rich data sources using AI and digital signal processing in highly energy constrained devices. Our GAP9 processor enables new-to-world features in hearable and IoT products.

GAP9 delivers extraordinarily low energy consumption on medium complexity neural networks such as the MobileNet series in both classification and detection tasks but also on complex, mixed precision recurrent neural networks such as our LSTM based audio denoiser.

The verified MLPerf™ Tiny benchmark results now show that although GAP9 is dimensioned to deal with much larger problems, it also excels when running very small neural networks (NNs), proving its ability to address the wide range of embedded NN processing tasks required by real life applications.

GAP9 has extremely low energy consumption for all four MLPerf™ Tiny benchmarks. These results are achieved thanks to GAP9's uniquely scalable architecture and to GreenWaves' state-of-the-art neural network toolchain, GAP*flow*, which transforms networks from ONNX and TensorFlow Lite formats into optimised, readable C source code.

GAP9 also has other leading edge computing capabilities that can be combined with Neural Networks. For example, an NN can be used to continuously update the parameters of GAP9's unique, sample by sample, highly configurable Smart Filtering Unit (SFU) enabling new NN steered Adaptive Active Noise Cancellation or Adaptive Audio Transparency features.

GAP9 provides an extremely powerful signal processing and AI computing platform for the next generation of hearable and IoT products.

For more information go to www.greenwaves-technologies.com.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

HLS4ML

The goal of the hls4ml open-source workflow is to enable researchers and engineers to codesign optimized neural networks for efficient dataflow architectures on a multitude of accelerator hardware platforms. It is developed and supported by the international Fast Machine Learning for Science research community. Originally developed for the Large Hadron Collider to do ultrafast sub-microsecond inference, hls4ml aims to serve the wider ML community to accelerate both the design process and neural network implementation across a broad range from low-power to high-performance devices.

With hls4ml, developers can rapidly build customized efficient neural networks for (SoC) FPGAs and ASICs using state-of-the-art quantization and pruning techniques. The workflow enables model optimization in open-source ML frameworks like TensorFlow, PyTorch, and Keras and translates models into deployable firmware in a single Python-based flow.

The hls4ml team, consisting of community members from Fermilab, Columbia University, UC San Diego, University of Washington, and CERN, have demonstrated the hls4ml workflow for the MLPerf Tiny benchmarks on the off-the-shelf TUL PYNQ-Z2 and Arty100-T boards which supports the Xilinx FPGAs. In the MLPerf Tiny rounds including 0.5 and 0.7, many developments, optimizations, and improvements have come about in hls4ml's open source workflow. In the MLPerf Tiny 1.0 Round, we've further improved upon our previous round's results, providing a greater than 3.5x latency improvement through dataflow optimization for the image classification benchmark.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

OctoML

OctoML (https://octoml.ai/) accelerates AI innovation by making it easy to harness the power of machine learning to build intelligent applications. Using OctoML, customers can automatically deploy ML models to a broad set of devices without specialized ML skills, using established DevOps workflows and tools. OctoML was founded by the creators of Apache TVM, an open source stack for ML portability and performance and a key part of the architecture of popular consumer devices like Amazon Alexa. Founded in 2019, OctoML is based in Seattle, WA, and backed by Amplify Partners, Madrona Ventures, Addition Capital, and Tiger Global.

OctoML's MLPerf™ Tiny benchmarks results show the inference performance on ARM Cortex-M4 and Cortex-M33. OctoML's submission uses two compilation flows. First, it uses the bring your own codegen (BYOC) path where after the model was imported to TVM and represented in Relay, TVM parses it into sub-graphs and matches them with the CMSIS-NN APIs. Second pipeline uses native microTVM schedules and AutoTuning approach. In autotuning, tunable sub-graphs would be extracted as tuning tasks. Tuning each task requires a schedule template with tunable parameters. Autotuning would use a set of values for each of the parameters, build the subgraph and examine the runtime on the device to optimize the runtime.

microTVM provides an easy way to integrate the model compilation into any embedded environment by providing a standard python API. This API allows you to create a project, build it using your embedded compiler/SDK and flash the built firmware on the device in only a few steps. In addition, microTVM uses ahead of time compilation to generate your model in C code and uses static memory allocation.

