

TensorFlow

SIG Micro Meeting Notes

This document is public. Self link: goo.gle/tf-micro-notes

meet.google.com/roz-tpim-tjt or join by phone: +1 706-622-8334 PIN: 956 as 708#

Meetings are generally the second Wednesday of every month at 8:30am Pacific Time.

 [Group email discussion](#)

 [Gitter chat channel](#)

 [SIG Micro charter](#)

SIG Micro Community Meeting

Time zone: America/Los_Angeles

Google Meet joining info

Video call link: <https://meet.google.com/roz-tpim-tjt>

Or dial: (US) +1 706-622-8334 PIN: 956 708#

More phone numbers: <https://tel.meet/roz-tpim-tjt?pin=1236753102492>

[Master Doc for all RFCs](#)

2024-08-14

AGENDA

Attendees

RJ Ascani, rjascani@google.com

Oleksiy Peutin, oleksiy.peutin@infineon.com

Ryan O'Shea

Notes

- Per-Channel Quantized FC kernels will be coming soon. Both Per-channel & Per-tensor will be supported in the kernels.
- Weight compression experimentation continues. Hopeful to have some benchmarks by the next meeting.

2024-06-12

AGENDA

Attendees

RJ Ascani, rjascani@google.com

Måns Nilsson, mans.nilsson@arm.com

Andrew Stevens, andrew.stevens@infineon.com

Notes

- TFLiteConverter now defaults to per-channel quantization for fully connected/dense layers (believe this was in 2.16).
 - This not supported in TFlite-Micro
 - There is now a runtime check within FullyConnected operators that fails tensor allocation if per-channel quantization is detected in FullyConnected.
 - On our roadmap to support the per-channel quantization for FullyConnected. Optimized kernels will come later.
 - For more details:
<https://github.com/tensorflow/tflite-micro/pull/2602>
- Compressing weight tensors in the flatbuffer
 - TFLM team is looking at adding framework support for compressed weights in the TFLite Flatbuffer.

- Nothing in the repo yet, but design is getting finalized and will start showing up shortly.
- General plan is to embed compression parameters into a Metadata Buffer using a separate TFLM-specific schema.
- Interpreter will parse this metadata and provide accessors to kernels for decompression, akin to INT4 unpacking.
- The runtime will provide helpers for decompressing into a scratch buffer, but kernels will also be able to access the compression parameters for on the fly decompression.
- Andrew Stevens (Infineon) has actually done something similar on their fork, where they use custom quantization parameters to embed compression parameters, and then kernels would be responsible for decompression.
- This raises the suggestion that we may want to consider allowing custom decompression schemes as well, so that users can add their own.
- AI Edge Torch
 - Google announced support for PyTorch models with TFLite using AI Edge Torch at I/O
 - <https://github.com/google-ai-edge/ai-edge-torch>
 - Does TFLM support this? Short answer: We think so, but we haven't really tried it out yet.
 - AI Edge Torch generates a TFLite flatbuffer file. TFLM supports the most common TFLite ops, but not all. If the ops generated by AI Edge Torch for a given PyTorch program are within the subset of TFLM supported ops, then it *should* work.

2024-05-08

AGENDA

Attendees

RJ Ascani, rjascani@google.com

Måns Nilsson, mans.nilsson@arm.com

Deepak Panickal, deepakpanickal@xmos.com

Notes

- No significant updates from Google side
- Request: Can we get meeting notes sent out to SIG Micro mailing list?
 - RJ: This seems reasonable and can be done.
- Will TFLM support other model formats beyond the TFLite file?
 - No plans on that at the moment.

2024-02-14

AGENDA

Attendees

Deepak Panickal, deepakpanickal@xmos.com

RJ Ascani, rjascani@google.com

Neil Tan, neil.c.tan@gmail.com

Notes

RJ to look into tflite-micro PyPi package for OSX & Windows

2023-12-13

AGENDA

Attendees

2023-11-08

AGENDA

Attendees

Notes

2023-10-11

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Notes

Readying the PyPl package

Updating examples to use new custom ops (TF Lite Micro Signal directory)

8-bit weight 16-bit activation, 2 step process:

2023-09-13

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Notes

Code for Signal Processing is growing

PiPy Packaging coming soon

2023-08-09

AGENDA

Attendees

Advait Jain, advaitjain@google.com

RJ Ascani, rjascani@google.com

Notes

What's New

- Signal Ops are now available in GitHub repository
 - End to end example is in progress
 - Would appreciate some feedback
 - Must use the custom signal TF ops from the TFLM repository.

- PyPi package is close, but not quite available yet

Q&A

- Does TFLM support integer only platforms?
 - No, the Init & Prepare stages will convert floating point to fixed point integer, but it does require at least floating point emulation. Eval does not require it.
 - Floating point libraries can be approximately 20 KiB in binary size as well.
- Merlin Carson ran into a possible bug in the requantization script. Will work on a PR.
- How will the TFLite converter handle the signal ops?
 - Converter treats these custom ops as a pass through

2023-07-11

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

2023-06-14

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

- TFLM Pypi package
- Audio processing as custom-op: first op -> window op
- TFLM model transform tool: python script manipulates flatbuffer; 5~20% size reduction for TFLM interpreter

2023-05-10

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

Tflite-compiler or edge impulse like tool to generate a C code from a model

2023-04-12

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

Example Clean-up, 16-bit LSTM completed, Python script to support 8-bit wight + 16-bit activation

2023-03-08

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

- New LSTM kernel with 16 bit activation + CMSIS-NN merged

2023-02-08

AGENDA

Attendees

Neil Tan, neil@utensor.ai

Proposed Agenda

- New LSTM kernel with 16 bit activation (WIP)
 - Related [CMSIS-NN PR](#)
- TFLM python interpreter (WIP)

2023-01-11

AGENDA

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agenda

Upcoming embedded DSP Ops TFLM (early release next month) 2022-11-9

AGENDA

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agenda

TF Lite vs TFLM reference kernel bit-exact?

