tinyML® EMEA

Enabling Ultra-low Power Machine Learning at the Edge

tinyML EMEA Technical Forum 2021 Proceedings

June 7 – 10, 2021

Virtual Event

TINY ML

www.tinyML.org

tinyML EMEA Technical Forum 2021
June 7-10, 2021

Squeeze-and-Threshold based quantization for Low-Precision Neural Networks

Presented by: Binyi Wu, PhD student, Infineon, Germany

June 7, 2021

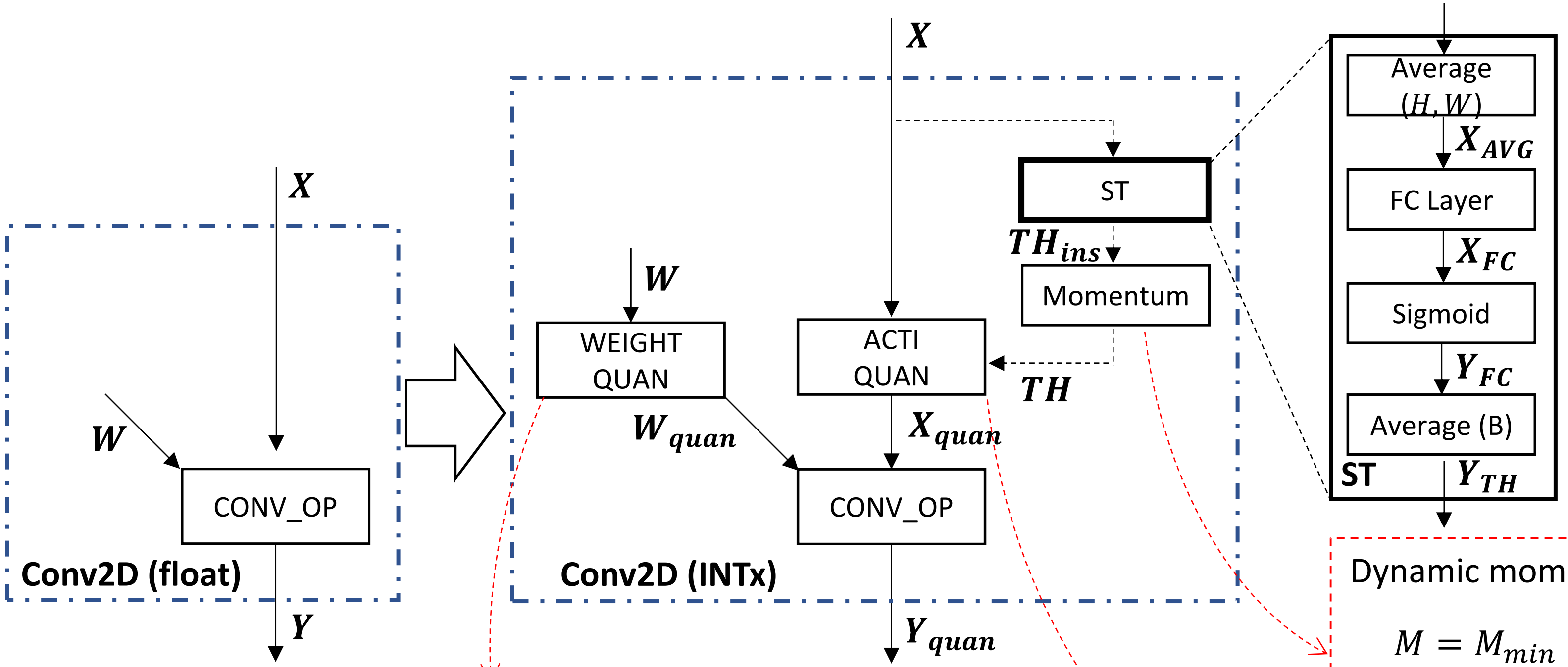
Introduction:

- Convolution neural networks quantization
- low-quality black-and-white photos are well enough to recognize. → Different features should be adjusted to different ranges, and a threshold should be should be learned to distinguish (quantized) them.

