

If you cite this paper, please use the PEMWN 2023 reference: Z. Huang, K. Zandberg, K. Schleiser, E. Baccelli. U-TOE: Universal TinyML On-board Evaluation Toolkit for Low-Power IoT. In Proc. of 12th IFIP/IEEE PEMWN, Sept. 2023.

U-TOE: Universal TinyML On-board Evaluation Toolkit for Low-Power IoT

Zhaolan Huang, Koen Zandberg, Kaspar Schleiser and Emmanuel Baccelli

arXiv:2306.14574v2 [cs.LG] 22 Aug 2023

Abstract—Results from the TinyML community demonstrate that, it is possible to execute machine learning models directly on the terminals themselves, even if these are small microcontroller-based devices. However, to date, practitioners in the domain lack convenient all-in-one toolkits to help them evaluate the feasibility of executing arbitrary models on arbitrary low-power IoT hardware. To this effect, we present in this paper U-TOE, a universal toolkit we designed to facilitate the task of IoT designers and researchers, by combining functionalities from a low-power embedded OS, a generic model transpiler and compiler, an integrated performance measurement module, and an open-access remote IoT testbed. We provide an open source implementation of U-TOE and we demonstrate its use to experimentally evaluate the performance of various models, on a wide variety of low-power IoT boards, based on popular microcontroller architectures. U-TOE allows easily reproducible and customizable comparative evaluation experiments on a wide variety of IoT hardware all-at-once. The availability of a toolkit such as U-TOE is desirable to accelerate research combining Artificial Intelligence and IoT towards fully exploiting the potential of edge computing.

I. INTRODUCTION

As Artificial Intelligence (AI) permeates our lives more and more, mechanisms such as deep neural networks [16] are put to use (or their deployment planned) in more and more places in various distributed systems. In particular, wireless sensor network (WSN) can improve its coverage and connectivity, reduce energy and bandwidth usage by deploying AI onto edge nodes [14].

The data pipeline with AI typically requires the creation and the use of a *model*, i.e. a layered structure of complex algorithms (also known as *operators*) which interpret data and make decisions based on that data. This model must first be trained (*learning* phase [16]), before it can be put in production (used for *inference*).

Recent work from the TinyML [13, 11] community forays into optimizing models to fit tinier resource budgets (and to perform efficiently nevertheless) on low-power microcontrollers in the Internet of Things (IoT). As a consequence, both learning and inference placement possibilities are extended to encompass ultra low-power terminals.

However, generic and convenient open source tools lack for designers tackling a combination of AI and IoT (AIoT), who are required to:

- evaluate the performance of their models when placed somewhere along the terminal-edge-cloud continuum, especially when including potential placement on different microcontroller-based devices;
- fine-tune their models, and identifying performance bottlenecks at model layer granularity, on different microcontrollers;
- select an adequate microcontroller to execute their model, for a targeted task running on a low-power device to-be-designed.

This paper thus introduces U-TOE, an open source AIoT toolkit which tightly combines a generic model compiler and a popular low-power IoT operating system to automatically compress, flash and evaluate arbitrary models (output of TensorFlow, PyTorch...) on arbitrary commercial off-the-shelf low-power boards (such as BBC:microbit, nrf52840dk, Arduino Zero, HiFive...) based on the popular microcontroller architectures (ARM Cortex-M, RISC-V, ESP32).

Paper contributions. Our contributions are as follows:

- We design a universal toolkit for TinyML on-board evaluation (U-TOE). It provides feasibility checks for the deployment of arbitrary model on a given IoT hardware platform. It allows researchers and developers to locate the performance bottleneck of a given model on a target device. The evaluation results enable co-design with other components at system level, help optimizing ML models and configurations for specific use cases, allowing to achieve the best possible performance on target devices.
- We released the code¹ of U-TOE under an open source licence. This implementation enables compilation, flashing, and evaluation of neural network (computational graph based) models from mainstream ML framework onto various low-power boards based on popular ISA.
- We provide benchmarks and a comparative experimental evaluation using U-TOE, reproducible both on an open-access IoT testbed and on personal workstations, which provide insights on inference performance with different models on different low-power hardware and demonstrate how U-TOE can be re-used by TinyML experimental researchers and developers to fine-tune IoT configurations.