Custom Offline-memory planner

2022-10-12

AGENDA

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agenda

- [Nvidia's study on effect of quantization](#) (Rafal Firlejczyk)
-

2022-9-14

AGENDA

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agenda

Sincnet and Sinc operator

- Fredrick: <https://github.com/ARM-software/CMSIS-DSP>
- Advait: Custom-Op or Composing built-in Ops
- MFCC currently using kiss-fft

2022-8-10

AGENDA

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agenda

4-bit quantization work-in-progress

Removing strings in flatbuffers significantly reduces the final sizes

2022-3-08

AGENDA

Roll call

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agendax

LSTM from TFL to TFLM, fused-op vs fine-grain-ops

Memory footprint script - idea/brainstorming on github before RFC

2022-2-09

AGENDA

Roll call

Attendees

Neil Tan, neil.c.tan@gmail.com

Proposed Agendax

TFLM memory footprint script - open for contribution (Advait <-> Fredrik)

2022-1-11

AGENDA

Roll call

Attendees

Neil Tan, neil.c.tan@gmail.com

Yi-Lin Tung, yilintung@on-device-ai.com

Pete Warden, petewarden@google.com

Fredrik Knutsson, fredrik.knutsson@arm.com

Michael Gielda, mgielda@antmicro.com

Yair Ehrenwald, yair.ehrenwald@ceva-dsp.com

Jacco, jacco@synopsys.com

Annie Tallund, annie.tallund@arm.com

Soham Chatterjee, 96soham96@gmail.com

Jens Elofsson, jens.elifsson@arm.com

Henrique Faria, hd.faria@poli.ufrj.br

Josh Morris, jmorris@dspconcepts.com

Proposed Agendax

- TF 2.7 MLIR converter problems
 - Workaround suggested with concrete function
- RNN work?
- Subgraphs and loops?
- Time-delayed neural network?
 - <https://arxiv.org/pdf/2005.06720.pdf>
-

2021-12-8

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Neil Tan, neil.c.tan@gmail.com

Fredrik Knutsson, fredrik.knutsson@arm.com

Andrew Stevens, andrew.stevens@infineon.com

Pete Warden, petewarden@google.com

Jens Elofsson, jens.elifsson@arm.com

Dmytro Korablyov, dmytro.korablyov@gmail.com

Robert John, robert.thas.john@gmail.com

Jacco, jacco@synopsys.com

Josh Morris, jmorris@dspconcepts.com

Michael Gielda, mgielda@antmicro.com

Vikram Dattu, vikram.dattu@espressif.com

Proposed Agenda

- Mini-item: status of “tagged versions as pseudo-releases” issue
 - Aligning with other people is useful with tags, better than SHA
- Mini-item: tflite converter tflite(u) ops selections = sane diagnostics.
 - Converter doesn’t care, requires a runtime test to detect
 - Has been a long-standing feature request internally at Google
 - Causes problems for implementors
- MLIR approach to converter adds extra ops like strided slice, and reshapes
 - Can someone external file a bug? That will help us boost the priority with the converter team
- Transformer usage question
 - Arm has had interest, and Infineon
 - Are transformers too large?
- Ambiq SDK integration, with SparkFun Edge board has broken
 - Robert will file a bug and send Pete an email
- Zephyr integration
 - initial integration exists in Zephyr for the LiteX/VexRiscv platform, but it’s mostly our demo merged as-is
 - a better integration would be useful for Antmicro’s Renode/Zephyr dashboard, now in matrix form:
<https://zephyr-dashboard.renode.io/matrix.html>
 - Fredrik: Ethos-U driver for Zephyr in the works now, coming in 1-2 months
 -

2021-11-10

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Neil Tan, neil.c.tan@gmail.com

Andrew Stevens, andrew.stevens@infineon.com

Yi-Lin Tung, yilintung@on-device-ai.com

Måns Nilsson, mans.nilsson@arm.com

Jacco, jacco@synopsys.com

Abi Aryan, goabiaryan@outlook.com

Dmytro Korablyov, dmytro.korablyov@gmail.com

Josh Morris, jmorris@dspconcepts.com

Jerry Kurata, jerryk@insteptech.com

Henrique Faria, hd.faria@poli.ufrj.br

Yair Ehrenwald, yair.ehrenwald@ceva-dsp.com

Chris Knorowski, chris.knorowski@sensiml.com

Proposed Agenda

- Version tags proposal?
 - It's integrated with ST, choose a commit.
- Memory planning in progress
- TF contributors' summit
- Memory size
 - Andrew - Moving init and prepare phases offline

- Arm working on an automatic op resolver code generator as part of the converter

2021-10-13

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Neil Tan, neil.c.tan@gmail.com

Abi Aryan, abiaryan.chat@gmail.com

Michael Gielda, mgielda@antmicro.com

Proposed Agenda

- **Jacco:** Synopsys has released [version 2.0](#) of its Machine Learning Inference (MLI) library, adding support for the [DesignWare ARC VPX DSP processors](#). The VPX is a powerful VLIW/SIMD vector core with support for 512, 256 and 128 bit vector words. The ARC-specific backend library for TensorFlow Lite for Microcontrollers has also been updated to add support for the VPX processor family
- Slides on TFL roadmap are at <https://goo.gle/3lDZnfs>
- Renode Zephyr Dashboard: <https://zephyr-dashboard.renode.io/>
- Official version tags for the repo would be very useful.
- Keil studio example, micro speech on TFLM
 - <https://github.com/KeilChris/tflmicrospeech>
 - <https://studio.keil.arm.com/>

- LSTM work is on-going, PR from Cadence at <https://github.com/tensorflow/tflite-micro/pull/473>
-

2021-09-8

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Neil Tan, neil.c.tan@gmail.com

Yi-Lin Tung, yilintung@on-device-ai.com

Proposed Agenda

Notes

2021-08-11

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Pete Warden, petewarden@google.com

Neil Tan, neil.c.tan@gmail.com

Yi-Lin Tung, yilintung@on-device-ai.com

Jens Elofsson, jens.elifsson@arm.com

Tim Callahan, tcal@google.com

Chris Knorowski, chris.knorowski@sensiml.com

Jerry Kurata, jerryk@insteptech.com

Marconi Jiang marconi.jiang@gmail.com

Yahya Tawil, yt@16lab.net

Jacco, jacco@synopsys.com

Proposed Agenda

- Pete Warden: LSTM prototype -
<https://github.com/tensorflow/tflite-micro/pull/281>
- Pete: Google engineers are making progress on CI for the new repo, and working on a new version of the Arduino library.
- Colab notebook for reading and writing TFL flatbuffer files from Python:
https://colab.research.google.com/drive/11R9hd_0zvW0y3mesitNtF8dCjf27xP9I?usp=sharing
-