Contributions:

- A novel quantization method based on attention mechanism
- A unified 1-bit and multi-bit activation quantization method

Activation Quantization: Squeeze-and-Threshold (ST) quantization



$$\boldsymbol{X_{AVG}} = \frac{1}{HW}\sum_{i=0}^{H}\sum_{j=0}^{W}\boldsymbol{X}[:,:i,j]$$

$$\boldsymbol{X_{FC}} = MatMul(\boldsymbol{X_{AVG}}, \boldsymbol{W})$$

$$\boldsymbol{Y_{FC}} = sigmoid(\boldsymbol{X_{FC}})$$

$$\boldsymbol{Y_{TH}} = \frac{1}{B}\sum_{i=0}^{B}\boldsymbol{Y_{FC}}[i]$$

Dynamic momentum

$$M = M_{min} + (M_{init} - M_{min}) * \cos\left(\frac{E_{cur}}{E_{tol}}\right)$$

$$\boldsymbol{TH} = (1 - M) * \boldsymbol{TH_{last}} + M * \boldsymbol{Y_{TH}}$$

$$\boldsymbol{X_{int}} = clip\left(round\left(\frac{\boldsymbol{X}}{\boldsymbol{TH}}\right), 0, 2^b - 1\right)$$

$$\boldsymbol{X_{quan}} = \boldsymbol{X_{int}} * mean(\boldsymbol{TH})$$

Weight quantization

1-bit weight binarization uses the scheme from BNN+[1].

Multi-bit weight quantization uses the scheme from LSQ[2].

[1] *Sajad Darabi, Mouloud Belbahri, Matthieu Courbariaux, andVahid Partovi Nia. BNN+: improved binary network training*

[2] *Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In 8th International Conference on Learning Representations (ICLR), 2020, Addis Ababa, Ethiopia, April 26-30, 2020*