Corresponding Author: Z. Huang. E-Mail: zhaolan.huang@fu-berlin.de

Z. Huang, K. Zandberg, and K. Schleiser are affiliated with Freie Universität Berlin. E. Baccelli is affiliated with Inria, France.

¹see <https://github.com/zhaolanhuang/U-TOE/>

II. RELATED WORK

Recent work has surveyed [12] the scope of ML frameworks, tasks, metrics, including a comprehensive review on TinyML stack and deployment pipeline. A number of challenges need to be met in order to fit the tiny resource budgets typical of microcontrollers (kiloBytes of memory, power consumption in mWatt, CPU frequency in MHz) while maintaining performance at an acceptable level and retaining portability to extremely polymorphic hardware in this category.

Embedded IoT Software Platforms – Various open source IoT operating systems are used to provide hardware abstraction, and resource sharing primitives as well as convenient peripheral access (e.g. sensor/actuator, network subsystem) on heterogeneous low-power microcontrollers. Prior work such as [7] surveys such operating systems, among which prominent examples include RIOT[2] and Zephyr[18]. However, such software platforms so far offer very limited support for ML frameworks – if at all. Moreover, the most advanced support so far are typically hardware- or vendor-specific e.g. with libraries provided by STM32CubeMx or ARM CMSIS-NN.

Low-power IoT Testbeds – Various testbeds offer remote access to fleets of reprogrammable microcontroller-based devices. Prior work such as [8, 6] survey such testbeds, among which prominent examples include the open access facility IoT-lab [1], which offers remote bare-metal access (serial over TCP) to a fleet composed of hundreds of popular low-power boards of various kind.

Benchmarking Suites for TinyML – Benchmarking ML on low-power hardware entails a number of challenges [4]. Prior work such as MLPerf Tiny [3] provides a standard benchmark suite (a fixed set of representative ML tasks) for evaluating the performance of given hardware, and an online platform for manufacturers to publish their comparative benchmark results. In contrast, U-TOE offers a more powerful and more customizable toolkit for performing feasibility’s check of user-defined machines learning models on low-power devices, with a greater degree of flexibility and customization.

TinyML Benchmarks – Prior work such as [9] focuses on performance comparison of different machine learning frameworks on two COTS low-power boards (Arduino Nano BLE 33 and STM32 NUCLEO-F401RE). In particular, it benchmarked two TinyML frameworks, TFLM and X-CUBE-AI over gesture recognition and wake word spotting. Other work such as [15] tested TFLM models on several microcontroller-based boards. While such papers provide a performance comparison of specific frameworks on specific boards for specific tasks, U-TOE offers greater flexibility and generality, allowing developers to evaluate a wider range of (user-specified) models on a larger variety of low-power devices, and to dive into the execution details of ML models.

TinyML Model Transpiler & Compilers – Compilers such as TVM [5] can be used to automate the transpilation and compilation of models provided by major ML frameworks (TFML, Pytorch etc.) so as to expose low-level routines and optimize them for execution on specific processing unit characteristics (CPU, GPU etc.). An extension of TVM called uTVM was

recently introduced, adding smaller hardware targets including a variety of MCUs (microcontroller units).

TinyML Model Profilers – ML-EXray[10] enables TinyML developers to gain visibility into the layer-level details of ML execution and diagnose cloud-to-edge deployment issues. Developers can analyze and debug edge deployment pipelines with high usability, using less than 15 lines of code for fully examination. However, the reliance on Tensorflow Lite restricts the capability to accommodate models from further ML frameworks and deploy on low-power devices. Major ML frameworks (TFML, Pytorch and MXNet etc.) provide internal profiler [17]. Such tools allows developers to measure the performance of their models. They can be used to collect metrics such as inference time and memory usage, which can then be analyzed to optimize the model’s performance. Though it can provide us execution details in layer level, it still lacks the supports for on-device deployment and evaluation on various IoT devices, while U-TOE is a more general-purpose toolkit that provides a comprehensive solution on a wide range of low-power devices. The above is summarized in Table I.

