The Future of Embedded ML

Alessandro Grande
Head of Product

ICTP, Trieste - July 3, 2023
Agenda

1. Intro to Edge Impulse
2. Customer challenges
3. What’s beneath the surface
4. Resources
5. Next steps
The edge AI platform
Number of Projects on Edge Impulse
Number of Projects on Edge Impulse
TinyML Use Cases

Health

Industrial

Wearables

Infrastructure

Buildings

ŌURA

NOWATCH

Brambles

Lexmark

Hyfe AI

ECOLAB

IZOELEKTRO
Production Challenges

1. Data collection
2. Data quality analysis
3. Feature extraction and DSP
4. Deployment
5. Monitoring performance
UI / Visible

Headless / Production

Edge Impulse Studio
dataset evaluation

Data Collection
10m 45s

Record new data

Device:
My device

Label:
Mode 1

Sensor:
Build-in accelerometer

Frequency:
100Hz

Start sampling
BYOM & Python SDK

- Profile on-device performance of any trained model
- Analyze the impact of architectural decisions
- Generate optimized C++ libraries
- Deploy to any edge device
CLI

Edge Impulse CLI tools

Command-line interface tools for Edge Impulse. We make things smarter by enabling developers to create the next generation of intelligent device solutions with embedded Machine Learning.

This package consists of four tools (click to see their respective documentation):

- **edge-impulse-daemon** - configures devices over serial, and acts as a proxy for devices that do not have an IP connection.
- **edge-impulse-uploader** - allows uploading and signing local files.
- **edge-impulse-data-forwarder** - a very easy way to collect data from any device over a serial connection, and forward the data to Edge Impulse.
- **edge-impulse-run-impulse** - show the impulse running on your device.
- **edge-impulse-blocks** - create organizational transformation blocks.
- **eta-flash-tool** - to flash the Eta Compute ECM3532 AI Sensor.
- **himax-flash-tool** - to flash the Himax WE-I Plus development board.
Embedded ML in the Real World

Automated pipeline

1. Create dataset
2. Develop & train model
3. Test & tune
4. Deploy
5. IoT Platform
6. Gateway
7. Algorithms
8. Machines

Studio & APIs
- New data
- Active learning

Edge Impulse Studio
- Engineering
- Business Intelligence
- Factory Manager

Dashboard

Alerts

Performance Monitoring
Interactive Feature Engineering

Real-time visualization of DSP

- Immediate feedback loop enabling tactile exploration by domain expert
- Service-based architecture for real-time DSP on individual samples (separate from job-based system for batched data)
Feature Importance

Don’t use everything

- Uses recursive feature elimination with cross-validation (RFECV)
- Only computed for relatively low-dimensionality data
FOMO: Faster Objects, More Objects

- 20x average performance improvement
- Object detection on MCUs
- Ultra fast on embedded Linux
- Better at detecting smaller and more numerous objects
- Capable of segmentation and counting objects

<table>
<thead>
<tr>
<th></th>
<th>Cortex-M4</th>
<th>Cortex-M7</th>
<th>Cortex-A</th>
<th>Nvidia</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOMO</td>
<td>2 fps</td>
<td>15-30 fps</td>
<td>60+ fps</td>
<td>150+ fps</td>
</tr>
<tr>
<td>SSD</td>
<td>NA</td>
<td>NA</td>
<td>3 fps</td>
<td>20 fps</td>
</tr>
</tbody>
</table>

Copyright © Edge Impulse Inc.
FOMO: Faster Objects, More Objects

- Remove classification head, replace with GMMs
- Only requires training on normal data
- Each cell tells you the chance that it's an anomaly
- Same performance as FOMO: Up to 30fps. on Cortex-M7, <200K RAM
Keras Expert Mode

- For advanced users
- Use Keras standard API
- Customize NN architecture and take full control over training procedure
EON Tuner

Establish a baseline quickly

- Search space based on prior knowledge of data modalities
- Reusable workers to minimize startup cost
- Customize search space
UI / Visible

- Edge Impulse Studio
- Python SDK
- CLI and API bindings
- DSP and advanced ML

Headless / Production

- Scalable data infrastructure
Data-Centric ML

Collect
- Rapidly build custom datasets

Design
- Develop models & algorithms

Test
- Validate your model performance

Deploy
- Easily deploy on any edge target

Edge device
- Update device firmware
- Store relevant samples in your cloud

Fetch new data samples
- Trigger webhook

Fetch new firmware
Active Learning with Edge Impulse

Active Learning with Edge Impulse

- **Data**: Rapidly build custom datasets
- **Design**: Develop models & algorithms
- **Test**: Validate your model performance
- **Deploy**: Easily deploy on any edge target

Flow:
1. Fetch new data samples
2. Store relevant samples in your cloud
3. Edge device
4. Update device firmware
5. Fetch new firmware
6. Trigger webhook
7. Update edge device
Working with Data

- Data pipelines and transformation - enabling data preparation at scale
- Data campaign dashboards - optimize performance and share learnings

**Data pipelines and transformations**

**Data dashboards**
Data preparation

- **Fetch data**
- **Basic checks:** Are all files present? Do all files start / end around the same time? All expected labels for the study present?
- **Advanced checks:** Correlation between different devices (e.g. HR from PPG, and HR from Polar)?
- Runs automatically at a set interval (or on-demand, or triggered from code)
- Sends email on new data
Visualize data and uncover critical insights

Data explorer

The data explorer shows a complete view of all data in your project. Use it to quickly label your data, or spot outliers. Learn more.

