tinyML EMEA
Enabling Ultra-low Power Machine Learning at the Edge

October 10-12, 2022
Limassol, Cyprus

www.tinyML.org
In Sensor and On-device Tiny Learning for Next Generation of Smart Sensors

Dr. Michele Magno

Credits: Prof. Dr. Luca Benini,
Introduction

- Internet of Things (IoT)
  - Variety of sensors
  - Connected to cloud (often wirelessly)

- Machine learning
  - Extract relevant information from data
  - High computational demand

- Data Processing and Learning
  - Mainly in the cloud
Edge Vs Cloud

- Latency/reliability
- Data Protection
- No Wireless Communication Needed – Lower Bandwidth requirements
- Lower Power Consumption
- Lower Cost

Figure reference: Accelerating Implementation of Low Power Artificial Intelligence at the Edge, A Lattice Semiconductor White Paper, November 2018
What about On-device Learning? Can we do at the edge?

- Customization and Personalization: TinyML devices need to continually adapt to new data collected from the sensors.
- Security: Data cannot leave devices because of security and regularization.
AI Workloads from Cloud to Edge (Extreme?)

**GOP+**

**MB+**

**Extreme edge AI challenge**

AI capabilities in the power envelope of an little cores:

- 100mW peak (1mW avg)
- 512KB RAM (best case 1MB)
New platform for edge computing every year. Edge-AI Platforms

- **GOPs/s**
  - 10000
  - 1000
  - 100
  - 1

- **Power/Thermal Envelope**
  - 10W
  - 1W

- **Near-Sensor Processing**
  - 6.2 GLOPS/s @6W 50$
  - 1 TOPS/s @2.5W 80$
  - 4 TOPS/s @3W 150$
  - 1 TOPS/s @20W 1000$

- **Multi-Core MCU**
  - 9 GLOPS/s @0.07W 5$
  - 100 GOPS/s @1W 70$

- **FPGA**
  - 32 TOPS/s @30W 1500$

- **GP-GPU**
  - 1 TOPS/s @20W 1000$

- **Embedded ASIC**
  - 6.2 GLOPS/s @6W 50$

- **Linux-Capable CPU**
  - 1 TOPS/s @20W 1000$

- **On Sensor**
  - 0.6 GLOPS/s @0.1W 4$

- **Near-Sensor Processing**

- **Trends:**
  - Optimized accelerator for NN
  - Parallel Solutions
  - Custom Architectures

D-ITET Center for Project-Based Learning, Integrated Systems Laboratory, ETH Zürich
New trend on-sensors capabilities

Sony IMX 501

IniVation Foveator
DVS-Sensor plus AI cores

Bosch BHI260AP

ST LMS6DSOX and new ISM330

6-axis iMEMO™ IMU with Machine Learning Core
Next generation of IoT devices: **Always-on Smart Sensors.**

1. ) Edge Signal Processing and AI

2. ) Energy harvesting

3. ) Low power system design

4. ) Low Power and long range communication

Smart devices for perpetual operation
Next generation of IoT devices: **Always-on Smart Sensors.**

1. ) Edge Signal Processing and AI

   ![Image](image)

2. ) Energy harvesting

   Smart devices for perpetual operation

3. ) Low power system design

4. ) Low Power and long range communication
On-Sensors ML and ODL

- Typical “Smart” Sensor
  - Traditional Sensors
  - MCU or other Edge processor for data collection and analysis
  - Wired/wireless interface to transmit findings (optional)

- In-sensor approach
  - Intelligence embedded in the sensor IC
  - Always-on analysis and learning
  - “Zero” latency
On-Sensor Learning

- Project Goal: anomaly detection based on Neural Network (encoder-decoder), parameters learned on the device.

- Advantages:
  - Plug-and-Play anomaly detection: no data acquisition, no training/deployment. Place-and-forget device.
  - Always on monitoring with low power consumption
In-Sensor Hardware: ST ISM330AILP

ST’s sensor puts together Sensor and Processing

- Established ST’s MEMS technology + proprietary ISPU (intelligent sensor processing unit)
- NN Core = small MCU with proprietary RISC architecture, FPU, BNN accelerator
- 40KB of memory (RAM+program)
- 4 cycle wake-up

- Previous work shows the performance use with supervised learning!
Challenges

- **Memory:** 8+32KB of memory for both .text and RAM
- **Computation resources**
  - Max 10MHz, but ideally only 5MHz
  - DSP instructions
- **Learning is memory intensive**
  - Only small models allowed as well as encoder/decoder
  - Partial learning (last layer, bias only update)

Current work

- **ODL options**
  - Full training (very memory intensive, some optimizations possible, discouraged) [1]
  - Update bias only (probably negligible effect in small networks) [2]
  - Output layer only (seems the best option) [3]

ODL Pipeline

Example of Dataset: frame error (blue) + smoothed (orange)


DSP library for STRed ISA!
Fancy functions supported in HW (FFT butterflies, circular buffers, SIMD)

*Using mini-batches for learning

Evaluation
On Airbus anomaly Detection Dataset [ETH]

- Evaluating on Anomaly Detection Task
  - Pipeline: RFFT $\rightarrow$ Accumulate/Avg $\rightarrow$ Autoencoder $\rightarrow$ MSE

- Evaluation of Feasibility:
  - Memory requirements
  - Network Inference/Training complexity (FLOPs)
  - Training latency considering real-time data

- Hyper Parameters:
  - FFT size
  - Network size
  - STFT length
  - Regularizations
  - Training time
  - Batch Size
  - …

Evaluation metric: ROC AUC
- Training time of 2 seconds
**Experimental results and evaluation**

**Training vs Inference vs Sampling rate.**

- High AUC/Accuracy can be achieved with
  - FFT preprocessing  672 flops
  - RAM usage: 4-15 KB + program (usually around 8KB with DPS functions) → should fit
  - FLOPS for training 232k
    - 1.6 mJ/training!
    - 0.23s of compute time at 5MHz with 5 cycles/FLOP
  - FLOPS per inference 1097

- We measured in ISPU ~5cycles/FLOP → training latency ~20s due to data to learn, of which ~0.25s of computation
  - Mostly dominated by data sampling rate, not by training
    - Sampling 1024Hz → 20s to acquire enough data for entire training
    - Energy for Sampling 1.8uJ/sample,

- Next steps: try with autoencoder in time domain (justifies the NN better, no need for FFT code (saves memory))
Conclusion and take away

- TinyML come become very tiny with in-sensors device
- ODL is an hot emerging top but even more challenging
- Anyway new technology are enabling both in sensors and ODL
- With a lot of limitation! :D
Thank you for your attention!

- TinyML and Embedded Control on Robots

- Self Sustaining Smart-Sensors for Indoor and outdoor Applications and Industrial Applications
Thank you
pbl.ee.ehtz.ch

Michele.magno@pbl.ee.ethz.ch
Copyright Notice

This presentation in this publication was presented at the tinyML® EMEA Innovation Forum 2022. 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.