tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

www.tinyML.org

tinyML EMEA Technical Forum 2021
June 7-10, 2021

A TinyML Platform for On-Device Continual Learning with Quantized Latent Replays

Presented by: *Leonardo Ravaglia*, *Ph.D. student in Data Science and Computation*, University of Bologna, Italy

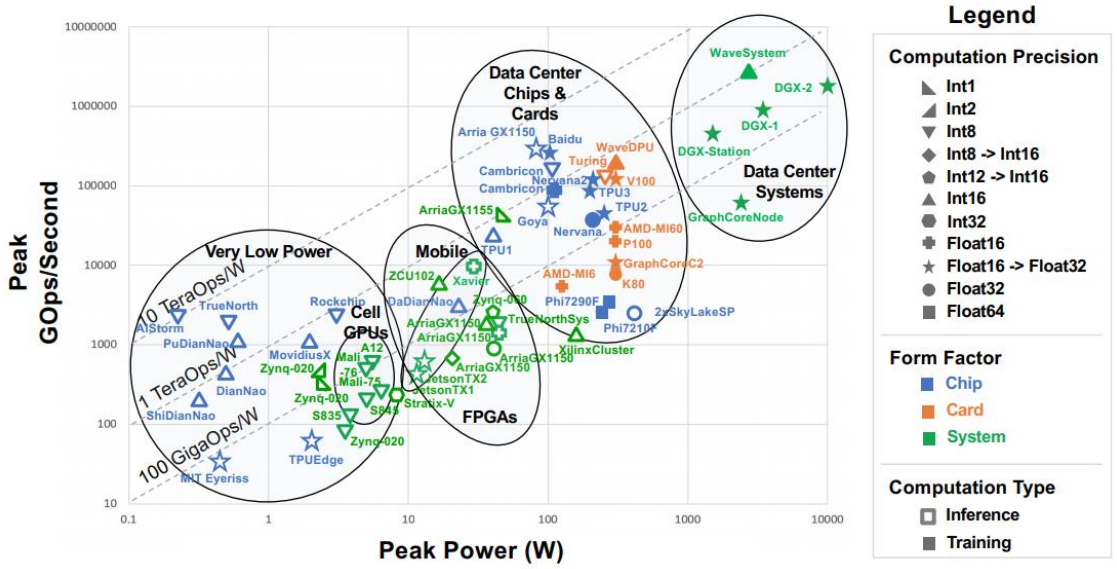
June 4, 2021

tinyML EMEA Technical Forum 2021
June 7-10, 2021

1. **DNN inference** capability has been already demonstrated, but the **training of DNN** models still relies on GPU-based machines.
2. Transfer Learning or incremental training lead to accuracy loss and **Catastrophic Forgetting**.

A way out of this limitations can be found in **Continual Learning** (CL) algorithms:

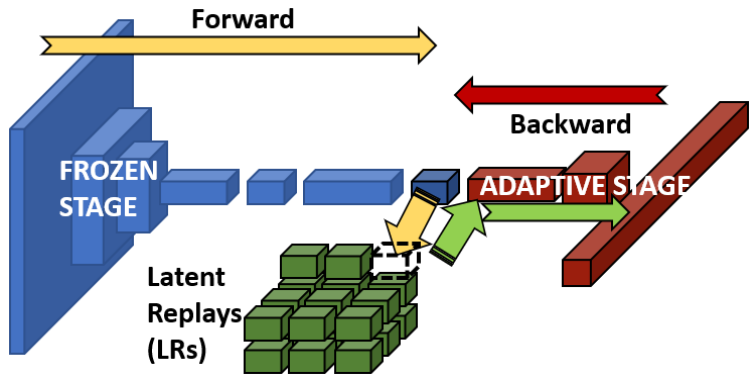
- **CL enables training** also on **TinyML** platforms, with very bounded memory and computational resources.



"Survey and benchmarking of machine learning accelerators", Reuther A. et al.