Notes

2021-07-14

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Pete Warden, petewarden@google.com

Neil Tan, neil.c.tan@gmail.com

Yi-Lin Tung, yilintung@on-device-ai.com

Nicholas Ollis, nic@oll.is

Marconi Jiang, marconi.jiang@gmail.com

[Dmytro Korablyov](mailto:Dmytro.Korablyov@gmail.com), dmytro.korablyov@gmail.com

Jacco van Gorp, jacco@synopsys.com

Andrew Stanford-Jason, andrew@xmos.com

Michael Gielda, mgielda@antmicro.com

Tim Callahan, tcal@google.com

Arijit Das, arijitdas180220006@gmail.com,

Henrique Faria, hd.faria@poli.ufrj.br

[Fredrik Knutsson](mailto:Fredrik.Knutsson@arm.com), fredrik.knutsson@arm.com

[Yair Ehrenwald](mailto:Yair.Ehrenwald@ceva-dsp.com) yair.ehrenwald@ceva-dsp.com

[Stefano Fabri](mailto:Stefano.Fabri@deeper.eu) (Deeper, s.fabri@deeperception.eu)

Proposed Agenda

[Repo Move](#) - Pete Warden

Courses and webinars on Zephyr - Michael Gielda @ Antmicro

Notes

Repo Move:

<https://github.com/tensorflow/tflite-micro/pulls>

<https://github.com/tensorflow/tflite-micro/actions>

Arduino project generation - David Davis @ BDTI

Zephyr courses:

- Zephyr webinar together with Nordic
- Also a separate RISC-V + Zephyr course, vol 2 should cover TFLM
- Will share links in the future
- also added Renode support for Bluetooth on nrf52840

<https://experiments.withgoogle.com/collection/tfliteformicrocontrollers>

2021-06-09

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Pete Warden, petewarden@google.com

Neil Tan, neil.c.tan@gmail.com

Arijit Das, arijitdas18022006@gmail.com

Yi-Lin Tung, yilintung@on-device-ai.com

Nicholas Ollis, nic@oll.is

Dmytro Korablyov, dmytro.korablyov@gmail.com

Andrew Stanford-Jason andrew@xmos.com

james.he@onsemi.com

jacco@synopsys.com

Yair Ehrenwald yair.ehrenwald@ceva-dsp.com

[Piotr Zierhoffer, pzierhoffer@antmicro.com](mailto:pzierhoffer@antmicro.com)

[Chris Knorowski chris.knorowski@sensiml.com](mailto:chris.knorowski@sensiml.com)

yair.ehrenwald@ceva-dsp.com

Proposed Agenda

CFU Playground -- Tim Callahan

Introducing [TensorFlow Forum](#)

Block floating point support - Andrew

Notes

Support for 'if' operator and subgraph

Legacy TFLM repo link - read only

2021-05-12

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

Pete Warden, petewarden@google.com

James He, james.he@onsemi.com

Andrew Stevens, andrew.stevens@infineon.com

Yi-Lin Tung, yilintung@on-device-ai.com

Henrique Faria, hd.faria@poli.ufrj.br

[Yair Ehrenwald, yair.ehrenwald@ceva-dsp.com](mailto:yair.ehrenwald@ceva-dsp.com)

[Fredrik Knutsson, fredrik.knutsson@arm.com](mailto:fredrik.knutsson@arm.com)

Andrew Stanford-Jason andrew@xmos.com

Proposed Agenda

- I/O is coming up! Workshop is at <https://events.google.com/io/session/ffd819da-d846-488d-a814-b383c8d8793c?lng=en>
- Google team is focused on binary size improvements, and moving the repo over in GitHub to make PRs easier.
- Are we in a freeze for optimized kernels still?
 - Pete - Yes for now, but can work in fork, repo move will help in the future
- Is there a list of C++11 features needed?
 - Will meet offline to discuss
- PR to get Micro into Zephyr has merged
 - Course for Linux Foundation is in-progress
- CFU Playground is bit.ly/cfuplay, an example of FPGA ML acceleration, Tim Callahan will present on this next month.
- Have updated Renode visualizer, Michael will add link, will demo next month. <https://github.com/antmicro/renode-board-visualization>
-

2021-04-14

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- Andrew Stevens, andrew.stevens@infineon.com
- **Christof Petig**, christof.petig@arcor.de
- Chris Knorowski, chris.knorowski@sensiml.com
- Arijit Das , arijitdas18022006@gmail.com
- Jacco, jacco@synopsys.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Fredrik Knutsson fredrik.knutsson@arm.com
- Vikram Dattu vikram.dattu@espressif.com
- Daniel Situnayake dan@edgeimpulse.com
- Tim Callahan, timo.callahan@gmail.com
- Michael Gielda, mgielda@antmicro.com
- Morten Opprud, opprud@gmail.com
- Jerry Kurata, jerry340@gmail.com
- **Marconi Jiang**, marconi1964@yahoo.com
- Stefano Fabri, s.fabri@deeperception.eu (Deeper/TF community Team)

Proposed Agenda

- BDDTI op porting has 12 ops checked in, several more coming
- TF Lite Micro being separated out into
<https://github.com/tensorflow/tflite-micro>
- Edge-case: efficient-but-portable kernels for simple "baseline" CPUs.
- Zephyr + TF Lite PR (mgielda@antmicro.com)
 - <https://github.com/zephyrproject-rtos/zephyr/pull/34057>
- [8x16 quantized kernel port RFC](#) ([dan@edgeimpulse](mailto:dan@edgeimpulse.com))

Notes

- Link to Petig/Stahl offline pre-interpreter is
https://github.com/cpetig/tflite_micro_compiler
Infineon working activity is in ifx/* branches this urgently needs to be updated. I'll try to get a sanitised push out ASAP (andrew.stevens@infineon.com).