OctoML's MLPerf™ Tiny benchmarks shows that using AutoTuning with native microTVM schedules we can achieve similar performance to CMSIS-NN libraries which are hand-tuned libraries.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

Plumerai

Plumerai makes deep learning tiny and radically more efficient to enable inference on small, cheap and low-power hardware. Plumerai's customers use its inference engine to accelerate any neural network on off-the-shelf microcontrollers. Plumerai again presents outstanding MLPerf™ Tiny Inference results for Arm Cortex-M CPUs.

Plumerai offers turnkey solutions for camera-based people detection for low power and cost efficient systems. Plumerai's customers integrate it into smart home, doorbell, security, and video conferencing cameras. Tiny sensing systems use it for smart lighting, occupancy detection, air conditioning, smart retail, smart city, robotics, and more.

Plumerai's inference engine runs any 8-bit neural network model in the smallest memory and power footprint. Plumerai's MLPerf™ Tiny Inference benchmark results show the four neural networks running on tiny off-the-shelf Arm Cortex-M33, M4, and M7 microcontrollers from STMicroelectronics and Infineon. Plumerai's inference engine supports any Arm Cortex-M microcontroller, Arm Cortex-A-based chips, x86, RISC-V, ARC EM, and accelerator architectures. The inference engine typically halves the memory footprint and increases inference speed by 70%, without affecting the accuracy of the AI.

Compared to the previous MLPerf™ Tiny 0.7 results, Plumerai further improved inference speed and reduced code size and memory usage. The latest version of Plumerai's inference engine also adds state-of-the-art support for time-series neural networks, such as LSTM-based RNNs. These are common in motion sensors, health sensors, speech, and audio applications.

Plumerai's inference engine unlocks new applications such as computer vision or IMU-based human activity recognition to run on tiny and battery-powered devices. Plumerai's inference engine enables AI tasks that typically run on much larger systems to now run on small edge devices. Plumerai is very proud of the MLPerf™ Tiny benchmark results that function as an independent endorsement of the highest efficiency that the inference engine reaches. You can try it out at plumerai.com/benchmark.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

Qualcomm

Qualcomm MLPerf™ Tiny v1.0 results have come to show the AI inference capabilities of our Qualcomm Sensing Hub.

At the heart of the Snapdragon mobile platform lies our Qualcomm AI Engine with the powerful Hexagon processor for dedicated AI processing. Nevertheless, the need for low power AI capabilities is increasing exponentially through the years. Having that in mind, in this round of submissions we chose to focus on another significant core inside our Qualcomm AI Engine that is responsible for ultra-low power AI processing, the Qualcomm Sensing Hub.

The Qualcomm Sensing Hub is a multi-core architecture that consists of a dedicated AI processor, a DSP, plus its own memory to process contextual data streams including voice, audio, sensors, and connectivity allowing for ultra-lower power AI processing.

With the newest generation Snapdragon platform we are pushing the boundaries of how much horsepower the Sensing Hub can bring at ultra-low power, reaching to 0.1ms in anomaly detection.

This is our first ever submission in this category and we will continue to drive innovation in the ultra-low power domain. Please keep an eye out for the announcement coming from our annual Snapdragon Summit: Snapdragon Summit 2022 | Snapdragon Tech Event (qualcomm.com)

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

Silicon Labs

Earlier this year Silicon Labs (NASDAQ: SLAB), a leader in secure, intelligent wireless technology for a more connected world, announced an end-to-end machine learning (ML) software development solution for its existing Series 1 and Series 2 wireless SoCs and also introduced the EFR32BG24 and EFR32MG24 Bluetooth and Multiprotocol SoCs, which feature integrated AI/ML hardware acceleration providing up to 8x faster processing and up to 6x lower power consumption for ML processing vs. processing only on an ARM Cortex M33 core. The Silicon Labs GSDK development environment supports TensorFlow Lite Micro with standard CMSIS-NN as well as HW accelerated kernels. Using Silicon Labs' hardware platform and ML development solution, designers can enhance embedded applications with AI/ML capabilities even in ultra-low-power wireless IoT devices.

