Chair of Real-Time Computer Systems
Department of Electrical and Computer Engineering
Technical University of Munich

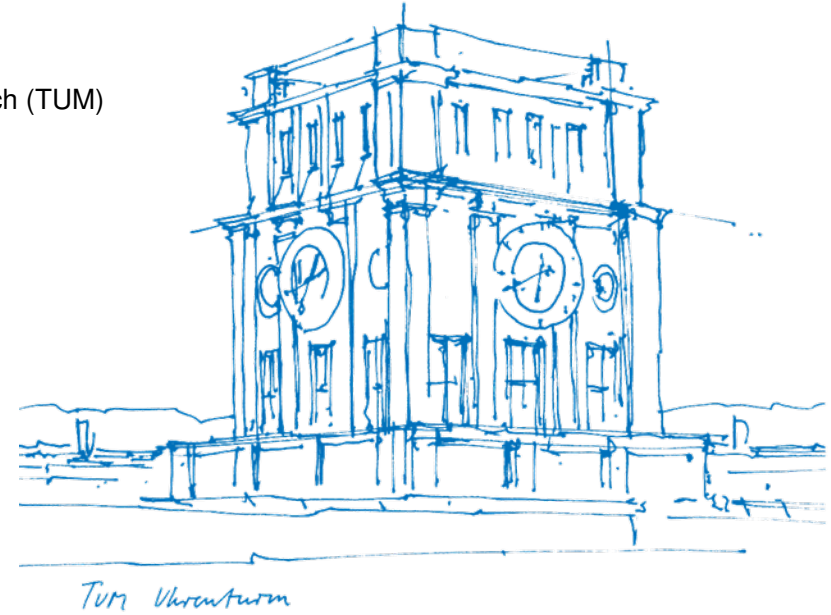
# Research Internship with RCS@TUM

## Implementation of tiny machine learning models on microcontrollers

**Philipp van Kempen**

Department of Electrical and Computer Engineering, Technical University of Munich (TUM)

**24.08.2020 - 25.10.2020**



# Motivation

- Machine learning in the past vs. today
- Compute-intensive algorithms
- State of the Art: ML/AI on smartphones
- Research topic: ML/AI on microcontrollers

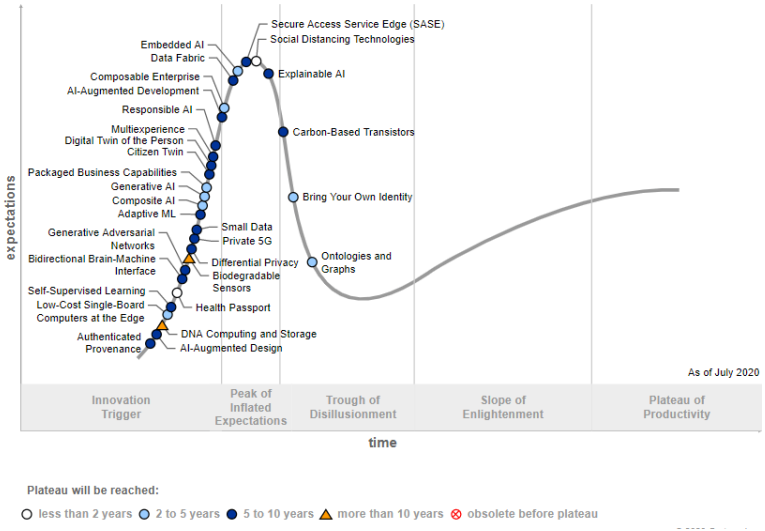
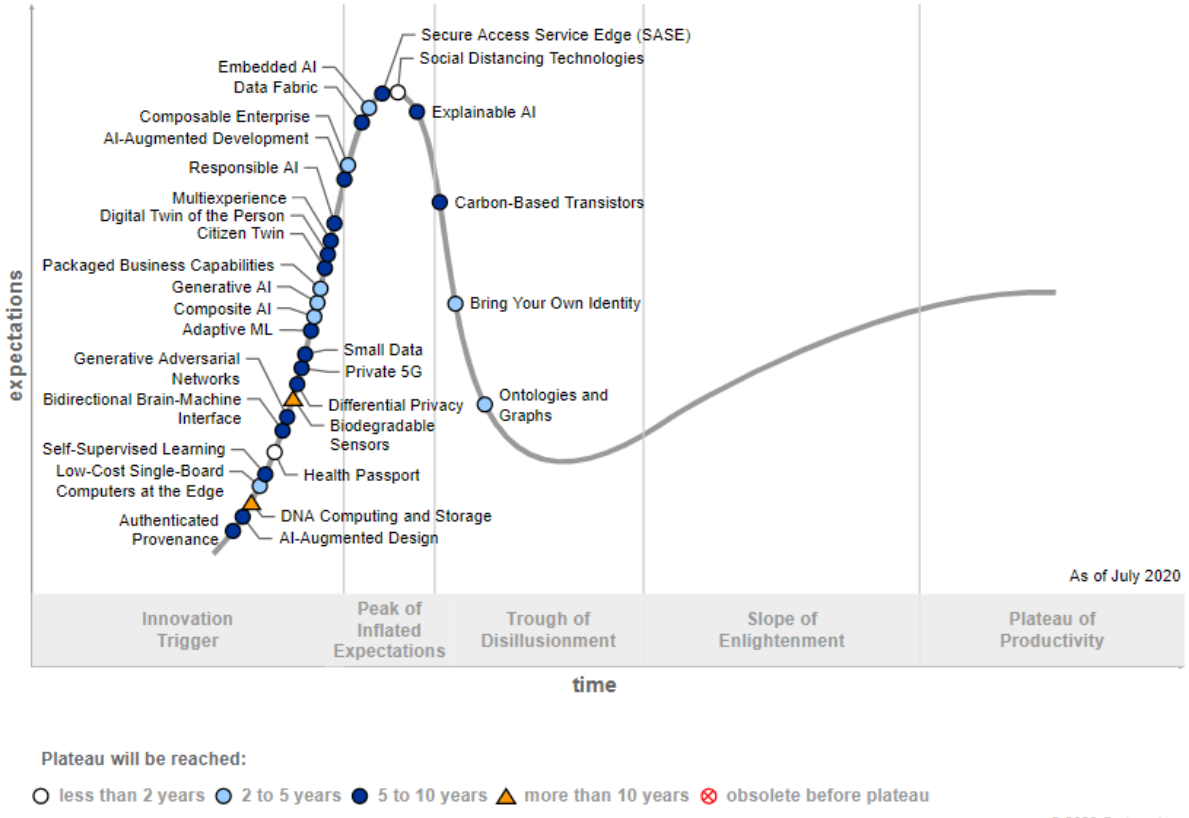



