“Amber: A Complete, ML-Based, Anomaly Detection Pipeline for Microcontrollers”

Brian Turnquist and Rodney Dockter – Boon Logic

November 24, 2020
tinyML Talks Sponsors

arm

Deeplite

tinyML Strategic Partner

EDGE IMPULSE

maxim integrated™

Qeexo

Reality AI®

SynSense

Additional Sponsorships available – contact Bette@tinyML.org for info
Optimized models for embedded
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

1. Connect to high-level frameworks
   - Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling
   - Application
   - Optimized models for embedded
   - 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

3. Connect to Runtime

Stay Connected
- @ArmSoftwareDevelopers
- @ArmSoftwareDev

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

© 2020 Arm Limited (or its affiliates)
WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

Automatically compress SOTA models like MobileNet to <200KB with little to no drop in accuracy for inference on resource-limited MCUs

Reduce model optimization trial & error from weeks to days using Deeplite's design space exploration

Deploy more models to your device without sacrificing performance or battery life with our easy-to-use software

BECOME BETA USER bit.ly/testdeeplite
TinyML for all developers

- **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**: Get your free account at [http://edgeimpulse.com](http://edgeimpulse.com)
Maxim Integrated: Enabling Edge Intelligence

www.maximintegrated.com/ai

**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.

**Low Power Cortex M4 Micros**
The biggest (3MB flash and 1MB SRAM) and the smallest (256KB flash and 96KB SRAM) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

**Advanced AI Acceleration**
The new MAX78000 implements AI inferences at over 100x lower energy than other embedded options. Now the edge can see and hear like never before.
Qeexo AutoML for Embedded AI
Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Wide range of ML methods: GBM, XGBoost, Random Forest, Logistic Regression, Decision Tree, SVM, CNN, RNN, CRNN, ANN, Local Outlier Factor, and Isolation Forest
- Easy-to-use interface for labeling, recording, validating, and visualizing time-series sensor data
- On-device inference optimized for low latency, low power consumption, and a small memory footprint
- Supports Arm® Cortex™- M0 to M4 class MCUs
- Automates complex and labor-intensive processes of a typical ML workflow – no coding or ML expertise required!

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

QEEXO AUTOML: END-TO-END MACHINE LEARNING PLATFORM

For a limited time, sign up to use Qeexo AutoML at automl.qeexo.com for FREE to bring intelligence to your devices!
Reality AI is for building products

Reality AI Tools® software
- Automated Feature Exploration and Model Generation
- Bill-of-Materials Optimization
- Automated Data Assessment
- Edge AI / TinyML code for the smallest MCUs

Reality AI solutions
- Automotive sound recognition & localization
- Indoor/outdoor sound event recognition
- RealityCheck™ voice anti-spoofing

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
SynSense builds **ultra-low-power** (sub-mW) sensing and inference hardware for **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
## Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
</table>
| Tuesday, December 8 | Ian Campbell  
CEO, OnScale  
Sakyasingha Dasgupta  
Founder, CEO & Managing Director, Edgecortix Inc. | Training Embedded AI/ML Using Synthetic Data  
Using AI to design energy-efficient AI accelerators for the edge |

Webcast start time is 8 am Pacific time  
Each presentation is approximately 30 minutes in length

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

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Brian Turnquist has worked in machine learning for the past twenty years developing numerous novel algorithms for automatically clustering biological signals in real-time. Turnquist is CTO of Minneapolis tech start-up, Boon Logic, and a previous visiting researcher at the Universities of Nürnberg and Heidelberg, and tenured professor at Bethel University. His Ph.D. is in Mathematics from the University of Maryland with fourteen refereed publications in neuroscience and mathematics.
Rodney Dockter

Rodney Dockter is an engineer with a broad background in robotics and machine learning. Application areas include industrial automation, autonomous off-highway vehicles, mobile robotics, and surgical robotics. Rodney is the Director of Computer Vision at Minneapolis-based Boon Logic, and a lecturer at the University of Minnesota. He holds a Ph.D. in Mechanical Engineering from the University of Minnesota.
Amber: A Complete, ML-Based, Anomaly Detection Pipeline for Microcontrollers

Brian Turnquist, Ph.D.
Rod Dockter, Ph.D.
24 November 2020
Comparison of Amber to a Common Tiny ML Approach

Three examples: Output current, Vibration, EEG

Core algorithm benchmark: The Boon Nano vs. K-means

Arm Cortex M7 and M4 Benchmarks

Implications for Tiny ML
Amber is an anomaly detection technology for single sensor and sensor fusion data streams.

