“On-device model fine-tuning for industrial anomaly detection applications”

Konstantin Meshcheriakov - Klika Tech

March 8, 2022
tinyML Talks Strategic Partners

AONdevices  arm  Deeplite  EDGE IMPULSE  emza visual sense

FotaHub  GREENWAVES TECHNOLOGIES  Grovety Inc.  HOTG  imagimob  itemis

KLIKA TECH  LatentAI  maxim integrated. ANALOG DEVICES  Micro.ai  NXP

Qualcomm  Qeexo  RealityAI  Renesas  REEXEN technology  SAP  seeed

SensiML™  Sony Semiconductor Solutions Corporation  STI  STREAM ANALYZE  SynSense  SYNTIANT

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
EDGE IMPULSE  The leading edge ML platform

www.edgeimpulse.com
Advancing AI research to make efficient AI ubiquitous

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

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

**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

Contact: partners@syntiant.com
Website: www.syntiant.com
Platinum Strategic Partners
Fastest Video Analytics Solutions on Arm CPUs
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.
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

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

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
LatentAI
Adaptive AI for the Intelligent Edge
Latentai.com
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
The Best Product of the Year and the Best Innovation of the Year awards are open for nominations between **November 15 and March 14.**

**tinyML Summit 2022**
Miniature dreams can come true…
March 28-30, 2022
Hyatt Regency San Francisco Airport
https://www.tinyml.org/event/summit-2022/

**tinyML Research Symposium 2022**
March 28, 2022
https://www.tinyml.org/event/research-symposium-2022

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/

8.5k members in
43 Groups in 34 Countries

2.6k members
&
5.4k 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>Tuesday, March 15</td>
<td>Matthias Hertel, Product Specialist, Arm</td>
<td>Accelerate ML development with cloud-based Arm Cortex-M models</td>
</tr>
</tbody>
</table>

Webcast start time is 8:00 am Pacific time

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
Konstantin Meshcheriakov

Konstantin is a solution architect at Klika Tech with strong experience in building embedded and machine learning solutions. Working closely with the clients, he is responsible for architecture creation and initiation of new IoT and ML-related projects. He also leads the machine learning competency and internal courses in the company.
On-device model fine-tuning for industrial anomaly detection applications

KLIKA·TECH
About us

End-to-End IoT Solutions
- Design, develop, deploy and support solutions and applications leveraging our customers’ and partners’ technologies and services.

Embedded and Cloud-native Development
- Focused on multi-architecture Embedded and Cloud-native serverless applications, delivery of all backend applications, data lakes, APIs, enterprise integrations, mobile, and UI/UX co-creating a fully-realized application.
- Multi-Cloud expertise: AWS, MS Azure, Alibaba, Hybrid.

Big Data AI/ML Experts
- Empowering the next wave of innovation by analyzing patterns and trends in data sets to drive new business insights and outcomes.
Global Partnerships
The project
Maintenance types

Reactive Maintenance

Preventive Maintenance

Prescriptive Maintenance

Predictive Maintenance
# Predictive maintenance

## Remaining Useful Life (regression)

**Question:** How long will it take for a failure to occur?

**Data:** Historical sensor data along with events labels

**Assumptions:**
- The RUL can be modeled
- Just one type of failure can be modeled with a given model
- Accurate labels are available

## Failure within a given time window (classification)

**Question:** Will the equipment fail within the next N days/cycles?

**Data:** Historical sensor data along with events labels

**Assumptions:**
- Multiclass classification can be used to model different potential failure types
- Enough data and accurate labels are available for each potential failure

## Anomaly detection

**Question:** Is the current equipment behavior normal?

**Data:** Historical sensor data with few possible failures recorded

**Assumptions:**
- It is possible to define a difference between normal and anomalous behavior
- Anomalous behavior is not always a failure
HVACR condition monitoring and anomaly detection

- Retrofitting of any existing HVACR systems
- Support of broad range of sensors
- Near-real time condition monitoring
- Near-real time anomaly detection
Infineon XENSIV™ predictive maintenance evaluation kit