2021-03-11

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- **Fredrik Knutsson**, fredrik.knutsson@arm.com
- Jens Elofsson, jens.elifsson@arm.com
- Tim Callahan, timo.callahan@gmail.com
- Michael Gielda, mgieldda@antmicro.com
- Yair Ehrenwald yair.ehrenwald@ceva-dsp.com
- Keith Moulton, keithm@xmos.com
- Scott Roset, scottroset@xmos.com

Proposed Agenda

Notes

- Op porting: Mobile to Micro ongoing
- Binary size-shrink
- Arm: RFC 16-bit quantized kernels

Antmicro: [Renode for education note](#)

2021-02-10

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- Fredrik Knutsson, fredrik.knutsson@arm.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- [Jacco van Gurp, jacco@synopsys.com](mailto:jacco@synopsys.com)
- Michael Gielda, mgielda@antmicro.com
- Tim Callahan, timo.callahan@gmail.com
- Andrew Stevens, andrew.stevens@infineon.com
- [Chris Knorowski, chris.knorowski@sensiml.com](mailto:chris.knorowski@sensiml.com)
- Keith Moulton keithm@xmos.com
- Yair Ehrenwald yair.ehrenwald@ceva-dsp.com
- Nicholas Ollis nic@oll.is
- Arijit Das arijitdas18022006@gmail.com
- Jens Elofsson, jens.elifsson@arm.com

Proposed Agenda

- [new magic wand demo](#) - gyroscope + accelerometer
- Potential streaming support in future - <https://arxiv.org/pdf/2005.06720.pdf>
- Raspberry Pi Pico : <https://github.com/raspberrypi/pico-tflmicro>

Notes

2021-01-13

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- Fredrik Knutsson, fredrik.knutsson@arm.com
- Daniel Situnayake, dan@edgeimpulse.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Jens Elofsson, jens.elifsson@arm.com
- Andrew Stevens, andrew.stevens@infineon.com
- Tim Callahan, timo.callahan@gmail.com
- Michael Gielda, mgielda@antmicro.com
- Henrique D. Faria, hd.faria@poli.ufrj.br
- Usama Hameed usamahameed@g.ucla.edu
- Keith Moulton keithm@xmos.com
- Rob Fisher robfxmos.com
- Christof Petig, christof.petig@arcor.de
- Yair Ehrenwald yair.ehrenwald@ceva-dsp.com

Proposed Agenda

- GCC 10.2 (Fredrik)
- RFC Preallocated tensors (Fredrik, Jens Elofsson)
- 8x16 quantized kernel port (Dan)
 - Planned RFC/work
 - Preferred way to avoid increasing code size?
- Arduino library release (Pete)
- PR Freeze Status (Andrew)
- TF Lite examples work

- <https://github.com/antmicro/tensorflow-arduino-examples>

Notes

- GCC 10.2: Cortex M55 ML workload optimization, improved performance.
- RFC Preallocated tensors: Tensor can address a different memory bank from the memory arena
- PR Freeze Status: on-going, progressing

2020-12-09

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- [Fredrik Knutsson](mailto:fredrik.knutsson@arm.com), fredrik.knutsson@arm.com
- [Yair Ehrenwald](mailto:yair.ehrenwald@ceva-dsp.com) yair.ehrenwald@ceva-dsp.com
- [Jacco van Gurp](mailto:jacco@synopsys.com), jacco@synopsys.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com,
- [Michael Gielda](mailto:mgielda@antmicro.com), mgielda@antmicro.com
- [Andrew Stevens](mailto:andrew.stevens@infineon.com) andrew.stevens@infineon.com
-
- Tim Callahan, tc@cal@google.com
- Henrique Faria, hd.faria@poli.ufrj.br
- Thomas Garcia, thomas.garcia@microchip.co
- Aditya Patwardhan

aditya.patwardhan@espressif.com

Proposed Agenda

- Corstone-300 FVP available (model with Arm Cortex-M55 and Ethos-U55) (Fredrik Knutsson)

- <https://developer.arm.com/tools-and-software/open-source-software/arm-platforms-software/arm-ecosystem-fvps>
- BDTI op porting work - petewarden@google.com
- Optimization strategy (e.g. <https://github.com/tensorflow/tensorflow/pull/45227>) - petewarden@google.com
- [mgielda] Arduino IDE simulation capability
- [mgielda] Tensorflow lite micro examples repo, colabs for video and audio

2020-11-11

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Dor Livne, dor.livne@dspg.com
- Niv Peled, niv.peled@dspg.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Yair Ehrenwald yair.ehrenwald@ceva-dsp.com
- [Jacco van Gurp, jacco@synopsys.com](mailto:jacco@synopsys.com)
- [Tim Callahan, tcal@google.com](mailto:tcal@google.com)
- [Jens Elofsson, jens.elifsson@arm.com](mailto:jens.elifsson@arm.com)
- [Fredrik Knutsson, fredrik.knutsson@arm.com](mailto:fredrik.knutsson@arm.com)
- [Christof Petig, christof.petig@arcor.de](mailto:christof.petig@arcor.de)
- [Daniel Situnayake, dan@edgeimpulse.com](mailto:dan@edgeimpulse.com)
- [Andrew Stevens, andrew/stevens@infineon.com](mailto:andrew/stevens@infineon.com)
- Keith Moulton, keithm@xmos.com
- Scott Roset, scottroset@xmos.com
- [Tarek Allam, t.allam.jr@gmail.com](mailto:t.allam.jr@gmail.com)

Proposed Agenda

- DSPG to present weight compression proposal
- [Michael] Antmicro to discuss Renode CI integration
 - focus on Nordic Cortex M4F-based Arduino Nano 33 BLE
 - also adding new RISC-V FPGA SoC platform
 - many sensors, ability to record audio and video and get it process it
 - ML metrics
 - also adding Arduino integration
 - creating a new repo with tested examples + Jupyter Notebooks