Silicon Labs is publishing MLPerf™ Tiny v1.0 benchmark results on its new EFR32MG24 platform, showing an improvement – 1.5 to 2x speed increase and 40-80% less energy on specific models – against the previously-published MLPerf™ Tiny v0.7 results. The improved results highlight the benefits of the on-chip acceleration that efficiently serves the growing needs of AI/ML-enhanced low-power wireless IoT solutions, allowing for devices to stay in the field for up to ten years on a single coin-cell battery. During inferencing calculations, the main CPU is offloaded and can execute other tasks if needed or can be in sleep mode to achieve even more power savings. Developers can get started on building AI/ML solutions for the IoT today by visiting https://www.silabs.com/applications/ai-ml.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

STMicroelectronics

Introduced at CES 2019, STM32Cube.AI has become an industry reference for building, validating, and deploying tiny Machine Learning and neural network models on STM32 microcontrollers. More than 60,000 customers using STM32 microcontrollers can benefit from this advanced technology at no charge. By collaborating closely with a wide spectrum of our customers we have managed to evolve our offering to understand their challenges and provide solutions that improve the way they effectively develop and deploy neural networks. With a continuous improvement in the number of layers and topologies supported, we are now measuring the results of this fruitful journey.

Since the previous MLPerf® Tiny Inference benchmark (v0.7), we improved our performance on inference time by up to 33% and on energy used by up to 37%. In addition, we observed at Customers using the latest version of the STM32Cube.AI on other Neural Network models improvements up to 75% on inference time. Key improvements driving these achievements include adding more flexibility on the parametrization of the Neural Network optimizations. Users can now optimize for RAM for applications that are memory footprint sensitive, optimize for time for applications that are inference time sensitive or choose the balanced optimization to get the best compromise of both.

The value delivered by embedding STM32's edge AI technology on edge devices is already visible in multiple applications such as wearables, industrial IoT, smart home objects, and more. The results presented to the MLPerf™ community in the closed category were obtained using standard STM32 software settings and hardware configuration. They can be reproduced very easily by anyone by visiting our site https://stm32ai.st.com/stm32-cube-ai/ and download our tools there.

Please follow the MLPerf Results Messaging Guidelines and MLCommons Trademark Usage Guidelines.

Syntiant

Syntiant is the recognized leader in delivering end-to-end deep learning solutions for always-on voice, audio, vision and sensing applications across a wide range of consumer and industrial use cases, from earbuds to automobiles. The company designed its technology as a complete turnkey system by combining purpose-built silicon with an edge-optimized data platform and training pipeline, bringing ultra-low power, cloud-free advanced machine learning solutions to the edge.

Introduced in 2021, its award-winning NDP120 is a powerful, highly efficient neural processor packaged in an extremely cost-effective, compact 3.1 mm x 2.5 mm WLBGA package, manufactured in widely available 40nm CMOS-based technology. The NDP120 enables the deployment of advanced ML solutions, such as voice and speech interfaces in severely energy- and size-constrained applications, while also supporting far-field audio.

The results of the MLPerf® Tiny v1.0 benchmarks demonstrate the NDP120's compelling throughput and energy performance. Its Syntiant Core 2™ neural network accelerator’s versatility is demonstrated by executing the benchmark’s visual wake word and image classification workloads. Just as in the previous version of the benchmark for keyword spotting, the Syntiant Core 2 again achieved latency and energy consumption orders of magnitude lower than other devices for the vision oriented tasks. The benchmarks also demonstrate the easy migration path from pre-trained models using Syntiant’s Training Development Kit (TDK).

The Syntiant Core 2 neural engine supports all commonly used layers while the included DSP provides flexible pre-processing and conventional signal processing. All of the benchmarks used less than one third of the on-chip resources, making the NDP120 ideal for sensor fusion or for tasks that require multiple independent networks to run concurrently. The NDP120 has been adopted to provide powerful neural processing in edge devices serving the security, healthcare, consumer and automotive industries.

More information on the company can be found by visiting www.syntiant.com.