tinyML EMEA Technical Forum 2021
June 7-10, 2021

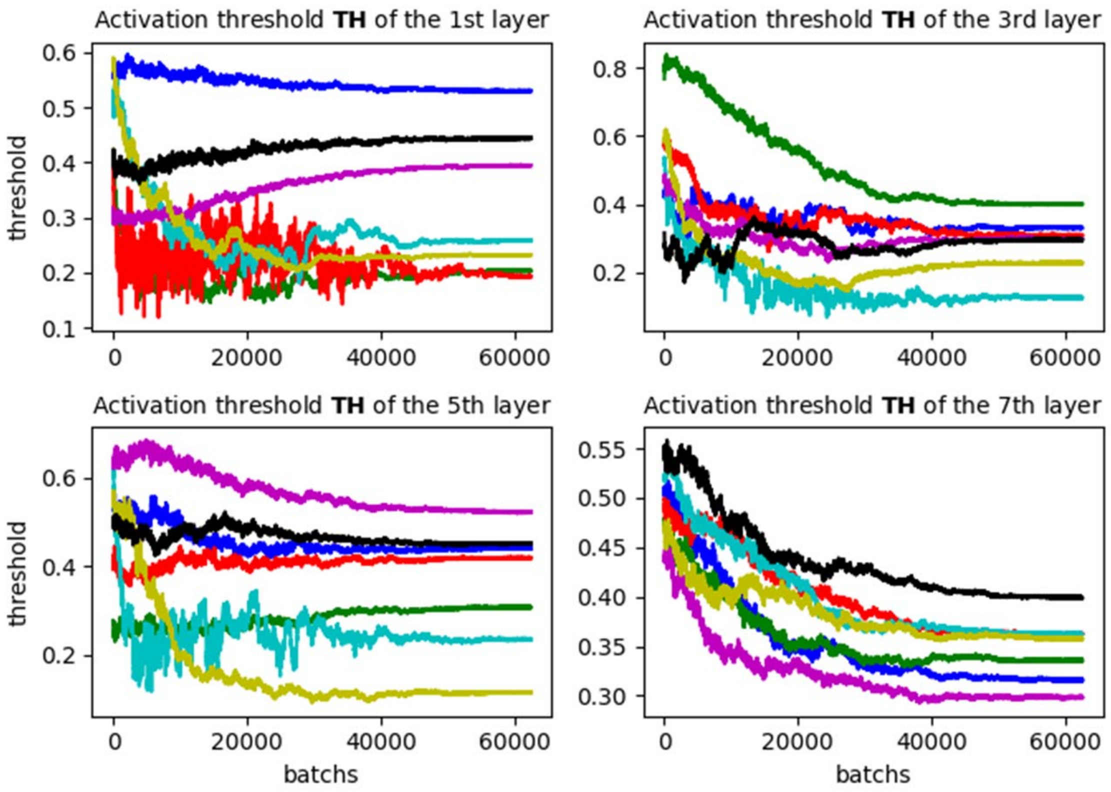
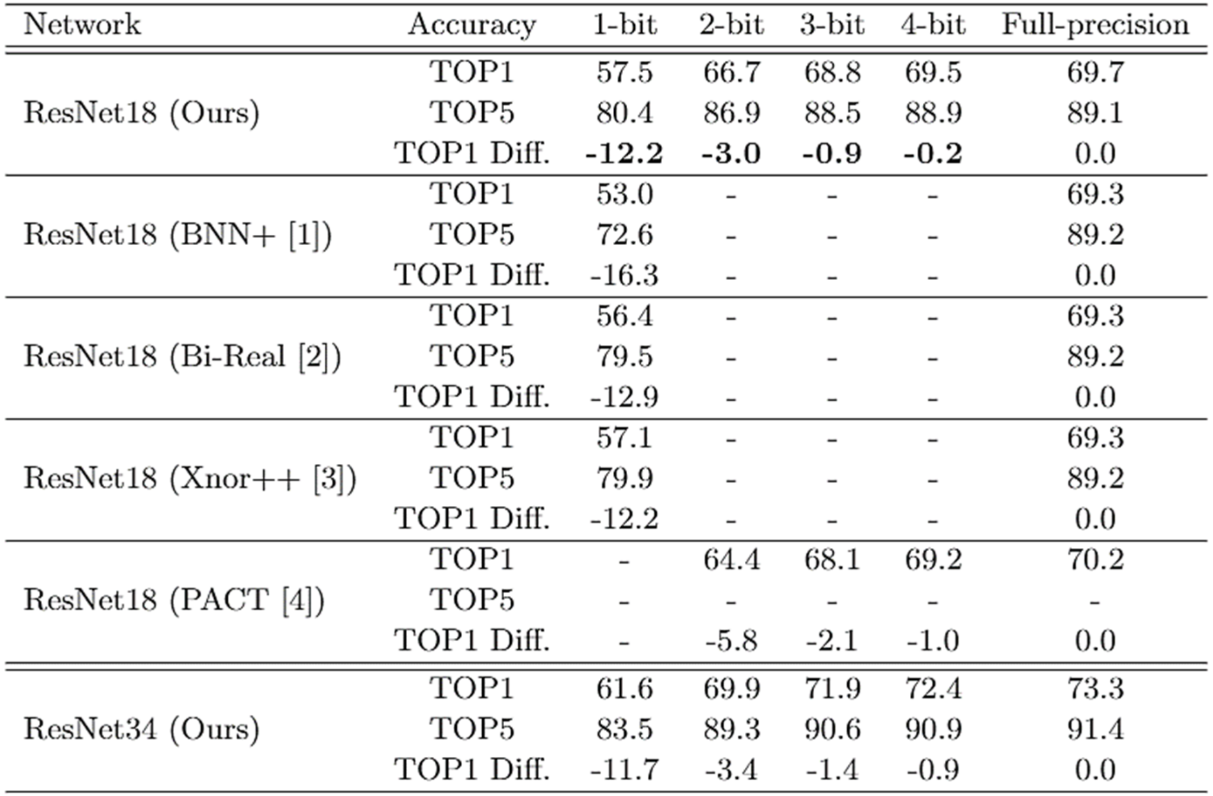