TABLE I
COMPARISON OF AIOT FRAMEWORKS AND TOOLKITS.

Framework	MCU	Model Type	Remote Eval.	Granularity	Model to Board Sol.
MLPerf	Yes	Specified	No	Model	No
ML-EXray	No	TFLite	No	Layer	No
TFLite	Yes	TFLite	No	Layer	No
Pytorch	No	Torch	No	Layer	No
uTVM	Yes	Universal	No	Operator	No
U-TOE	Yes	Universal	Yes	Operator	Yes

Eval. – Evaluation, Sol. – Solution

III. BACKGROUND ON TINYML PERFORMANCE ANALYSIS

One the one hand, as the most immediate limiting resource budget on microcontrollers concern memory limitations, typically in the order of kiloBytes, TinyML performance evaluation typically focuses primarily on metrics measuring memory consumption – while keeping an eye on execution speed – as described below. On the other hand, TinyML performance analysis can be tackled at different granularity levels: at the global model level, or at the operator level, for finer granularity, as described in the following.

A. Performance Metrics

The considered metrics offer insights into the feasibility, efficiency and resource utilization of offloading model inference burden to low-power devices. By analyzing these metrics, users can make initial decisions regarding model selection, optimization techniques, and hardware configurations to maximize performance and minimize the resource footprint on low-power devices.

Memory (RAM) Consumption – This metric measures the amount of dynamic memory space (primary RAM) consumed by the model during inference. It reflects the memory footprint of the model activation and is important for low-power devices

that have limited memory resources. Efficient memory utilization allows for the deployment of larger and more complex models on such devices.

Storage (Flash memory) Consumption – This metric quantifies the amount of storage space, typically in terms of Flash memory region, required to store the compute instruction and associated parameters. It reflects the model’s storage footprint on the low-power device. Minimizing storage consumption allows for accommodating multiple models on the device or orchestrating with other essential applications.

Computational Latency – This metric measures the time consumption of performing inference for each input sample, either at the model level or at the level of individual operators within the model. It reflects the inference speed of the model on the low-power device and plays crucial role in real-time or latency-sensitive applications. Core clock frequency, cache strategies and communication latency between memory and working core have great impact on this indicator.

SoC Price – This metric considers the cost of the System-on-a-Chip (SoC) used in the low-power device. The price of the SoC affects the overall affordability and feasibility of deploying model in large-scale distributed system. Lower-cost SoCs can make the deployment more accessible and cost-effective.

B. Measurement Granularity

As for performance analysis of machine learning in other domains, TinyML performance can be measured at different granularity levels:

Per-Model Evaluation – At this coarse level, one measures performance of the model as a whole, i.e. the resource footprint incurred by the execution of the model including all its layers and operators. For example, this allows for evaluating the resource consumption for inference with the production-ready code, on a particular industrial hardware setup.

Per-Operator Evaluation – At this level, one measures separately the performance of one or more operators (i.e. one or more components of the model). This per-operator measurement can help identifying specific operators that contribute to performance or inefficiencies, in optimizing the model’s efficiency and spotting potential bottlenecks.

IV. U-TOE TOOLKIT DESIGN & IMPLEMENTATION

A. Toolkit Architectural Design

U-TOE integrates uTVM and RIOT to perform model compilation, flashing, and evaluation of arbitrary models from mainstream ML frameworks onto various low-power boards. The toolkit is composed of the following key components:

- Model Compiler. U-TOE leverages the uTVM compiler to convert arbitrary neural network models into efficient C code. This compiler enhances the efficiency of the models and enables them to be run on low-power devices.
- RPC Mechanism. To evaluate the resource consumption at operator level, U-TOE utilizes the Remote Procedure Call (RPC) mechanism of uTVM. The RPC mechanism enables to upload and launch functions on to IoT boards over serial. This is useful for remote testing and profiling,

enabling U-TOE to wrap the model operators for measuring the computational latency and memory usage. It is composed of a client on the host and a server on the target device, receiving commands and executable instructions from the host.