Figure: 2020 Gartner Hype Cycle[1]

[1] https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/


Secure Access Service Edge (SASE)
Social Distancing Technologies
Embedded AI
Data Fabric
Explainable AI
Composable Enterprise
AI-Augmented Development
Responsible AI
Carbon-Based Transistors
Multiexperience
Digital Twin of the Person
Citizen Twin
Packaged Business Capabilities
Generative AI
Bring Your Own Identity
Composite AI
Adaptive ML
Small Data
Private 5G
Generative Adversarial Networks
Bidirectional Brain-Machine Interface
Differential Privacy
Ontologies and Graphs
Biodegradable Sensors
Self-Supervised Learning
Low-Cost Single-Board Computers at the Edge
Health Passport
Authenticated Provenance
DNA Computing and Storage
AI-Augmented Design
expectations
As of July 2020
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

# Goals

- 9 weeks
- Support Electronic System Level (ESL) research group (EDA+RCS)
- Implement reference implementations of TinyML models on STM32 Hardware
- Work based on previous attempts by Alex Hoffman
- Extend toolchain with new features and documentation
- Summarize results at the end of the internship

# Steps

**Major topics:**

- Reading books and web pages
- Toolchain setup and extension
- Training of examples
- Implementation of examples
- Documentation and handover

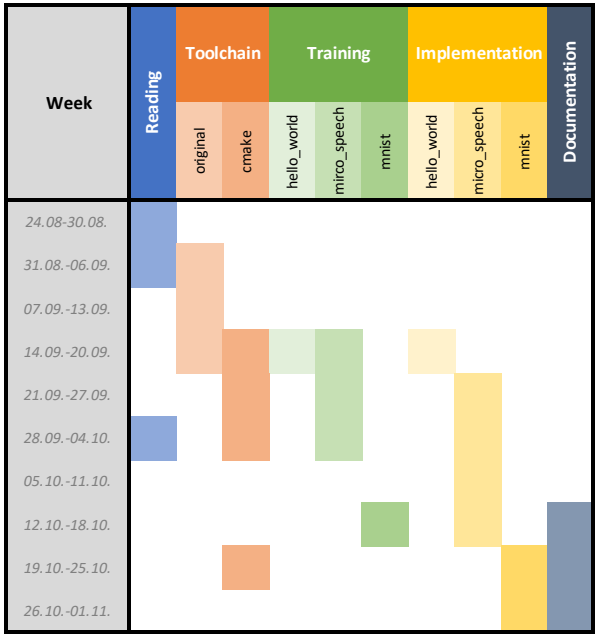



Figure: Schedule

# Reading

- Book by Pete Warden and Daniel Situnayake
- Introduction to Tensorflow Lite for Microcontrollers (TFLM) framework
- Referencing examples located in the Tensorflow source tree
  - **Hello World**
  - **Micro Speech**
- Official documentation of Tensorflow (v1.5 and latest)