Figure 1. Activation thresholds on the different layers during training



Network	Accuracy	1-bit	2-bit	3-bit	4-bit	Full-precision
ResNet18 (Ours)	TOP1	57.5	66.7	68.8	69.5	69.7
	TOP5	80.4	86.9	88.5	88.9	89.1
	TOP1 Diff.	**-12.2**	**-3.0**	**-0.9**	**-0.2**	0.0
ResNet18 (BNN+ [1])	TOP1	53.0	-	-	-	69.3
	TOP5	72.6	-	-	-	89.2
	TOP1 Diff.	-16.3	-	-	-	0.0
ResNet18 (Bi-Real [2])	TOP1	56.4	-	-	-	69.3
	TOP5	79.5	-	-	-	89.2
	TOP1 Diff.	-12.9	-	-	-	0.0
ResNet18 (Xnor++ [3])	TOP1	57.1	-	-	-	69.3
	TOP5	79.9	-	-	-	89.2
	TOP1 Diff.	-12.2	-	-	-	0.0
ResNet18 (PACT [4])	TOP1	-	64.4	68.1	69.2	70.2
	TOP5	-	-	-	-	-
	TOP1 Diff.	-	-5.8	-2.1	-1.0	0.0
ResNet34 (Ours)	TOP1	61.6	69.9	71.9	72.4	73.3
	TOP5	83.5	89.3	90.6	90.9	91.4
	TOP1 Diff.	-11.7	-3.4	-1.4	-0.9	0.0

Table 1. Top-1 and top-5 accuracy (in percentage) of our quantization scheme and prior state-of-the-art quantization methods on ImageNet dataset

[1] Sajad Darabi, Mouloud Belbahri, Matthieu Courbariaux, andVahid Partovi Nia. BNN+: improved binary network training

[2] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yand, Wei Liu, and Kwang-TingCheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved rep-resentational capability and advanced training algorithm

[3] Adrian Bulat and Georgios Tzimiropoulos. Xnor-net++: Improved binary neuralnetworks

[4] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce IJen Chuang, Vijay-alakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: parameterized clippingactivationfor quantized neural networks