How should we generate the data explorer?

- Using a pretrained keywords model
  Great for keywords that fit in a 1 second window.

- Using your trained impulse
  Works great if you have collected some labeled data already and have a trained model.

- Using the preprocessing blocks in your impulse
  Use this if you don't have any labels for your data yet, and thus can't train a full model.

Dimensionality reduction technique

- t-SNE
  Recommended for your dataset. Separates best, but takes a significant amount of time on large datasets.

- PCA
  Separates less well, but works on any dataset size.
Validating data: correlation

Fixes many issues: data uploaded for wrong participant, device failure and can be used to collect clock drift. Here implemented for PPG (derive HR) + Polar H10.
Resources
(Courses)

tinyml.seas.harvard.edu
# Curriculum and Content

## Full Courses

<table>
<thead>
<tr>
<th>Organization</th>
<th>Course Name</th>
<th>Date of Course</th>
<th>Target Audience</th>
<th>Language of Instruction</th>
<th>Language of Materials</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>edX tinyML Specialization by Harvard University</td>
<td>Launched 2020-2022</td>
<td>Everyone</td>
<td>English</td>
<td>English</td>
<td><a href="#">Course 1-3 Website</a></td>
<td><a href="#">All Materials</a></td>
</tr>
<tr>
<td>Embedded Machine Learning on Coursera by Edge Impulse</td>
<td>Launched 2021-2022</td>
<td>Everyone</td>
<td>English</td>
<td>English</td>
<td><a href="#">Course 1</a></td>
<td><a href="#">Course 2</a></td>
</tr>
<tr>
<td>ESE3600: Tiny Machine Learning by the University of Pennsylvania</td>
<td>Fall 2022</td>
<td>Undergraduate and Graduate Students</td>
<td>English</td>
<td>English</td>
<td><a href="#">Website and Materials</a></td>
<td></td>
</tr>
<tr>
<td>MIT 6.5965 TinyML and Efficient Deep Learning</td>
<td>Fall 2022</td>
<td>Graduate Students</td>
<td>English</td>
<td>English</td>
<td><a href="#">Website Materials</a></td>
<td></td>
</tr>
<tr>
<td>UNIFEI EESTI01 TinyML – Machine Learning for Embedding Devices</td>
<td>Jan 2021 - Present</td>
<td>Undergraduate Students</td>
<td>Portuguese</td>
<td>English</td>
<td><a href="#">2022 Website and Materials</a></td>
<td><a href="#">2021 Website and Materials</a></td>
</tr>
<tr>
<td>Harvard CS249r Tiny Machine Learning</td>
<td>Sept 2020 - Present</td>
<td>Graduate Students</td>
<td>English</td>
<td>English</td>
<td><a href="#">2022 Website and Assignments</a></td>
<td><a href="#">2020 Website Assignments</a></td>
</tr>
</tbody>
</table>
Resources
(Projects)

www.edgeimpulse.com/projects
docs.edgeimpulse.com/experts
Project: Smart HVAC

- **Creator:** Jallson Suryo
- **Description:** Set heating/cooling based on number of people in each room
- **Hardware:** Arduino Nicla Vision
- **Model:** FOMO

[Link to project documentation](docs.edgeimpulse.com/experts/featured-machine-learning-projects/arduino-nicla-vision-smart-hvac)
Project: Artificial Nose

- **Creator:** Benjamin Cabé
- **Description:** Classify different odors based on gas data
- **Hardware:** Seeed Studio Wio Terminal
- **Model:** DNN

github.com/kartben/artificial-nose
Project: Motor Anomaly Detection

- **Creator:** Avi Brown
- **Description:** Identify anomalies based on motor current and voltage
- **Hardware:** Raspberry Pi Pico
- **Model:** K-means clustering

Project: Concrete Surface Crack Detection

- **Creator:** Naveen Kumar
- **Description:** Identify surface cracks in concrete structures
- **Hardware:** TI TDA4VM
- **Model:** MobileNetV2 with CAM

docs.edgeimpulse.com/experts/prototype-and-concept-projects/surface-crack-detection-ti-tda4vm
University Program

edgeimpulse.com/university

1. Free hardware kits
2. Content to build curriculum
3. Access to expert network
4. Discount to enterprise edition

Deadline July 16
Let’s simplify embedded ML for the next generation of engineers together

Thanks!