2020-10-14

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
 - Neil Tan, neil.c.tan@gmail.com
 - Yi-Lin Tung, yilintung@on-device-ai.com
 - Fredrik Knutsson, fredrik.knutsson@arm.com
 - Jacco van Gurp, jacco@synopsys.com
 - Michael Gielda, mgielda@antmicro.com
 - Yair Ehrenwald yair.ehrenwald@ceva-dsp.com
 - Aditya Patwardhan aditya.patwardhan@espressif.com
 - Andrew Stevens, andrew.stevens@infineon.com
-
- Tim Callahan, tc@google.com
 - Daniel Müller-Gritschneider, daniel.mueller@tum.de

- Keith Moulton, keithm@xmos.com

Proposed Agenda

- Pre-allocated tensors (Fredrick)
- CMSIS patching (Fredrick)
- CMSIS-NN SVDF kernel merged (Fredrick)
- Kernel API changes (Advait Jain)
- Op Porting Guide (Advait Jain)
- TFL-Micro paper coming up (Pete)

2020-09-09

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Fredrik Knutsson, fredrik.knutsson@arm.com
-
- Yi-Lin Tung, yilintung@on-device-ai.com
- Yair Ehrenwald, yair.ehrenwald@ceva-dsp.com
- Michael Gielda, mgielda@antmicro.com
- Tim Callahan, timo.callahan@gmail.com
- Andrew Stevens, andrew.stevens@infineon.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Jens Elofsson, jens.elifsson@arm.com
- Christof Petig, christof.petig@arcor.de
- Keith Moulton, keithm@xmos.com

Proposed Agenda

- [New contribution guidelines](#)

- Pete: redoing bash/make -> Python
- Andrew: separating TF Lite out as a module would be great
- Advait: put ideas for code reorg in issues! -- TF Lite team
- [Renode colab work](#) - working on a TF Lite one

Notes from chat:

2020-08-12

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- **Pete Warden**, petewarden@google.com
- **Neil Tan**, neil.c.tan@gmail.com
- Jens Elofsson, jens.elifsson@arm.com
- [Fredrik Knutsson](mailto:fredrik.knutsson@arm.com), fredrik.knutsson@arm.com
- Yi-Lin Tung, yilintung@on-device-ai.com
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Lukas Geiger, lukas@plumerai.com
- [Peter Torelli](mailto:peter.torelli@eembc.org), EEMBC, peter.torelli@eembc.org
- [Sergii Khomenko](mailto:khomenko@brainscode.com), khomenko@brainscode.com
- [Jacco van Gorp](mailto:jacco@synopsys.com), jacco@synopsys.com
- **Andrew Stevens**, andrew.stevens@infineon.com
- [Michael Gieldda](mailto:mgieldda@antmicro.com), mgieldda@antmicro.com
- [Aditya Patwardhan](mailto:aditya.patwardhan@espressif.com) aditya.patwardhan@espressif.com
- [Advait Jain](mailto:advaitjain@google.com), advaitjain@google.com
- Ian Nappier, iannappier@google.com
- Keith Moulton, keithm@xmos.com

Proposed Agenda

- Breaking changes to tensor access in eval stage
- Pull request pause
- MLPerf progress
- Harvard/EdX course announcement:
<https://programs.edx.org/harvard-tiny-ml>
-

Notes from chat:

2020-07-08

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Pete Warden, petewarden@google.com
- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Michael Gielda, mgielda@antmicro.com
- [Andrew Cavanaugh, andrewc@xmos.com](mailto:andrewc@xmos.com)
- Chris Knorowski, chris.knorowski@sensiml.com
- Jens Elofsson, jens.elifsson@arm.com
- Daniel Situnayake, dan@edgeimpulse.com
- Andrew Stevens , andrew.stevens@infineon.com
- Stephan Uphoff, ups@stups.com
- Tim Callahan, timo.callahan@gmail.com
- [Lennart Heim, leheim@ethz.ch](mailto:leheim@ethz.ch), ETH Zurich
- [Christof Petig, christof.petig@aptiv.com](mailto:christof.petig@aptiv.com)

- [Ajay Gopi, ajay@edgeimpulse.com](mailto:ajay@edgeimpulse.com)
- [Daniel Mueller-Gritschneider, daniel.mueller@tum.de](mailto:daniel.mueller@tum.de)
- Ian Nappier, iannappier@google.com

Proposed Agenda

- CMSIS optimizations and Arduino - Pete Warden
- The "right place" to pack narrow constants? - Andrew Stevens
- Quick intro to Ajay Gopi - Dan Situnayake
- Vision model training with Edge Impulse in case you need models to test with - Dan Situnayake
- Optimized SVDF kernel in CMSIS-NN in progress (Jens Elofsson, arm)
- New speech training data release by Mozilla,
<https://discourse.mozilla.org/t/common-voice-dataset-release-mid-year-2020/62938>, Pete Warden

Notes from chat:

Optimized kernels are at

<https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/micro/kernels>

Examples of rewriting flatbuffers:

https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/tools/randomize_weights.py

<https://colab.research.google.com/drive/1h0eFudI7gy5AGYdxBFdoVpggwBN5ftMu?usp=sharing>

https://github.com/uTensor/utensor_cgen/pull/93

2020-06-10

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Jens Elofsson, jens.elfsson@arm.com
- Fredrik Knutsson, fredrik.knutsson@arm.com
- Neil Tan, neil.c.tan@gmail.com
- Pete Warden, petewarden@google.com
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- [Jacco van Gorp, jacco@synopsys.com](mailto:jacco@synopsys.com)
- [Chris Knorowski, chris.knorowski@sensiml.com](mailto:chris.knorowski@sensiml.com)
- Henrique Faria, hd.faria@poli.ufrj.br
- Dan Goncharov, gdanfx@gmail.com
- [Andrew Stevens, andrew.stevens@infineon.com](mailto:andrew.stevens@infineon.com)
- Vijay Janapa Reddi (vj@eecs.harvard.edu)
- Michael Gielda, mgielda@antmicro.com
- Christof Petig, christof.petig@arcor.de, Aptiv Services
- Daniel Mueller-Gritschneider, daniel.mueller@tum.de, Technical University of Munich
- Ian Nappier, iannappier@google.com