Premier Sponsor

Intelligent Agent
Neuton

Automated TinyML

Zero-code SaaS solution

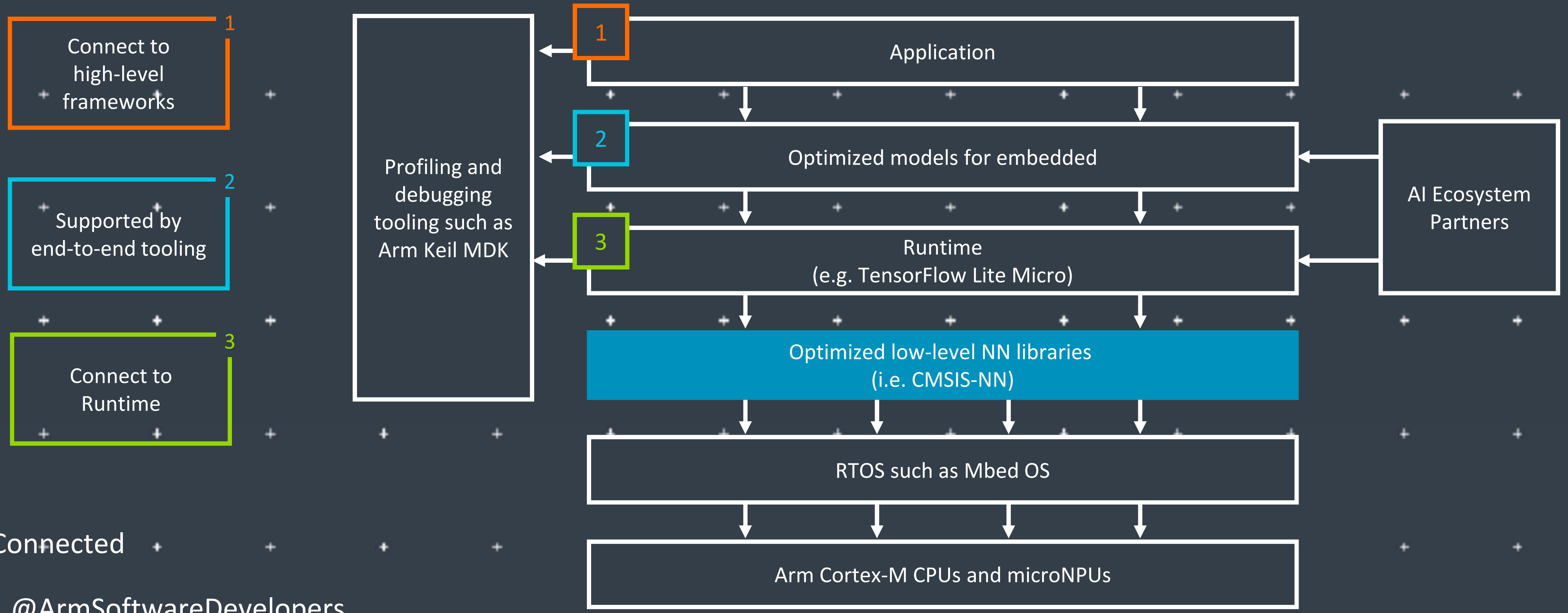
Create tiny models, ready for embedding, in just a few clicks!

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

Build Fast. Build Once. Never Compromise.

Executive Sponsors

Arm: The Software and Hardware Foundation for tinyML


1
Connect to high-level frameworks
2
Supported by end-to-end tooling
3
Connect to Runtime
Profiling and debugging tooling such as Arm Keil MDK
1
Application
2
Optimized models for embedded
3
Runtime
(e.g. TensorFlow Lite Micro)
Optimized low-level NN libraries
(i.e. CMSIS-NN)
RTOS such as Mbed OS
Arm Cortex-M CPUs and microNPUs
AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