- OS Environment. RIOT was chosen to provide a lightweight runtime environment for model execution and evaluation on microcontrollers, with its advantages in extensibility and wide-spectrum support for low-power boards.
- Evaluation Module. It contains two units: measurement worker and analyser. As shown in Fig. 3, the measurement worker is deployed on the MCU for acquiring performance metrics at model or operator level. Besides carrying out the measurement of resource footprint, It is in charge of the randomization of model input, and responsible to report metrics data to host device. The analyser runs on the host, statisticizes the uploaded metrics from device and provides human-readable frontend for users.

Additionally, U-TOE provides a *connector* for cloud-based IoT testbed which enables seamless interaction with remote boards using serial over TCP. As depicted in Fig. 1, users can evaluate and benchmark their models on local device, or on a remote, scalable testbed with wide-range of MCUs via U-TOE connector.

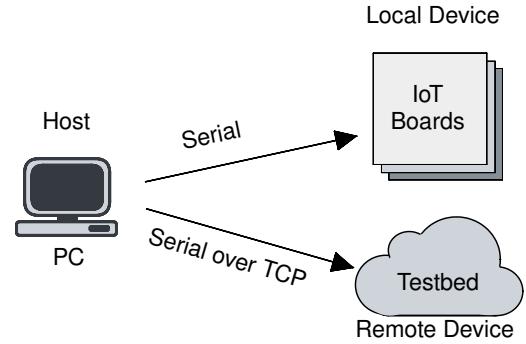


Fig. 1. Hardware setup of U-TOE. Users can connect local boards with host PC via serial, or use remote board service on IoT-Testbed.

Compilation and Deployment – U-TOE first gathers the specification of target device to decide the compilation options for uTVM and RIOT. Then, as presented in Fig. 2, uTVM generates sub-optimal model library of the input model without scheduling². Some static optimization strategies are applied in this stage according to target device type. This uTVM-generated library is jointly compiled with RIOT, RPC server and measurement worker into executable firmware, which is then automatically flashed on the device.

Evaluation – After deploying the program on the target device, a bidirectional channel – in Per-Operator evaluation –, or a unidirectional channel – in Per-Model evaluation –

²A schedule specifies low-level optimization for loop execution, enhancing cache hit and memory access. The optimal one is co-determined by model and device specification, and identified by heuristic search algorithms based on measurements on device. [5]

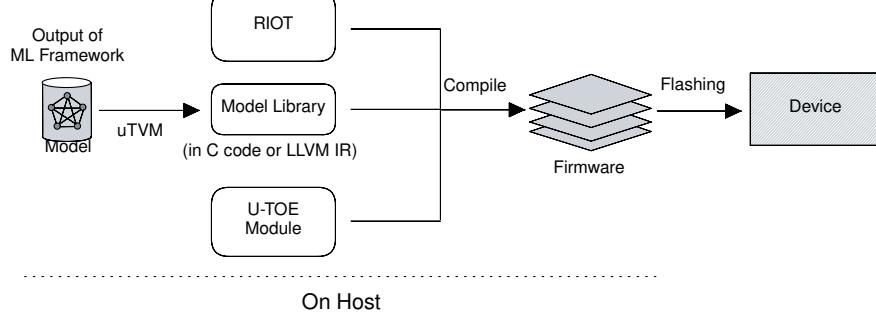


Fig. 2. Compilation and deployment workflow of U-TOE. uTVM optimizes and translates model from mainstream ML framework into model library, which is co-compiled and flashed with RIOT and U-TOE components onto target boards.