- **Continual Learning with Latent Replays** is our choice, since offers a solution compatible with embedded devices resources.

tinyML EMEA Technical Forum 2021
June 7-10, 2021

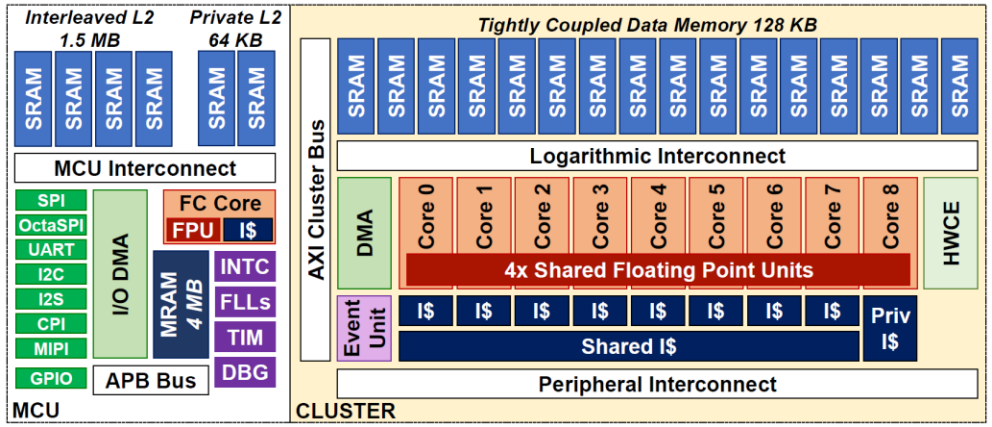


"Latent replay for real-time continual learning", Pellegrini et al.

Contributions

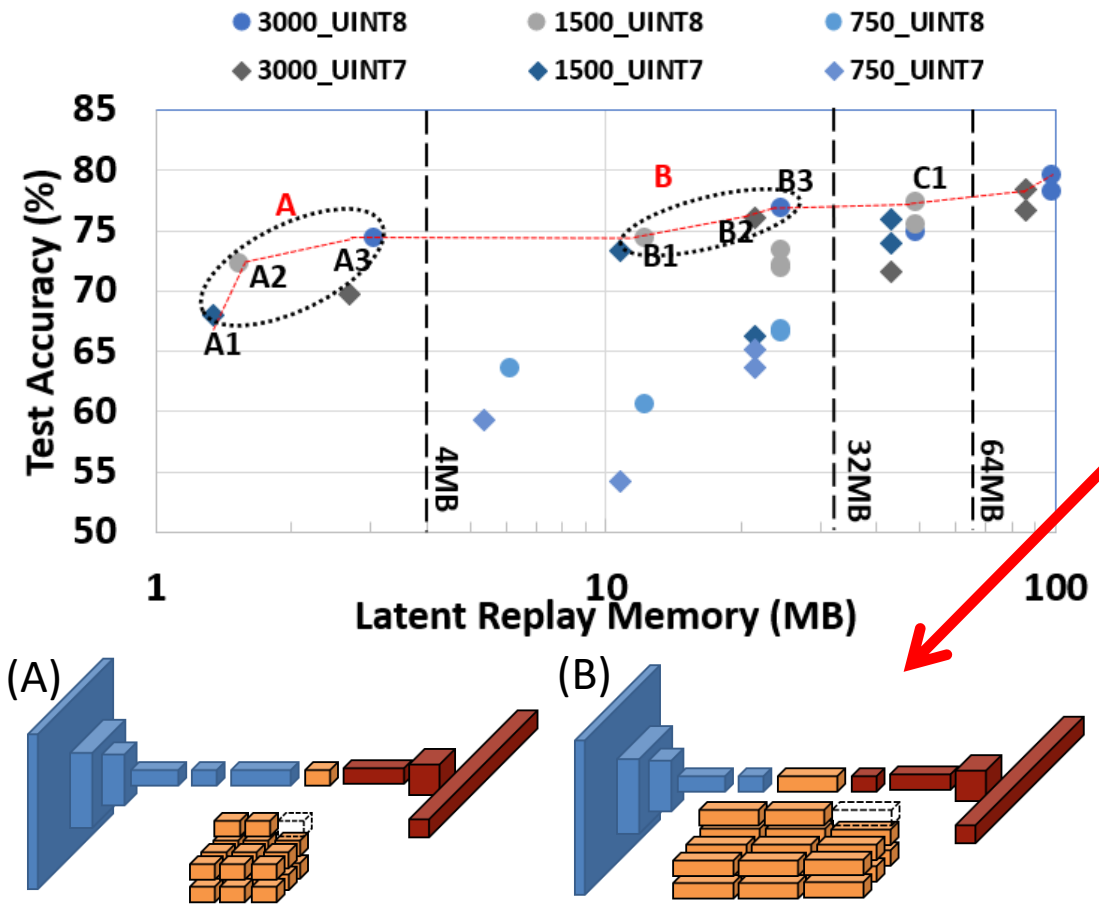
- We **extend the CL with LR algorithm** to work with 8-bit quantized frozen stage.
- We propose a set of CL primitives for forward and backward propagation of common layers: **convolution**, **depth-wise convolution** and **fully connected layers.**