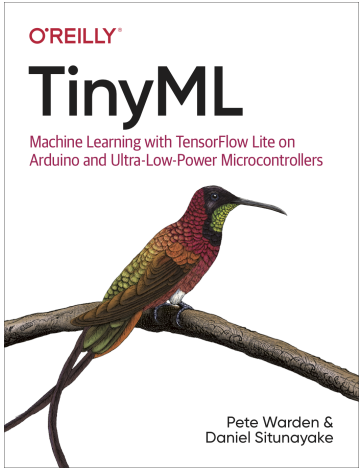



Figure: TinyML Book[2]

[2] https://tinymlbook.com

# Toolchain

- Original ARM mBed toolchain used as a reference
- CMake based - Initially developed by Konstantin Oblaukhov [3], Improved and extended by Alex Hoffman
- Reference Application: `STM3240G-EVAL-TensorFlow-MNIST`
- Some workarounds required to fix upstream bugs

(a) STM32F413H-DISCOVERY

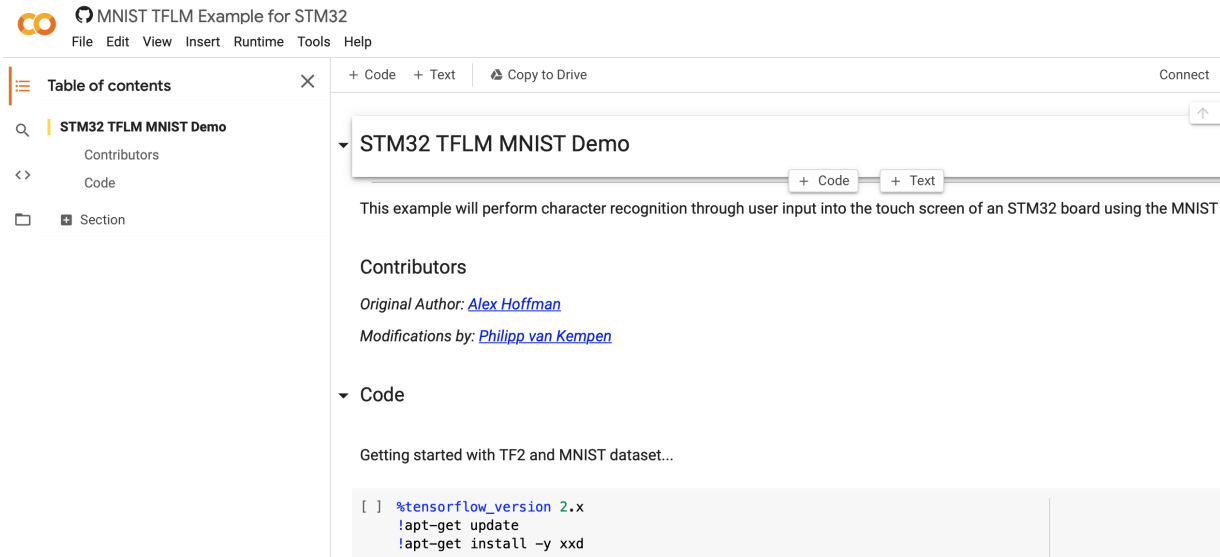
(b) STM32F769I-DISCOVERY

Figure: Target Boards[4]

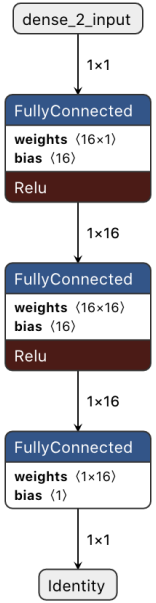
[3] `https://github.com/ObKo/stm32-cmake`
[4] Images taken from `https://www.st.com`