- Current monitoring (up to 120 A), e.g. for overall current anomaly detection
- Vibration and position sensing, e.g. for drives and compressor
- Air-flow and pressure measurement, e.g. for filters
- Open/closed lid detection, e.g. for detecting the status of service lids
- Hall speed sensors, e.g. for blocked fan and fan speed monitoring
- Sound anomaly detection, e.g. for gears and joints

The cloud architecture

1. HVAC device
   - MQTT sensors data

2. AWS IoT Core
3. IoT Rule
4. AWS Lambda (enrich metrics)

5. Amazon Kinesis Data Firehose
6. Amazon Kinesis Data Stream (initial metrics)
7. Amazon Kinesis Data Analytics (anomaly detection)
8. AWS Lambda (anomaly to ES)
9. Kibana

10. Amazon Elasticsearch Domain
11. Amazon CloudWatch
12. AWS CloudFormation
13. AWS IAM
14. Amazon Cognito
15. Amazon S3 Bucket (logs)

16. User

- A sketching algorithm
- Works really well on streaming data
- Assemble of Random Cut Trees with all operations designed to maintain the distribution of data
- Uses averaged collusive displacement (CoDisp) to calculate the anomaly score
Cloud-based solution advantages and disadvantages

**PROS:**

- Ease of deployment
- Near real-time data visualization
- Streaming data collection for datasets creation
- Adaptation of the models to customers’ environment

**CONS:**

- Infrastructure and data transfer costs
- The necessity to pretrain the cloud anomaly detection algorithm
Moving to TinyML
Anomaly detection in the tiny world

- K-Means
- DROCC: Deep Robust One-Class Classification (EdgeML)
- One Class SVM
- ...

Models comparison, PyOD documentation, by Yue Zhao, 2019
**Autoencoder**

- Neural network with a bottleneck layer
- Is trained with label equal to inputs
- The reconstruction error metrics can be used for anomaly detection

**Anomaly Score:**

\[
MAE(x, x') = \frac{1}{N} \sum_{i=1}^{N} |x_i - x'_i|
\]
Concept Drift

Concept drift (dataset shift) occurs when the distribution of the data on which the model was trained differs from the one on which it is used. It can be caused by:

- Sample selection bias (biased data collection)
- Nonstationary environment
  - Environmental changes
  - Manufacturing process non-repeatability
  - Modification of the equipment
  - User behavior changes

TinyML related papers:

Adding fine-tuning to a TFLite model

Fine-tuning stages:

1. Convert the model to TFLite Format
2. Deploy the model to the device using the usual method
3. On first boot, the ModelEditor parses the model, separates a specified number of trainable layers from the model, and stores the modified model and trainable layers to the device flash
4. During the inference the Interpreter performs the inference using the reduced model and uses the model's head to compute the results
5. If fine-tuning is enabled, the Interpreter runs the back-propagation for the model's head and stores the modified head to the device's flash.
Demo #2
Catastrophic forgetting is a tendency of a neural network to “forget” previously learned when learning new information.

It can be encountered when:
- The multiclass classification model is continuously trained on the data from a single class
- Anomaly detection model works on data with multimodal distribution and is continuously retrained with data from a single mode

Methods to avoid it:
- **Rehearsal** or **Pseudo-rehearsal** - mixing the learning on the specific class and learning on the remembered data
- **Orthogonality** - “disentangling” paths of the neural network intended for different tasks
Catastrophic forgetting
Further steps

- Change the architecture to provide support for any inference engine
- Add quantization support
- Add hardware-specific operations support
- Support additional operations
Thank You

Joseph Zaloker
Head of Global Alliances and Enterprise Accounts
Email: jzaloker@klika-tech.com
Phone: +1 651 261-4860

Gennadiy Borisov
Co-CEO / President
Email: gborisov@klika-tech.com
Phone: +1 917-770-3133

Konstantin Meshcheryakov
Solution Architect
Email: kmeshcheryakov@klika-tech.com
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.tinyml.org