- We fine-tune and optimize the **execution on VEGA**, a TinyML platform for Deep Learning based on **PULP**. We also introduce a **tiling scheme** to manage data movement for the CL primitives.
- We **compare the performance of our CL primitives on VEGA with that on other devices** that could in the future target on-chip at-edge learning (e.g. STM32 L4).



PULP paradigm

tinyML EMEA Technical Forum 2021
June 7-10, 2021



- Memory and Accuracy requirements **vary depending on the LR layer**.
- Quantization of LRs to **INT8 leads to no accuracy loss** with respect to FP32.
- The **LR layer depth** and the **precision** impacts on the accuracy.
- We report the bounds **4, 32 and 64 MB**, which are sizes compatible with embedded devices memories.
- We are able to **retrain on a new class of object**, within few seconds (about 6s).

Journal under review is available at:
https://drive.google.com/drive/folders/1Htabat8Fc_ttMYydm2PIO_f_eCFvjjrR?usp=sharing

TINY ML EMEA
Virtual
JUNE 7-10,2021

Premier Sponsor

Intelligent Agent
Neuton

Automated TinyML

Zero-code SaaS solution

Create tiny models, ready for embedding, in just a few clicks!

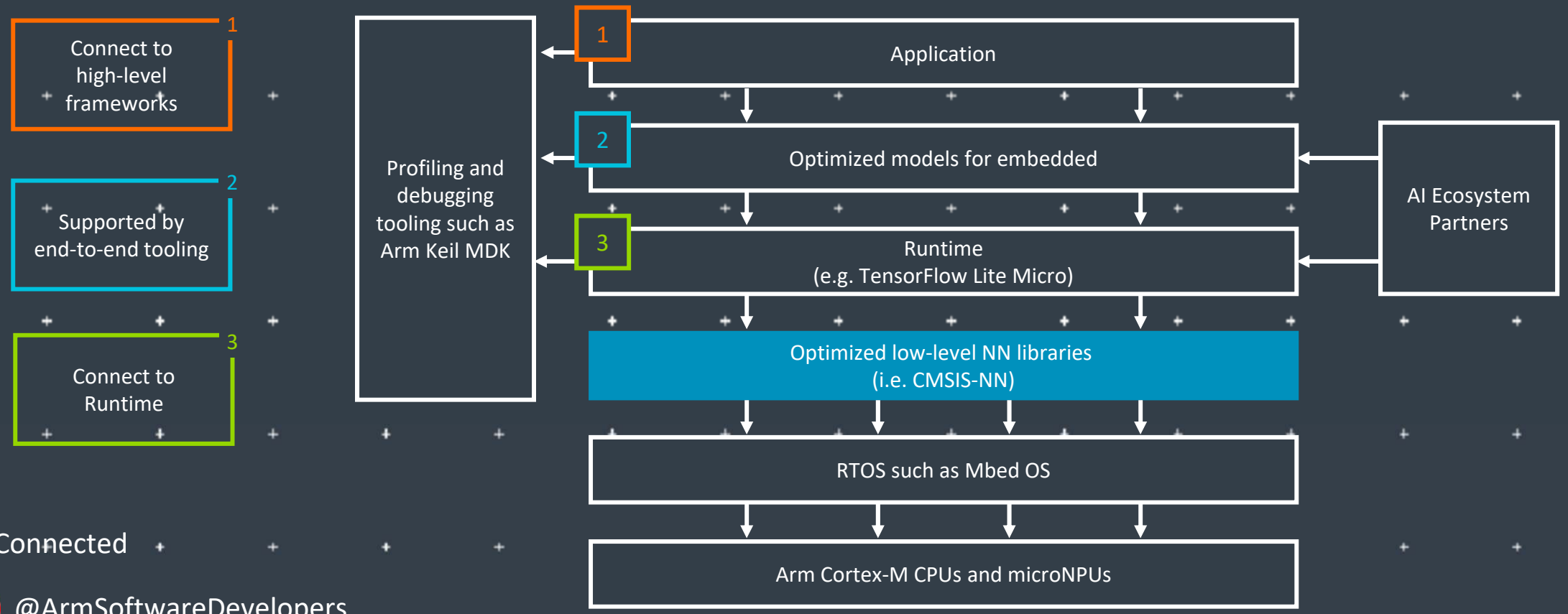
Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Build Fast. Build Once. Never Compromise.