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

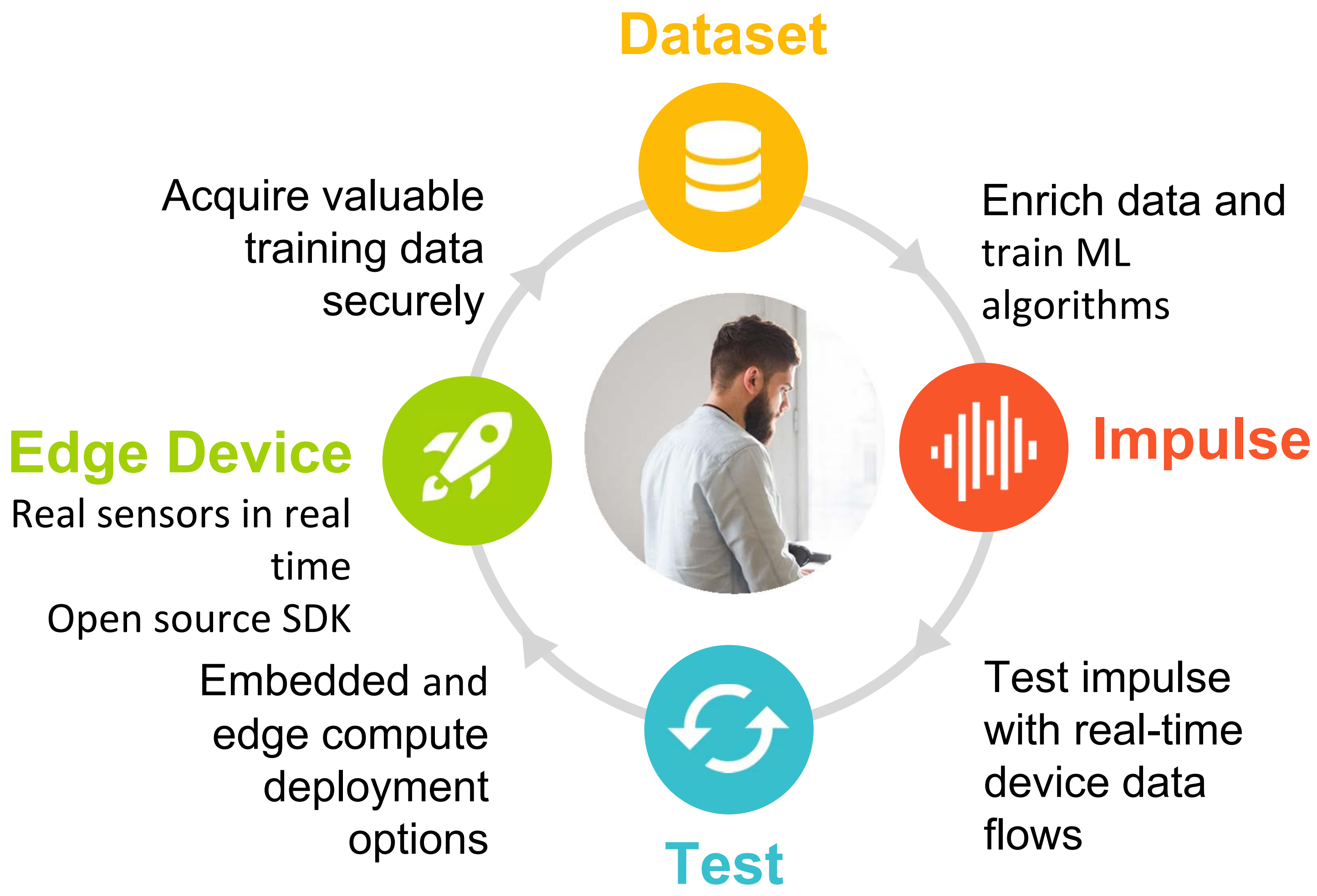
© 2020 Arm Limited (or its affiliates)

TinyML for all developers

C++ library

Arduino library

WebAssembly



Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Impulse

Test impulse with real-time device data flows

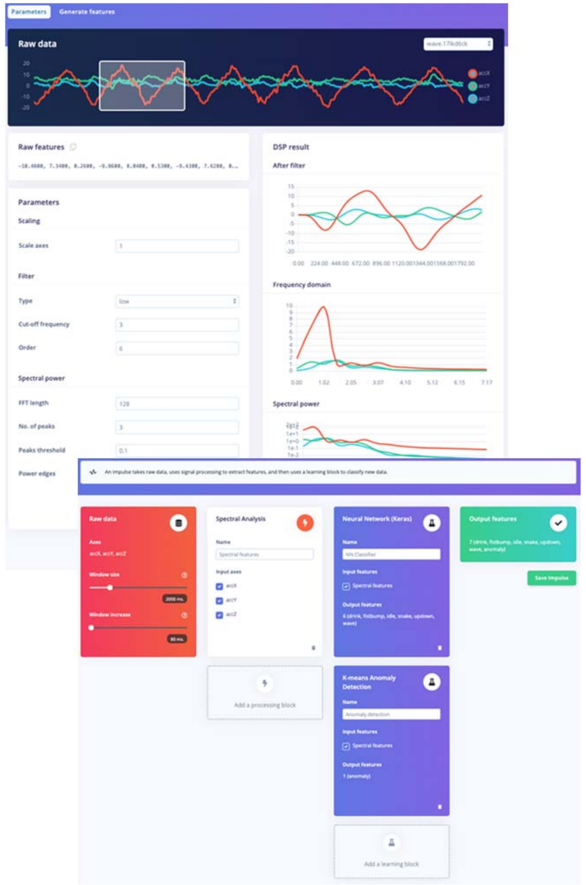
Test

Embedded and edge compute deployment options

Edge Device

Real sensors in real time

Open source SDK



Raw features

-10.4000, 7.3400, 0.2600, -0.0600, 0.0400, 0.5300, -5.4300, 7.6200, 0..