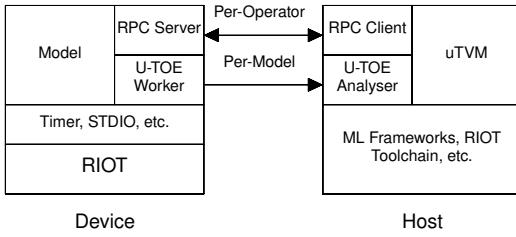


Fig. 3. Software architecture and components of U-TOE.

is set up between host and device, as depicted in Fig. 3. The measurement worker starts collecting performance metrics at user-specified level and uploads metrics data to analyser. Eventually, users can obtain the overall statistics of model metrics, or catch the performance bottleneck with execution details of each operator. All the raw metrics data are saved in log file for further, user-customized analysis.

B. Measurement Procedure

We designed two measurement procedures to support evaluation at different granularity. The procedures run across multiple components of the toolkit and the most workloads are primary on the target board. Following steps describes the measurement routine after compilation of executable program. The steps marked with **bold number** are executed on the target board.

Per-Model Evaluation – This mode focuses on the model performance in actual production environment. Here is the corresponding measurement routine:

- 1) Calculate model memory and storage consumption based on ELF file. We disable dynamic memory allocation to enable static analysis of memory footprint.
- 2) Deploy executable program to local or remote IoT board.
- 3) Repeat model inference based on the user-specific number of trials with randomized input on uniform distribution.
- 4) Record computational latency of each trial.
- 5) Upload records to host device for further analysis and archive.

At the end of evaluation, results with statistics (e.g. 95% confidence interval, median, maximum and minimum) are

presented on the host device, including computational latency and consumption of memory and storage.

Per-Operator Evaluation – In contrast to per-model evaluation, this mode focuses on the efficiency and resource footprint of each operator, enabling to discover the performance bottleneck inside models. Thanks to the time evaluator inside the uTVM’s RPC mechanism, we can measure computational latency at operator level. The high abstraction of timer and serial in RIOT allows us to unified the implementation of time measurement and RPC communication on arbitrary IoT boards. Here is the corresponding measurement routine:

- 1) Analysis memory footprint at operator level utilizing internal API of TVM.
- 2) Deploy executable program to local or remote IoT board.
- 3) Start RPC Server on IoT board.
- 4) Launch RPC client to benchmark and record execution performance of each operator.

It is noted that the operator structure constructed by uTVM are usually inconsistent with the hand-crafted version in ML framework. That’s because uTVM as model compiler applies model optimization (i.e., operator fusion) and inserts execution details (i.e., quantization arithmetic) during convert and compilation, which potentially merge multiple operator into single one or insert additional operators. Nevertheless, we annotate the operators from uTVM with model parameters (weights, biases etc.), so that users can associate a specific operator to the corresponding layer.

V. EXPERIMENTS USING U-TOE

We conducted experiments to validate the functionality and compatibility of U-TOE on model side (universal support for model structure and ML frameworks) and on device side (wide-spectrum support for IoT devices). Hence, we dived into two orthogonal directions: For device support, we evaluated a quantized LetNet-5 on various IoT boards; For model compatibility, we evaluated multiple models on a local STM32F746G discovery board.

- **Model Selection** We selected pre-trained, quantized models from open source repositories³, which target on typical

³see <https://github.com/ARM-software/ML-zoo> and <https://github.com/mlcommons/tiny>

TinyML tasks (Image Classification, Key Word Spotting, Visual Word Wake, Noise Suppression and Abnormal Detection). The weights and activations of the model were quantized to 8-bit integer, yet the inputs and outputs remain in IEEE FP32 format.

- **Model Optimization** We only used built-in, rule-based optimization in uTVM. Thus, all heuristic optimization strategies like model scheduling were disabled.
- **MCU Configuration** We disabled data and instruction cache to observe the "memory wall" effect in ML model. The core clock frequency was pre-set by CPU initialization code in RIOT and is presented with experiment results in Section VI.
- **Hybrid Deployment** The experiments were conducted both on local and remote IoT boards provided by FIT-IoT Lab.