TINY ML EMEA
Virtual
JUNE 7-10,2021

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

© 2020 Arm Limited (or its affiliates)

arm

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

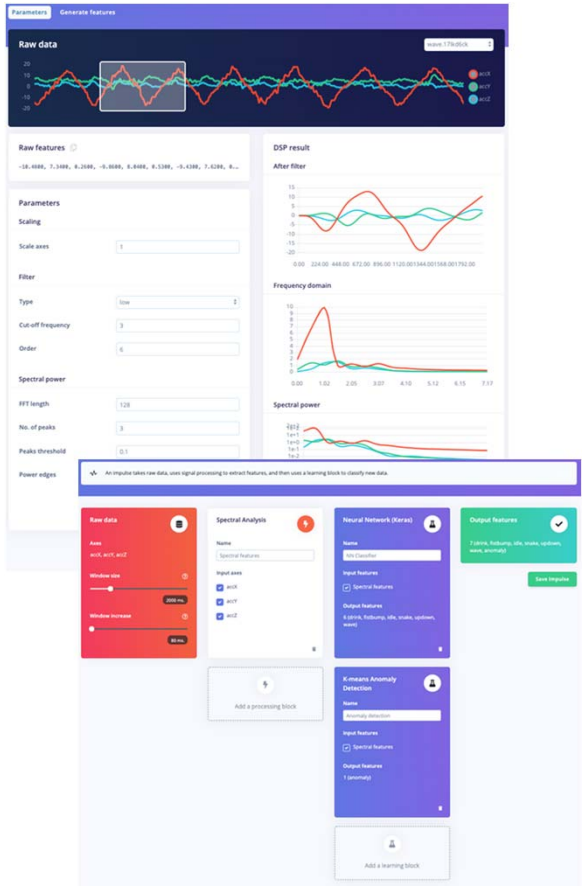
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



www.edgeimpulse.com

Qualcomm
AI research

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

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

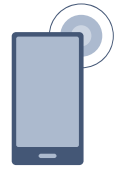
Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive



Mobile

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

SYNTIANT

Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant's chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com

@Syntiantcorp

TINY ML EMEA
Virtual
JUNE 7-10,2021

Platinum Sponsors

Part of your life. Part of tomorrow.