Amber can be trained and deployed entirely on the microcontroller without compute support from gateways or the cloud.

Amber is NOT a neural network and not based on neural networks.

Amber is based on unsupervised ML segmentation of the sensor signal.

Amber is an application of Boon Logic’s core segmentation algorithm, the Boon Nano.
A Common Approach to Tiny ML

1. Record data needed for model training and upload it to a general purpose computer

2. Label anomalies (if supervised solution)

3. Apply cloud and/or GPU resources to build inference model, prune model

4. Deploy custom model to MCU to begin monitoring
The Amber Approach to Tiny ML

**Autotuning**
Automatically find optimal clustering parameters (~ 1000 samples)

**Learning**
Automatically build segmentation model using unsupervised learning (Boon Nano)

**Monitoring**
Use the segmentation model to find anomalies

No Cloud
No Gateway

on MCU
The Amber Approach to Tiny ML

Core Per-Sample Analytic Outputs

ID: The cluster ID of the subsignal (most recent n samples)
AI: The frequency-based anomaly index of the subsignal (our focus for this talk)
DI: The distance-based anomaly index of the subsignal
AH: The ongoing count of anomalies in the local observation window
AM: Expectation of anomalies above background anomaly activity
AW: Criticality of the anomaly level
   0 = green = signal/asset is compliant/nominal (within normal variation)
   1 = amber = signal/asset is changing
   2 = red = signal/asset is critical
Data source: Customer
— Allen Bradley drive + motor
— Bottle-capping profile
— Use PLC “Output Current” tag
— Intel Atom Dual-core 1.46 GHz processor (Amber had CPU 4% utilization)
— Do universal models work? No.
— Amber detection of transient, drift, and shift anomalies
— Transient anomalies introduced every 30 seconds
Example: Vibration

Data source: Synthetic

— 40 principal frequencies
— Gaussian noise
— Transient anomalies (200 ms) at 60 and 75 seconds
— Transient “shudder” (2 seconds) at 90 seconds
— New frequency component at 105 seconds
Example: Brain Waves

Data source: MIT data set for ML applied to seizure detection

—Shoeb and Guttag 2010

—23 cranial electrodes per patient, all with epilepsy

—24-hour recordings

—Seizure episode ground truth scored by domain experts
Clustering results of K-means (L2 distance) comparable to Boon Nano (L1 distance)

Boon Nano is much faster (~1000x) and much smaller (up to 1024 clusters on MCU)

<table>
<thead>
<tr>
<th>Per Pattern Inference Time (microseconds)</th>
<th>100 Clusters</th>
<th>500 Clusters</th>
<th>1000 Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means (Scikit-learn/Python)</td>
<td>737</td>
<td>3409</td>
<td>6746</td>
</tr>
<tr>
<td>K-Means (MLPack/C++)</td>
<td>233</td>
<td>1197</td>
<td>2397</td>
</tr>
<tr>
<td>Boon Nano</td>
<td>0.26</td>
<td>1.1</td>
<td>2.15</td>
</tr>
</tbody>
</table>

Four core 32G RAM EC2 instance on AWS KDDCup99 Network Intrusion Benchmark
Arm Cortex M7
— 480 MHz CPU
— 1 MB Internal RAM
— 8MB External SDRAM
— Wi-Fi/Bluetooth Radio Module
— Single Cell Li-Po Battery
Arm Cortex Benchmarks: Inference Rate

M7 @ 480 MHz
~50 Clusters
40 +/- 24 us / Inference
Sensor Max Sample Rate = 25 Khz

M4 @ 240 Mhz
~50 Clusters
165 +/- 112 us / Inference
Sensor Max Sample Rate = 6 Khz

Max Clusters: 1024
External SDRAM: 2.1 MB
Internal RAM: 73 KB
Amber builds a customized inference engine for each individual sensor stream.

Amber builds its inference engine in real-time without human insight or intervention.

Amber operates fully autonomously as it monitors each sensor.

Amber is scalable (think millions of sensors).

Amber is deployable entirely at the edge.
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.tinyML.org