“AutoFlow - an open source Framework to automatically implement neural networks on embedded devices”

Daniel Konegen - Hahn-Schickard
Marcus Rüb - Hahn-Schickard

April 5, 2022
tinyML Talks Strategic Partners

Additional Sponsorships available – contact Olga@tinyML.org for info
Executive Strategic Partners
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
Advancing AI research to make efficient AI ubiquitous

- **Power efficiency**
  - Model design, compression, quantization, algorithms, efficient hardware, software tool

- **Personalization**
  - Continuous learning, contextual, always-on, privacy-preserved, distributed learning

- **Efficient learning**
  - Robust learning through minimal data, unsupervised learning, on-device learning

- **Perception**
  - Object detection, speech recognition, contextual fusion

- **Reasoning**
  - Scene understanding, language understanding, behavior prediction

- **Action**
  - Reinforcement learning for decision making

A platform to scale AI across the industry
SYNTIANT

Neural Decision Processors
- At-Memory Compute
- Sustained High MAC Utilization
- Native Neural Network Processing

ML Training Pipeline
- Enables Production Quality Deep Learning Deployments

Data Platform
- Reduces Data Collection Time and Cost
- Increases Model Performance

End-to-End Deep Learning Solutions for TinyML & Edge AI

partners@syntiant.com
www.syntiant.com
Platinum Strategic Partners
Fastest Video Analytics Solutions on Arm CPUs
KLIKA·TECH
GLOBAL IOT SOLUTIONS
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
BROAD AND SCALABLE EDGE COMPUTING PORTFOLIO

Microcontrollers & Microprocessors

**Arm® Core**
- Arm® Cortex®-M 32-bit MCUs
  - Arm ecosystem, Advanced security, Intelligent IoT
- Arm®-based High-end 32 & 64-bit MPUs
  - High-resolution HMI, Industrial network & real-time control
- Arm® Cortex®-M0+ Ultra-low Power 32-bit MCUs
  - Innovative process tech (SOTB), Energy harvesting

**Renesas Core**
- Ultra-low Energy 8 & 16-bit MCUs
  - Bluetooth® Low Energy, SubGHz, LoRa®-based Solutions
- High Power Efficiently 32-bit MCUs
  - Motor control, Capacitive touch, Functional safety, GUI
- 40nm/28nm process Automotive 32-bit MCUs
  - Rich functional safety and embedded security features

Renesas Synergy™
- Arm®-based 32-bit MCUs for Qualified Platform
  - Qualified software and tools

Core technologies

**AI**
- A broad set of high-power and energy-efficient embedded processors

**Security & Safety**
- Comprehensive technology and support that meet the industry’s stringent standards

**Digital & Analog & Power Solution**
- Winning Combinations that combine our complementary product portfolios

**Cloud Native**
- Cross-platforms working with partners in different verticals and organizations

© 2021 Renesas Electronics Corporation. All rights reserved.
Gold Strategic Partners
Maxim Integrated: Enabling Edge Intelligence

**Advanced AI Acceleration IC**

The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.

[www.maximintegrated.com/MAX78000](http://www.maximintegrated.com/MAX78000)

---

**Low Power Cortex M4 Micros**

Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

[www.maximintegrated.com/microcontrollers](http://www.maximintegrated.com/microcontrollers)

---

**Sensors and Signal Conditioning**

Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

[www.maximintegrated.com/sensors](http://www.maximintegrated.com/sensors)
LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com
seeed
The IoT Hardware Enabler
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com
SynSense builds sensing and inference hardware for ultra-low-power (sub-mW) embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
Silver Strategic Partners
The Best Product of the Year and the Best Innovation of the Year awards are open for nominations between November 15 and March 14.

More sponsorships are available: sponsorships@tinyML.org
Our next tinyML Trailblazers Series
Success Stories with Eric Pan
(Founder, Seeed Studio and Chaihuo Makerspace)

LIVE ONLINE April 6th, 2022 at 8 am PST

Register now!
Join Growing tinyML Communities:

tinyML - Enabling ultra-low Power ML at the Edge

The tinyML Community
https://www.linkedin.com/groups/13694488/

8.8k members in
45 Groups in 35 Countries

2.7k members
&
5.7k followers
Subscribe to tinyML YouTube Channel for updates and notifications
(including this video)
www.youtube.com/tinyML
Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thursday, April 7</td>
<td>Łukasz Szelejewski&lt;br&gt;CTO, Kontakt.io</td>
<td>Battery optimized people counting using FIR and AI</td>
</tr>
<tr>
<td></td>
<td>Dawid Crivelli&lt;br&gt;R&amp;D Lead Engineer, Kontakt.io</td>
<td></td>
</tr>
</tbody>
</table>

Webcast start time is 7:00 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting
Reminders

Slides & Videos will be posted tomorrow

Please use the Q&A window for your questions

tinyml.org/forums  youtube.com/tinyml
Daniel Konegen studied Mechanical Engineering and Mechatronics (B.Sc.) and Mechatronic Systems (M.Sc.) at Furtwangen University. In his master's thesis, he worked on the automated implementation of neural networks on embedded systems. After completing his Master's degree in 2020, he worked at the Karlsruhe Institute of Technology at the Institute of Telematics from October 2020 to August 2021. Since September 2021, he has been responsible for the areas of embedded AI and data science at Hahn-Schickard.
Marcus Rüb studied electrical engineering (B.Sc.) and mechatronic systems (M.Sc.) at Furtwangen University from 2015 to 2020. Since 2018, he has been employed at Hahn-Schickard as a research assistant and conducts research in the field of TinyML. He is currently doing his PhD at the Technical University of Munich.
AutoFlow
Bring your AI to the Edge

Marcus Rüb + Daniel Konegen,
Hahn-Schickard-Gesellschaft für angewandte Forschung e.V.
What is AutoFlow and why do we need it?
How does AutoFlow work?
Which features are included in AutoFlow?
Demonstration
How can you use it?
How can you take part in the development?
What is AutoFlow and why do we need it?