Parameters

Scaling

Scale axes 1

Filter

Type low

Cut-off frequency 3

Order 6

Spectral power

FFT length 128

No. of peaks 3

Peaks threshold 0.1

Power edges

An impulse takes raw data, uses signal processing to extract features, and then uses a learning block to classify new data.

Raw data

Axes

accX, accY, accZ

Window size

2000 ms.

Window increase

80 ms.

Spectral Analysis

Name

Spectral features

Input axes

☑ accX

☑ accY

☑ accZ

Add a processing block

Neural Network (Keras)

Name

NN Classifier

Input features

☑ Spectral features

Output features

6 (drink, fistbump, idle, snake, updown, wave)

K-means Anomaly Detection

Name

Anomaly detection

Input features

☑ Spectral features

Output features

1 (anomaly)

Add a learning block

Output features

7 (drink, fistbump, idle, snake, updown, wave, anomaly)

Save impulse

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

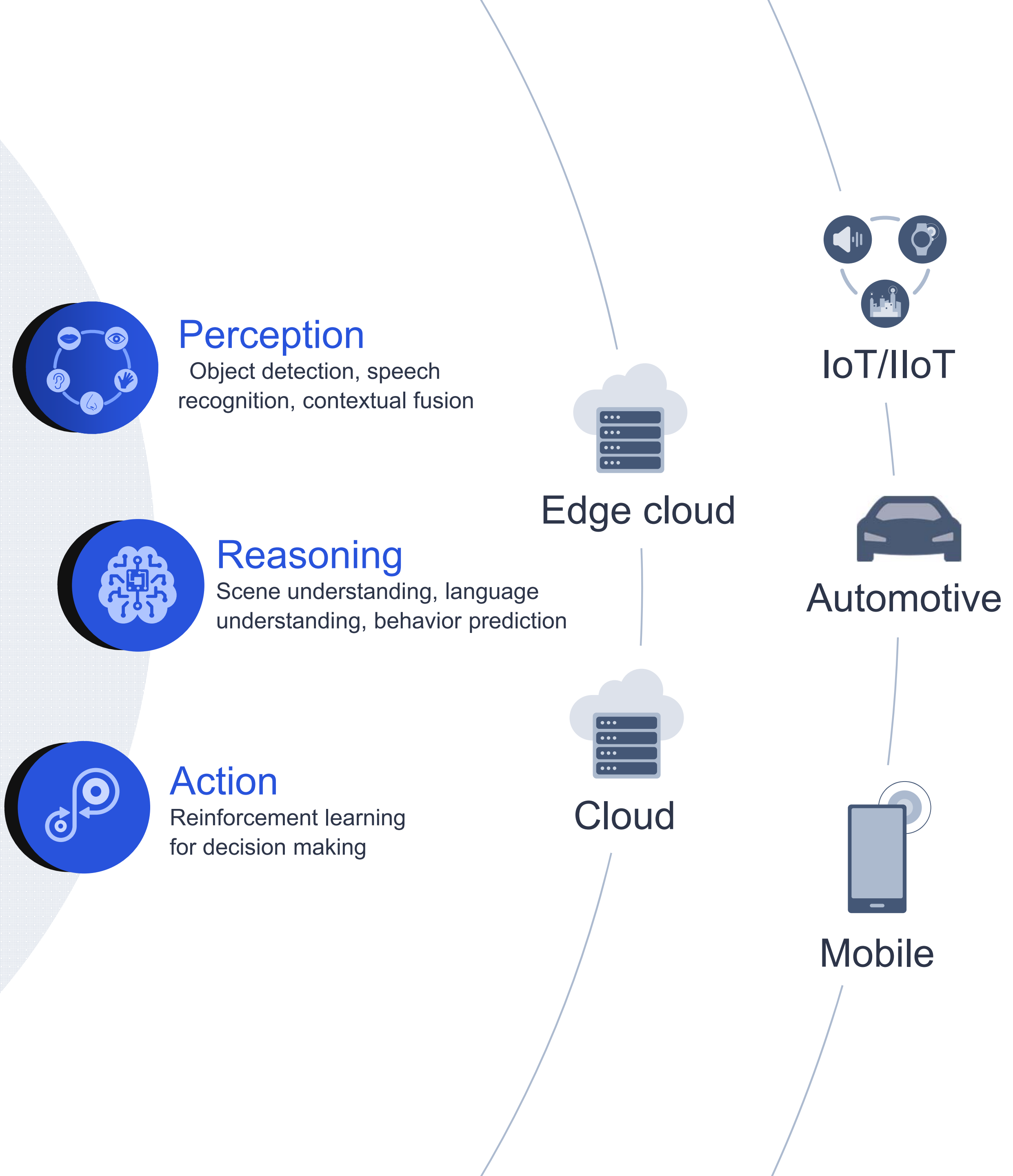
Personalization

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