# Training

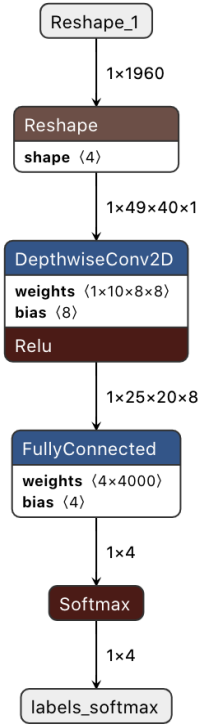
- Training scripts: Interactive notebooks hosted on Google Colaboratory[5]
- Deployment on Microcontrollers:
  - Quantization
  - Optimizations
  - Conversion to TFLite
  - Interpreter vs. Compiled offline model



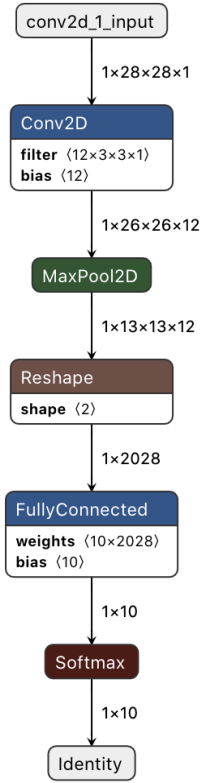

[5] https://colab.research.google.com




(a) Hello World




(b) Micro Speech

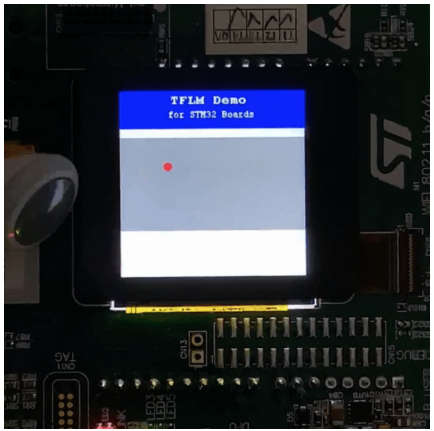



(c) MNIST

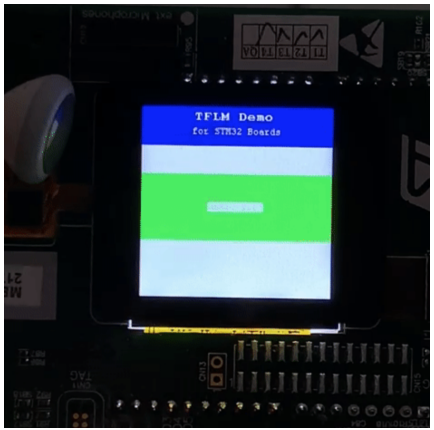
Figure: TFLite Model Graphs per Example

# Implementation

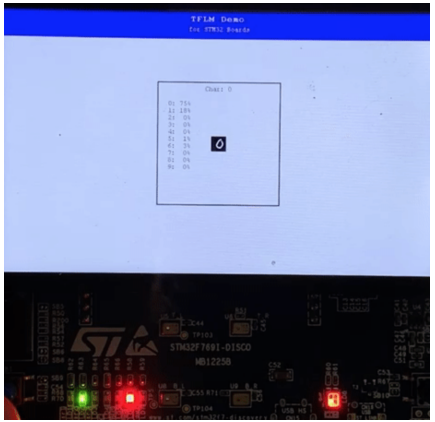
- **Hello World:** Ran almost out of the box
- **Micro Speech:**
  - Hardest challenge: Streaming real-time audio
  - Boards to slow? $\Rightarrow$ Solution: Enable CMSIS-NN
  - Further Tuning of parameters was required to detect any voice commands $\Rightarrow$ More false positives!
- **MNIST:**
  - Dataset: ten-thousands of handwritten digits
  - Quantization issue: Latest TFLM does not support unsigned uint8 inputs $\Rightarrow$ transform to $\{-128, \ldots, 127\}$
  - Challenge: Transform touchscreen input into 28x28 pixel grayscale images
  - Bug in Alex's implementation: Inverted colors
  - Some digits can not be detected at all $\Rightarrow$ Improve network architecture
- Common:
  - Added Benchmarking module
  - Added Memory Reporting support
  - Added possibility to choose TFLM interpreter or compiler
  - For easier Testing: SD-Card support to feed real samples to the network