- TensorFlow models usually cannot be run directly on embedded devices like MCUs or FPGAs
- Some manual steps to execute them are necessary
- Data scientists need knowledge in the areas of:
  - Machine learning
  - Embedded systems
  - Data understanding
What is AutoFlow and why do we need it?

- AutoFlow automates this process → Easy and fast use of TinyML
- Cover entire workflow of a data scientist, from creating a ML model to optimizing and implementing the models on the target platform

 ![Diagram of the AutoFlow workflow](image)

- Train model
- Select target
- Optimization
- Conversion
- Compile model

Marcus Rüb & Daniel Konegen - Tuesday, April 5, 2022 - AutoFlow
How does AutoFlow work?

AutoFlow contains two parts:

- Optimization and implementation of NNs for different target platforms
- Automized generation of neural networks for embedded devices
How does AutoFlow work?

Automatized generation of neural networks for embedded devices:

1. Input the data for which a NN should be trained
2. Selection of task to be solved by the model
3. Train different NNs automatically
4. Save best model

Image classification/regression
Data classification/regression

Knowledge in machine learning + underlying data

AutoKeras

Marcus Rüb & Daniel Konegen - Tuesday, April 5, 2022 - AutoFlow
How does AutoFlow work?

Optimization and implementation of neural networks for various target platforms:

Select the model to be used, project name, output path

Choose target platform (MCU, FPGA, SBC)

Select optimization algorithms Pruning and/or Quantization (optional)

Generate files for the selected target platform

Optimize the model using the implemented algorithm(s)

Pass training data if at least one optimization algorithm is selected
Optimization algorithm pruning

Before Pruning

After unstructured Pruning

Pruned weights

Pruned neurons

After structured Pruning

Marcus Rüb & Daniel Konegen - Tuesday, April 5, 2022 - AutoFlow
**Two versions implemented:**

**Factor**
- Define prune factor for dense and conv layers
- Remove neurons and filters
- Retrain model

**Accuracy**
- Define min accuracy or max loss of accuracy
- Remove neurons and filters (start factors of 5%)
- Retrain model

*If accuracy reached ➔ increase factors*
Quantization reduces the number of bits needed to represent the value of the weights → Reduction of memory space

Weights of neural networks are represented as 32-bit float values by default
Optimization algorithm quantization

Two versions implemented:

**int8 + float32**

- Model input: 32-bit-Float
- Model weights: 8-bit-Integer
- Model output: 32-bit-Float

**int8 only**

- Model input: 8-bit-Integer
- Model weights: 8-bit-Integer
- Model output: 8-bit-Integer
Which features are included in AutoFlow?

- AutoFlow is based on TensorFlow
- Automatized generation, optimization and conversion is only possible with TensorFlow models
- The main features of AutoFlow are:
  - Graphical user interface - AutoFlow is implemented with a graphical user interface to make it easy to get started with TinyML.
  - AutoML - Using AutoML techniques, ML models can be generated without the need for ML experience.
  - Automatic compression - To reduce the size of ML models and their execution time, the optimization algorithms pruning and quantization can be applied.
  - Automatic code generation - Code is automatically generated for the selected target platform.
  - Various target platforms - ML models can be executed on MCU, FPGA, Raspberry Pi, ...
Demonstration

AutoFlow

Load or train a model?

- Train a new model
- Load a trained model

Graphical user interface: With this framework we offer a GUI, which should facilitate the entry into the AI as far as possible.

AutoML: With the help of AUTOML techniques, ML models can be generated automatically. No experience with

Marcus Rüb & Daniel Konegen - Tuesday, April 5, 2022 - AutoFlow
How can you use it?

1. **Download tool from GitHub**
   - `git clone https://github.com/Hahn-Schickard/AutoFlow`

2. **Install required libraries**
   - `pip install -r requirements.txt`

3. **Customize AutoKeras**
   - `python src/automl/customize_autokeras.py \ C:/Users/.../Anaconda3/envs/AutoFlow`
How can you take part in the development?

- As mentioned before, AutoFlow is an open source tool and can be found on GitHub [https://github.com/Hahn-Schickard/AutoFlow](https://github.com/Hahn-Schickard/AutoFlow)

- New issues can be opened, e.g. with the following content:
  - What are new features which should be implemented?
  - What should be changed?
  - What is not working?
  - ...

- You are also welcome to implement new functions and features by yourself and to create a pull request → If this fits our vision of AutoFlow, we will be happy to implement your changes/new features.
Are you a german SME and interested in AI?

Free AI offers for SMEs

Our offer for small and medium-sized enterprises (SMEs):
- On-site visits to develop suitable fields of application for AI.
- AI trainings
- AI events
- AI development projects for your application
Thank you for your attention!
Copyright Notice

This multimedia file is copyright © 2022 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org
Copyright Notice

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

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

tinyML is a registered trademark of the tinyML Foundation.

www.tinylm.org