Proposed Agenda

- Benchmarks -- [tinyMLPerf](#) (Vijay Janapa Reddi, Harvard/MLPerf)
- Ethos-U55: Integration landed. Driver open sourced. (Fredrik Knutsson)
- Offline memory planner, merge ongoing
- Updated CMSIS-NN interface
- Eventrecorder updates
- Tech talk: CMSIS-NN & Ethos-U55 integration to TFLu
- Update on pull request process (Pete Warden)
- TF Lite + Renode [note](#) up! (Michael Gielda)
- Vijay: teaching using Renode, Arduino IDE with Renode

- Awesome announcement next week around open source FPGA SoC platform from QuickLogic (Michael Gielda)

Notes from chat:

https://docs.google.com/document/d/1XPkeg3_RTQ_uN1y6_PXWYf0aV1vYwP0pW4LRgAe30_o/edit

Ethos-U55 presentation: <https://www.youtube.com/watch?v=5j1WqV01VbU>

The tech talk I mentioned:

<https://developer.arm.com/solutions/machine-learning-on-arm/ai-virtual-tech-talks>

Arduino setup challenges:

https://docs.google.com/document/d/1sPVGiuyByJUdUK4A-x3U80RB-NRwsStdIF16vjkXo_c/edit

LSTM support: LSTM and GRU ops need to be ported

2020-05-13

AGENDA

[Video Recording](#) for 2020-05-13 meeting!

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Pete Blacker, Pete.Blacker@Surrey.ac.uk
- Jens Elofsson, jens.elifsson@arm.com
- Phil Valencia, philip.valencia@csiro.au

- [Pete Warden, petewarden@google.com](mailto:petewarden@google.com)
- [Daniel Situnayake, dan@edgeimpulse.com](mailto:dan@edgeimpulse.com)
- [Michael Gielda, mgielda@antmicro.com](mailto:mgielda@antmicro.com)
- [Chris Knorowski, chris.knorowski@sensiml.com](mailto:chris.knorowski@sensiml.com)

Proposed Agenda

- Question on TF Lite Interpreter Tensor life-cycle (Neil Tan)
- Layer-by-layer measurement (Jens Elofsson)
- Scratch memory API (Pete Warden)
- Benchmarks (Pete Warden)
- SVDF kernel (Jens Elofsson)
-

Notes

- Example of loading, modifying and saving model file in Python Colab is at
<https://colab.research.google.com/drive/1h0eFudI7gy5AGYdxBFdoVpggwBN5ftMu?usp=sharing>
- “Zip” automatic test generation comparing TF and TFL op results can be found at
https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/testing/generate_examples.py
- Scratch buffer API is documented in comments at
<https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/c/common.h#L591>
- Benchmarks are at
<https://github.com/tensorflow/tensorflow/tree/master/tensorflow/lite/micro/benchmarks>

2020-04-08

AGENDA

[Video Recording](#) for 2020-04-08 meeting!

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Dmytro Korablyov, dmytro.korablyov@gmail.com
- Daniel Situnayake, dan@edgeimpulse.com, Edge Impulse
- Michael Gielda, mgielda@antmicro.com, Antmicro
- Andrew Cavanaugh, andrewc@xmos.com, XMOS Ltd.
- [Ian Nappier, Google](#)
- [Fredrik Knutsson, Arm](#)
- Jacco van Gorp (jacco@synopsys.com), Synopsys, Netherlands
- Andrew Stevens, andrew.stevens@infineon.com , [Infineon Technologies AG](#)
- Stephan Uphoff, ups@stups.com
- Pete Warden, petewarden@google.com
- Jens Elofsson, jens.elifsson@arm.com, ARM
- Chris Knorowski, chris.knorowski@sensiml.com, SensiML
- Joana Carraqueira joanafilipa@google.com

Proposed Agenda

- Are tensor names optional in the converted .tflite?
 - See https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/tools/strip_strings.py
- CI testing in Renode portable
 - no need for Docker in Docker now
- benchmarking in Renode
- [Zephyr support PR](#)
- Pre-calculating Tensor Arena size

- Determining model compatibility before runtime
 - Ethos-U55 integration PR
 - Generate Projects with full make file PR
- <https://github.com/tensorflow/tensorflow/pull/38175/files>

Notes

- updates: great work from many people: Arm, Cadence, Synopsys etc.
- working with ML perf group, hired VJ (?) for couple days a week to work on benchmarks for embedded ML, hoping for some benchmarks in the coming months
- Daniel: is there a possibility to collaborate
- Pete: yes, the ML perf meetings are open, happy to connect anyone who wants to VJ directly
- Pete: had to scale down work in this quarter, so need to ask for even more patience as they work on things, many people have lost childcare
- Michael: portable Renode package with Robot tests capability ready, sent to Pete; we can also do some benchmarking in Renode so interested to talk to VJ too
- merging Zephyr PR would offer good marketing opportunities

2020-03-11

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Neil Tan, neil.c.tan@gmail.com
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd. Andrew Cavanaugh, andrewc@xmos.com, XMOS Ltd.

- [Fredrik Knutsson, fredrik.knutsson@arm.com](mailto:fredrik.knutsson@arm.com)
- Pete Warden, petewarden@google.com
- Jacco van Gurp (jacco@synopsys.com), Synopsys, Netherlands
- Dominic Symes, dominic.symes@arm.com, Arm, UK
- Stephan Uphoff (ups@stups.com)
- Anuraj R (anuraj.rp@gmail.com)
- Jens Elofsson, ARM, Sweden (jens.elifsson@arm.com)
- Michael Gielda, Antmicro (mgielda@antmicro.com)
- Morten opprud opprud@gmail.com
- Ian Nappier iannappier@google.com

Proposed Agenda

- Fixed-point input (input/output type)
- Memory allocation API added
- Renode test CMSIS-NN (x86 test)
- Offline memory planner status

Notes

- Profiling feature for TFLM
- Passing-in buffer size for scratch buffers

2020-02-12

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Neil Tan, neil.tan@arm.com, Arm
- Joana Carrasqueira (joanafilipa@google.com)TF Team @Google

- Pete Warden (petewarden@google.com)

Proposed Agenda

Cortex M55 and NPU U55 via CMSIS-NN?

Mbed library [setup](#) with standalone library (turn this into a makefile build, check for existing work on the CMSIS-NN folks)

Notes:

TF Lite Mobile has just added an experimental cancellation function:

<https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/interpreter.h#L351> - The cancellation happens in between operations.