www.infineon.com

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

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

TINY ML EMEA
Virtual
JUNE 7-10,2021

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

Latentai.com

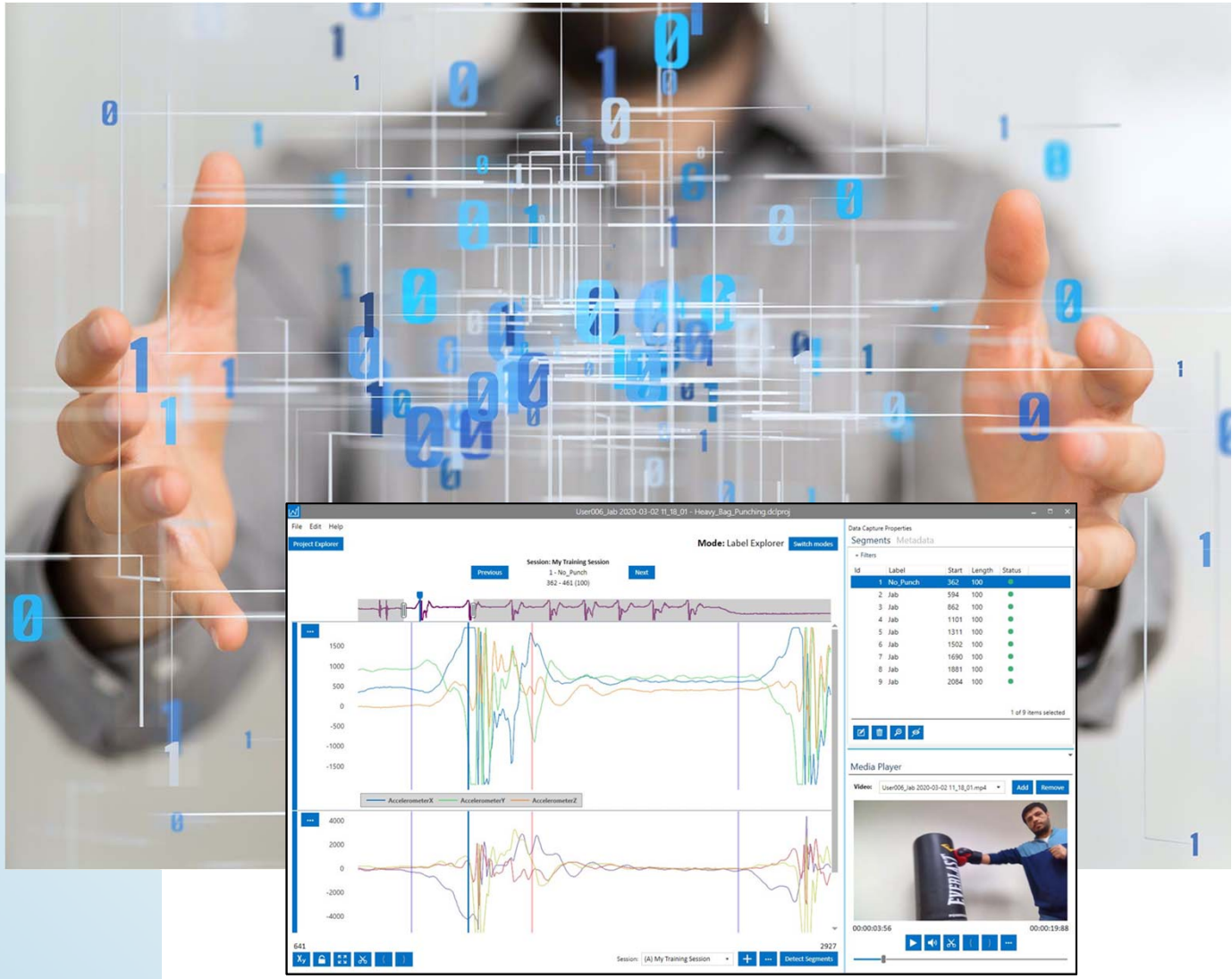
SensiML™

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

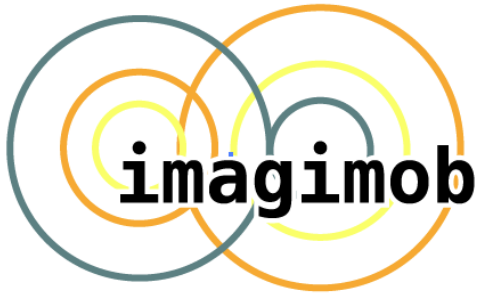
We enable the creation of production-grade smart sensor devices.



sensiml.com

TINY ML EMEA Virtual JUNE 7-10,2021

Silver Sponsors



Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® EMEA Technical Forum 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at tinyML EMEA. 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.tinyML.org