It is noted that for each evaluation we preset the number of trials to ten in order to address random error.

VI. ANALYSING U-TOE MEASUREMENTS

Per-Model Evaluation – Table II presents the resource consumption of LeNet-5 model on various IoT boards, generated by Per-Model evaluation. ARM Cortex-M series MCUs show no significant difference on memory and storage usage, and the computational latency declines as the core frequency increases. Benefits from fully support of DSP and Thumb-2 instruction set, Cortex-M3, -M4 MCUs perform better than Cortex-M0+ with the same core clock frequency. An **outlier** was discovered on SiFive RISC-V MCU. With the highest core clock frequency this won the least favorable ranking on computational latency and the memory usage. The SiFive RISC-V MCU uses an external, SPI NOR flash for data and program storage, causing a huge performance regression while we disabled the cache.

Table III presents the results of various ML models on representative TinyML tasks on single IoT board, proving the universal support of ML framework and model structure. Except for LeNet-5 trained on local host device with Pytorch, all the others came from open source model zoo. Memory and storage columns refer to the resource consumption.

Per-Operator Evaluation – We here used a tiny model with only three layers from TFLite as an example to avoid unnecessary complexity in demonstration, with output results presented in Table IV. The computational bottlenecks are located in operator *add_nn_relu* and *add_nn_relu_1*, and with the highest memory and storage consumption as well. We can trace down the corresponding layers of original model with the hints of associated parameters, which are the weights, bias or other trainable parameters of the model, making it possible to apply optimization strategies on well-targeted layers.

VII. REPRODUCIBLE & CUSTOM U-TOE EXPERIMENTS

We released the full source code of the U-TOE toolkit on Github at <https://github.com/zhaolanhuang/U-TOE/> under an open source LGPL v3 license. For further details on how to start with U-TOE hands-on, the reader is referred to the comprehensive *Readme.md* in the repository.

On the one hand, researchers and practitioners who possess IoT hardware that is supported by the open source operating system RIOT (currently 250+ types of boards, using 60+ types of CPUs⁴ can use U-TOE out-of-the-box, directly on their boards.

On the other hand, combined with the use of the free open-access testbed IoT-Lab⁵, even researchers and practitioners who do not have such hardware on premises can conduct large scale experimental evaluation campaigns using U-TOE.

Perspectives – As RIOT board and CPU support expands and improves over time, and as uTVM also expands support to other architectures in parallel (both open source communities are very active) U-TOE can in a very short time expand its support for new use cases, automatically adding the support of uTVM for new boards, and the support of RIOT for new models. As such U-TOE may organically grow and become a useful link between the two communities.

Moreover, while the work on U-TOE in this paper has been focused on inference only on single-core microcontrollers, there is strong potential to extend the toolkit provided by U-TOE to support on-device learning scenarios, and for optimizing exploitation of multi-core microcontrollers.

VIII. CONCLUSION

In this paper we presented U-TOE, a novel toolkit we designed to substantially facilitate the task of AIoT practitioners, by enabling universal TinyML model evaluation. Using U-TOE, the output of arbitrary machine learning frameworks can be evaluated on arbitrary low-power hardware based on different microcontroller architectures, all at once. The wide availability of such a toolkit is indeed desirable to accelerate the field of AIoT. We thus provided a highly re-usable, documented and customisable open source implementation of U-TOE, jointly harnessing the active open source communities around RIOT and uTVM. Finally, we demonstrate the use of U-TOE, by providing initial experimental evaluation results on popular low-power boards used in the TinyML community, for a wide variety of models from popular model zoo.

ACKNOWLEDGMENT

The authors would like to thank Cedric Adjih and Nadjib Achir for useful discussions and suggestions. The research leading to these results partly received funding from the MESRI-BMBF German/French cybersecurity program under grant agreements No. ANR-20-CYAL-0005 and 16KIS1395K. The paper reflects only the authors' views. MESRI and BMBF are not responsible for any use that may be made of the information it contains.