(a) Hello World




(b) Micro Speech



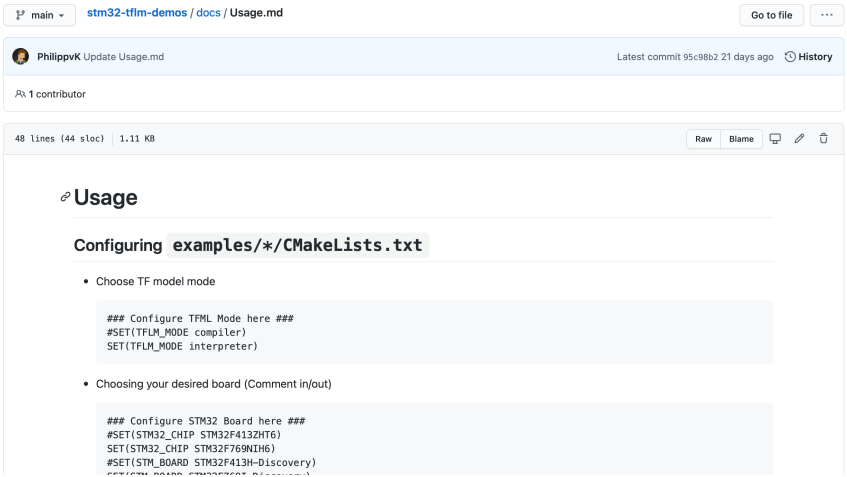

(c) MNIST

Figure: Examples running on the STM32 Boards[6]

[6] See `https://github.com/PhilippvK/stm32-tflm-demos` for GIFs!

# Documentation

- Common toolchain parts as submodules ⇒ less redundant code
- Wrapper repository[7]
- Common documentation at a single place
- Great Extend-ability




[7] https://github.com/PhilippvK/stm32-tflm-demos

# Evaluation

See report PDF or Documentation[8] for details! Only showing tables here...

| | **Boards** | | |
|---|---|---|---|
| **Metrics** | STM32F413HDISCOVERY | STM32F769IDISCOVERY | **Units** |
| Clock Frequency | 100 | 216 | *MHz* |
| Special Features | - | Double Issue, I/D-Cache | - |
| Flash Memory | 1.5 | 2 | *MB* |
| SRAM Memory | 256 | 512 | *kB* |

Figure: Board Metrics

[8] `https://github.com/PhilippvK/stm32-tflm-demos/blob/main/docs/Metrics.md`

# Evaluation - Memory Usage & Number of Ops

| | **Examples** | | | |
|---|---|---|---|---|
| **Type** | hello_world | mirco_speech | mnist | **Units** |
| Model Size (FLASH) | 2 | 18 | 23 | *kB* |
| TensorArena Size (SRAM) | 1 | 7 | 11 | *kB* |

Figure: Memory Usage (approx.)

| **Examples** | | |
|---|---|---|
| hello_world | mirco_speech | mnist |
| 41 *FLOPS* | 689980 *FLOPS* | 202810 *FLOPS* |

Figure: Number of Ops (after quant.)

# Evaluation - Runtime Measurements & CMSIS-NN

| | | Examples | | | |
|---|---|---|---|---|---|
| **Section** | **CPU** | hello_world | mirco_speech | mnist | **Units** |
| Populate | F4 | ∼0 | 38 | 132 | *ms* |
| | F7 | ∼0 | 11 | 88 | |
| Invoke | F4 | ∼0 | 49 | 34 | *ms* |
| | F7 | ∼0 | 52 | 13 | |
| Respond | F4 | ∼0 | ∼0 | 125 | *ms* |
| | F7 | ∼0 | ∼0 | 93 | |

Figure: Runtime Measurements (approx.)

| | | Examples | | | |
|---|---|---|---|---|---|
| **Settings** | | hello_world | mirco_speech | mnist | **Units** |
| CMSIS_NN | OFF | ∼1 | 413 (unusable!) | 52 | *ms* |
| | ON | ∼0 | 52 | 13 | *ms* |
| Difference | | (?) | (-87) | (-75) | *%* |

Figure: CMSIS-NN Improvements (approx.)

# Conclusions and Outlook

- Main goal fulfilled
- MNIST was optional but also implemented successfully
- Models need more tuning
- Demonstrated capabilities of TinyML on microcontroller platforms

[9] https://github.com/munober/thesis/blob/master/digital_edition.pdf

# Conclusions and Outlook

- Main goal fulfilled
- MNIST was optional but also implemented successfully
- Models need more tuning
- Demonstrated capabilities of TinyML on microcontroller platforms

**Future Work:**

- Support USB-Storage as alternative to SD-Card
- Add TinyFace Model[9]
- Merge Deployment and Evaluation Flow with EDA RISCV-toolchain
- Enable usage of FreeRTOS instead of baremetal
- Add possibility to parse results via UART
- Feed samples via UART for automated testing

  $\Rightarrow$ Extract data on the accuracy of the models

[9] `https://github.com/munober/thesis/blob/master/digital_edition.pdf`