Pete is supportive of this and he hopes that they can follow the TF Lite API.

We need to make sure the cooling sequence is right.

2020-01-08

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Attendees

- Neil Tan, neil.tan@arm.com, Arm
- Joana Carrasqueira, Google

Proposed Agenda

- (Neil to share his example) (**uint8 to int8**)
- A centralized place for all the RFCs? (Pete W. to help)

2019-12-12

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

TensorFlow

SIG Micro Meeting Notes

This document is public. Self link: goo.gle/tf-micro-notes

meet.google.com/roz-tpim-tjt or join by phone: +1 706-622-8334 PIN: 956 708#

Meetings are generally the second Wednesday of every month.

✉ [Group email discussion](#)

💬 [Gitter chat channel](#)

📄 [SIG Micro charter](#)

2019-11-13

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

Pete Warden, petewarden@google.com, Google

Michael Gielda, mgieldda@antmicro.com, Antmicro

Stephan Uphoff, ups@stups.com

Neil Tan, neil.tan@arm.com

[Brandon Satrom, brandon@particle.io](mailto:brandon@particle.io), Particle

[Sandeep Mistry, s.mistry@arduino.cc](mailto:s.mistry@arduino.cc), Arduino

Jason Zaman (SIGBuild jason@perfinion.com)

Leigh Johnson, hi@leighjohnson.me, ML GDE, Slack

Items to discuss

Op support roadmap

- 41 supported as reference
- Arm delivered and verified CMSIS-NN optimizations
- Arm also working on Helium optimizations

User Feedback (Neil to Ian Nappier)

- Keen to hear feedback from contributors and users, iannappier@google.com

Step Invoke (Neil to Stephan Uphoff)

(https://docs.google.com/document/d/1hMhTTw5nBp3kRnHSBiQnEHin8crHwV6_4o7U3xRAU6c/edit?usp=sharing)

Pull request process update from petewarden@

Discussion of separate repos and nightly build process

2019-10-09

9 October 2019 / 8:00 AM Pacific Time

[Video recording](#)

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

- Edd Wilder-James, ewj@google.com, TensorFlow Team @ Google
- Pete Warden, petewarden@google.com, TensorFlow Team @ Google
- Neil Tan, neil.tan@arm.com
- [Brandon Satrom](mailto:brandon@particle.io), brandon@particle.io, DevX @ Particle
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Joana Carrasqueira, joanafilipa@google.com, TensorFlow @Google
- Andrew Cavanaugh, andrewc@xmos.com, XMOS Ltd.
- Jens Elofsson, jens.elifsson@arm.com, ARM, Sweden
- Fredrik Knutsson, fredrik.knutsson@arm.com, ARM, Sweden
- Tiezhen WANG, wangtz@google.com, TensorFlow @ Google
- Dominic Symes, dominic.symes@arm.com, Arm, UK
- Niranjan Yadla, nyadla@cadence.com, Cadence Design systems

Items to discuss

- Update on last month's actions
 - Ops support roadmap (Pete)
 - Motion based gesture recognition (Tiezhen)
- Dominic Symes & Fredrik Knutsson: Proposal on combined offline + online memory planner
- `main(args,argv)`, arguments are not passed to `setup()` function as part of `main()` functions
- Hifi4_ops placeholder (`tensorflow/lite/kernels/internal/hifi4_ops`)
- Scratch buffer api to use it for hifi4 optimized ops

Questions on quantization: offline and symmetric quantization status (Neil Tan)

NOTES

2019-09-11

11 September 2019 / 8:00 AM Pacific Time

[Video recording](#)

AGENDA

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

- Edd Wilder-James, ewj@google.com, TensorFlow Team @ Google
- Pete Warden, petewarden@google.com, TensorFlow Team @ Google
- [Chris Hall \(Vivint Smart Home\)](#) Chris.hall@vivint.com
- Neil Tan, neil.tan@arm.com
- Joana Carrasqueira, joanafilipa@google.com, TensorFlow Team @ Google
- Nat Jeffries, njeff@google.com, Tensorflow Team @ Google
- Fernando García-Redondo fernando.garciaredondo@arm.com, Arm Research
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Brad Larson, bradlarson@google.com, TensorFlow Team @ Google
- Pete Blacker Pete.Blacker@surrey.ac.uk Surrey Space Centre
- Morten Opprud opprud@gmail.com Aarhus University
- Fredrik Knutsson fredrik.knutsson@arm.com. Arm Engineering, Sweden

Items to discuss

- Update on last month's actions
 - Ops support roadmap (Pete)
- Neil Tan: Operator test cases (data-driven): Conservative accuracy threshold vs. Operator versioning
- Dominic Symes: Offline memory planning/tflite metadata status
- Fredrik Knutsson: Scratch buffer registration to memory allocator
- Runtime statistic - binary size (Neil)

NOTES

- Runtime memory allocator checked in (making TFL Micro deployable on smaller platforms and enabling use-cases)
-

ITEMS FOR NEXT MONTH

- items

2019-08-14

14 August 2019 / 8:00 AM Pacific Time

DISCUSSION LEADERS

Neil Tan, Pete Warden

AGENDA

Housekeeping

As with all TensorFlow forums, we believe in an open and welcoming environment, so please review and abide by our [code of conduct](#).

This meeting and notes are public. Feel free to ask questions, and let us know who you are when you speak.

Roll call

So we can understand who is here and how to get in touch, please leave your name, company and email address here. This is not mandatory.