REFERENCES

- [1] Cedric Adjih et al. "FIT IoT-LAB: A large scale open experimental IoT testbed". In: *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*. IEEE. 2015.

⁴see <https://github.com/RIOT-OS/RIOT/tree/master/boards>

⁵see <https://www.iot-lab.info/>

TABLE II
EVALUATION OF LENET-5 MODEL ON VARIOUS IOT BOARDS.

Board / MCU	Core	Memory (KB)	Storage (KB)	Computational Latency (ms)			
				95%-CI	Median	Min.	Max.
b-l072z-Irwan1 / STM32L072CZ	M0+ @ 32 MHz	11.288	64.340	[261.829, 262.249]	262.187	261.350	262.216
samr30-xpro / ATSAMR30G18A	M0+ @ 48 MHz	11.208	65.168	[176.936, 176.965]	176.958	176.924	176.975
arduino-zero / ATSAMD21G18	M0+ @ 48 MHz	11.292	64.940	[182.061, 182.082]	182.068	182.051	182.098
rpi-pico / RP2040	M0+ @ 125 MHz	28.704	109.504	[70.108, 70.130]	70.117	70.091	70.151
openmote-b / CC2538SF53	M3 @ 32 MHz	11.100	66.080	[200.337, 200.384]	200.367	200.323	200.404
IoT-LAB M3 / STM32F103REY	M3 @ 72 MHz	11.296	62.260	[97.740, 97.757]	97.751	97.733	97.764
nucleo-wl55jc / STM32WL55JC	M4 @ 48 MHz	11.288	63.180	[98.649, 98.668]	98.661	98.637	98.679
nrf52840dk / nRF52840	M4 @ 64 MHz	11.348	61.332	[66.078, 66.112]	66.088	66.087	66.163
b-l475e-iot01a / STM32L475VG	M4 @ 80 MHz	11.288	61.604	[52.900, 52.901]	52.901	52.900	52.902
stm32f746g-disco / STM32F746NG	M7 @ 216 MHz	11.076	64.712	[39.600, 39.602]	39.601	39.599	39.604
esp32-wroom-32 / ESP32-D0WDQ6	ESP32 @ 80 MHz	115.958	157.719	[85.580, 85.583]	85.582	85.576	85.584
esp32c3-devkit / ESP32-C3FN4	RISC-V @ 80 MHz	258.874	222.272	[54.947, 54.957]	54.953	54.938	54.961
sipeed-longan-nano / GD32VF103CBT6	RISC-V @ 108 MHz	103.108	106.422	[37.783, 37.789]	37.789	37.779	37.791
hifive1b / SiFive FE310-G002	RISC-V @ 320 MHz	60.884	66.492	[153.621, 154.166]	153.747	153.717	154.938

TABLE III
EVALUATION OF VARIOUS QUANTIZED MODELS ON STM32F746-DISCO BOARD.

Model	Task	Memory (KB)	Storage (KB)	Computational Latency (ms)			
				95%-CI	Median	Min.	Max.
DS-CNN Small INT8**	Keyword Spotting	68.992	71.796	[461.395, 461.396]	461.396	461.396	461.397
MobileNetV1-0.25x INT8*	Visual Wake Words	185.352	491.668	[1435.937, 1435.938]	1435.938	1435.938	1435.939
LeNet-5 INT8	Image Classification	12.068	65.851	[39.599, 39.603]	39.601	39.598	39.605
Deep AutoEncoder INT8*	Anomaly Detection	6.532	292.696	[35.637, 35.638]	35.638	35.638	35.639
RNNNoise INT8**	Noise Suppression	4.688	119.652	[12.151, 12.157]	12.154	12.148	12.160

* These models originate from MLPerf Tiny Benchmarks repository on <https://github.com/mlcommons/tiny>.