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^{TM} 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

Platinum Sponsors

infineon

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

Gold Sponsors

LatentAI

Adaptive AI for the Intelligent Edge

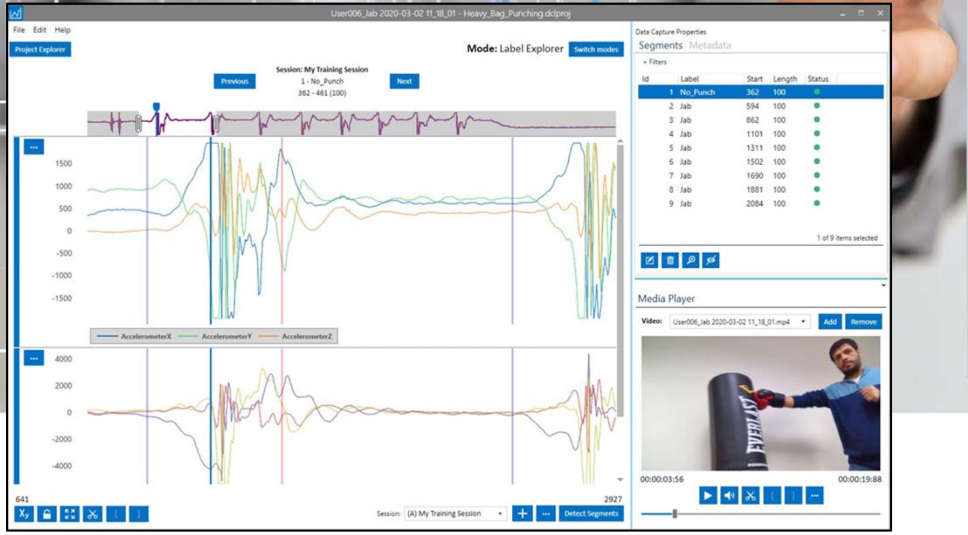
Latentai.com

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



Session: My Training Session
1 - No_Punch
362 - 461 (100)

Session: (A) My Training Session

Id	Label	Start	Length	Status
1	No_Punch	362	100	●
2	Jab	594	100	●
3	Jab	862	100	●
4	Jab	1101	100	●
5	Jab	1311	100	●
6	Jab	1502	100	●
7	Jab	1690	100	●
8	Jab	1881	100	●
9	Jab	2084	100	●

1 of 9 items selected

Video: User006_Jab 2020-03-02 11_18_01.mp4

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

sensiml.com

Silver Sponsors

emza
visual sense

GREENWAVES
TECHNOLOGIES

HOTG

imagimob

Qeexo

seeed studio
The IoT Hardware Enabler

ST
life.augmented

TINY ML EMEA
Virtual
JUNE 7-10,2021

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