- Edd Wilder-James, ewj@google.com, TensorFlow Team @ Google
- Brandon Satrom, brandon@particle.io, Particle
- Jason Zama (SIG-Build) <jason@perfinion.com>
- Mehdi Amini <aminim.google.com>, MLIR @ Google
- Jan Jongboom <janjongboom@gmail.com>
- Shashikant Ghangare (DRDO) , shashikant.ghangare@outlook.in
- Margaret Maynard-Reid (ML GDE), margaretmz@gmail.com
- Carl Turner (independent contractor), carl@turner.ca
- Morten Opprud (Teaching Embedded Systems, Aarhus University, DK)
opprud@gmail.com

- [Chris Hall \(Vivint Smart Home\) Chris.hall@vivint.com](#)
- Ian Nappier, iannappier@google.com, TensorFlow Team @ Google
- Brad Larson, bradlarson@google.com, Swift for TensorFlow Team @ Google
- Yi-Lin Tung, yilintung@on-device-ai.com , On-Device AI Co., Ltd.
- Zhe He, zhe@au-zone.com, Au-Zone Technologies
- Joe Antognini, joe@whisper.ai, Whisper AI
- Sandeep Mistry (Arduino), s.mistry@arduino.cc
- Fernando García Redondo <fernando.garciaredondo@arm.com> (Arm)
- Anuraj R, anuraj.rp@gmail.com, TactoTek Group

Introduction to SIG

1. What is it, and what will it be doing?

Roadmap

- What ops are we prioritizing for porting, and who's doing the work?
- What ops are we prioritizing for optimization, and who's responsible?
- What quantization scheme are we using?

NOTES

- Dominic: Q: how do we deal with decisions around operations and testing, governance.
- Pete: Still an open question how we test operations. We're looking to discover consensus from these meetings and the community along the right way to go forward. We want agreement across the stakeholders.

On the question of operator accuracy, my perspective is “does the operator when placed in a full model hurt the accuracy in a significant way?”. If it doesn't hurt the accuracy, it's the job of the unit test owner to loosen the tests so they pass. Unit tests are guard rails, but what we really care about is overall model accuracy.

- Dominic (ARM): would agree with that. Follow on Q: what sort of example models would then be used to test those results? Accuracy requirements can depend on the depth of the model.

- Pete: we should have models checked into the repo which will be the ones we want to use: get more models in the examples section. We might need to expand that out to a model zoo.
- Jason Zaman:
 - https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/experimental/micro/kernels/all_ops_resolver.cc most (all?) of the ops are listed here too if you want to follow the repo
- Pete Blacker: I've got an MNIST example ready do go
- Fredrik (ARM): want an app that would let us test multiple models without adding extra examples
- Pete Warden: challenge is pre-processing, and also the metrics. For things like classification, we typically care about the top-1 class or the order of the outputs, not so much the exact score.
- Sebastian (SI labs): accuracy is one thing. Will we look at other metrics, around performance for example, within some envelope of accuracy or power consumption.
 - Pete Warden: would love to see it, don't know how yet to implement that. We don't have the resources or expertise to create a device farm. Looked at ??? for functional testing, but it doesn't work for latency. Does the group have other ways to do this?
- Sebastian: number of ways to solve that. We need to drive optimized systems. You can measure the final application but that's not as useful. There are probably ways to instrument models so for a particular platform you can assess performance. It might end up being individual vendors have to do this, but the SIG can potentially define the envelope.
 - Pete W: at CVPR this year we ran a contest to see who can make the most accurate model on person/not-person in image classification within 250kb of space. This was a one-off thing.
- Sebastian: Maybe there could be tiers and profiles of power constraints vs memory constraints etc. The SIG could try and define these.

- Pete Blacker: how mature is memory allocation and management? I've been trying to optimize this. This is a core functionality that might not be easy to do later.
 - Pete Warden: I've got a pending change which does a much better job at allocating memory. Hoping to get this in for review this week or early next week. It is a big problem currently, we are using way too much SRAM than we should be.
 - Neil Tan: ARM is also working on memory issues. Currently trying a lot of brute force methods to find a good layout
- Neil Tan: can we tweak the flatbuffer schema for experimentation? Perhaps to add hints about memory allocation for the interpreter.
 - Pete: There is a metadata field that we can use for a lot of these extra pieces of information. Also hoping to put in better tools at the python level to mess around with the flatbuffer format. Would like to be able to swizzle weight formats. Or replace ops with special custom ops. And use the file format in creative ways. We're not like android where you don't know what platform you're deploying to. We generally know what we're deploying to or have custom modifications to the interpreter, so having the serialization format hackable would help.
 - Neil: There is relevant work in the uTensor group. TFLite importer/exporter, should help with experimentation and help migration and an example for people.
 - Dominic Symes: I'd like to agree with memory comments and controlling from the flatbuffer.
 - Pete: We've tried to use the current allocator from TFLite Mobile but its hard to debug.
- Dominic Symes: Question about relationship between "TFLite" and "TFLite Micro".
 - Pete: What we're trying internally, is reduce the code duplication. Initially we're adding to the duplication since we have op boilerplate code is different and even worse is the

testing code. Experimenting with providing APIs for ops to use (eg tflite ops call malloc a lot, so can't be used for tfliteMicro) so ops can be unified. Hope is to make TFLite modular enough so TFLiteMicro is the same code (just maybe less of it).

- John Brothers (ARM): There could still be other kinds of ops, for some there are power optimizations that would matter.
- Jason Zaman: Having more things shared with TFLite would help the testing coverage too. And I could see that even on mobile it might be useful to be able to choose low power implementations instead of faster ops, so those kinds of ops could still be shared.
- Pete Warden: rather than emailing me direct, use the group we have: micro@tensorflow.org
- Neil Tan: as we progress, we'll add more examples into the test suite. We'll need to relax or tighten the accuracy for operators, depending on models. How will we deal with this? As multiple versions?
 - Pete: would prefer to move in the direction of loosening accuracy. Otherwise we'd have previously good ops that no longer work and if we need to tighten accuracy that would be a nasty migration.
 - Jason: use op versioning
 - Mehdi: not sure that's a good practice for issues of precision
 - John Brothers: sounds like we'll get a lot of clutter. There are some examples of developing an API for hw against capabilities... Other cases where this has been done.
 - Mehdi / Jason conversation: New ops for differing precision would be complicated, eg. would the op still allocate the full word but return lower precision? What do you do with the weights, etc?
- Thanks everyone! Hope everyone knows about the SIG-Micro mailing list and the gitter channel: <https://gitter.im/tensorflow/sig-micro>

ACTION ITEMS

1. Doc somewhere that lists the currently supported/not-supported ops and the roadmap for porting new ops AI: petewarden@google.com
2. Figure out gitter and list it in the doc. DONE.

NEXT MONTH'S AGENDA

1. Operator test cases (data-driven): Conservative accuracy threshold vs. Operator versioning