** These models originate from ARM Model Zoo on <https://github.com/ARM-software/ML-zoo>.

All models were pre-trained and quantized by TFLite, except LeNet-5 INT8 was by Pytorch.

TABLE IV
PER-OPERATOR EVALUATION OUTPUT OF TFLITE SINUS MODEL ON STM32F746-DISCO BOARD.

Operators	Time (us)	Time (%)	Params	Memory	Storage
add_nn_relu	8.856	15.22%	p0, p1	0.128	0.128
add_nn_relu_1	46.682	80.23%	p2, p3	0.128	1.088
add	2.646	4.54%	p4, p5	0.068	0.068

The uTVM auto-generated prefix `tvmgen_default_fused_nn_dense_` of operator name is not presented for the purpose of clarity. Memory and storage consumption are presented in KB.

- [2] Emmanuel Baccelli et al. “RIOT: An open source operating system for low-end embedded devices in the IoT”. In: *IEEE Internet of Things Journal* 5.6 (2018).
- [3] Colby Banbury et al. “MLperf tiny benchmark”. In: *arXiv preprint arXiv:2106.07597* (2021).
- [4] Colby R Banbury et al. “Benchmarking tinyml systems: Challenges and direction”. In: *arXiv preprint arXiv:2003.04821* (2020).
- [5] Tianqi Chen et al. “TVM: An automated end-to-end optimizing compiler for deep learning”. In: *arXiv preprint arXiv:1802.04799* (2018).
- [6] Alexander Gluhak et al. “A survey on facilities for experimental internet of things research”. In: *IEEE Communications Magazine* 49.11 (2011), pp. 58–67. DOI: 10.1109/MCOM.2011.6069710.
- [7] Oliver Hahm et al. “Operating systems for low-end devices in the internet of things: a survey”. In: *IEEE Internet of Things Journal* 3.5 (2015), pp. 720–734.

- [8] Luis Eduardo Lima et al. “Experimental environments for the Internet of Things: A review”. In: *IEEE Sensors Journal* 19.9 (2019), pp. 3203–3211.
- [9] Anas Osman et al. “Tinyml platforms benchmarking”. In: *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2021*. Springer, 2022, pp. 139–148.
- [10] Hang Qiu et al. “ML-EXRay: Visibility into ML deployment on the edge”. In: *Proceedings of Machine Learning and Systems* 4 (2022), pp. 337–351.
- [11] Partha Pratim Ray. “A review on TinyML: State-of-the-art and prospects”. In: *Journal of King Saud University-Computer and Information Sciences* 34.4 (2022), pp. 1595–1623.
- [12] Swapnil Sayan Saha et al. “Machine learning for microcontroller-class hardware-a review”. In: *IEEE Sensors Journal* (2022).
- [13] Ramon Sanchez-Iborra et al. “Tinyml-enabled frugal smart objects: Challenges and opportunities”. In: *IEEE Circuits and Systems Magazine* 20.3 (2020), pp. 4–18.
- [14] Himanshu Sharma et al. “Machine Learning in Wireless Sensor Networks for Smart Cities: A Survey”. In: *Electronics* 10.9 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10091012. URL: <https://www.mdpi.com/2079-9292/10/9/1012>.
- [15] Bharath Sudharsan et al. “Tinyml benchmark: Executing fully connected neural networks on commodity microcontrollers”. In: *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*. IEEE. 2021, pp. 883–884.

- [16] Vivienne Sze et al. “Efficient processing of deep neural networks: A tutorial and survey”. In: *Proceedings of the IEEE* 105.12 (2017), pp. 2295–2329.
- [17] Ehsan Yousefzadeh-Asl-Miandoab et al. “Profiling and Monitoring Deep Learning Training Tasks”. In: *Proceedings of the 3rd Workshop on Machine Learning and Systems*. EuroMLSys ’23. DOI: 10.1145/3578356.3592589.
- [18] Zephyr Operating System. <https://www.zephyrproject.org/>. Accessed: 2023-05-15.