A Battery-Free Long-Range Wireless Smart Camera for Face Detection: An accurate benchmark of novel Edge AI platforms and milliwatt microcontrollers

Dr. Michele Magno. ETH Zurich. D-ITET Center for Project-based Learning
Credits: Marco Giordano, Philipp Mayer, Xiaying Wang.
Overview

- Introduction
- System Architecture
- Neural Net Deploy
- Experimental Benchmark
- Demo
Introduction

Miniaturized camera devices are today a commercial reality, widely used by:
- Surveillance
- Monitoring
- Controlling access

They rely on batteries with few hours of operation time and few of them are smart.

The wave of IoT is pushing the limit of battery-less devices and Tiny ML.
Internet of Things pushes AI and ML at the edge

The world is producing excessive amounts of "unstructured data" that need to be reconstructed

(IBM’s CTO Rob High)
“A PC will generate 90 megabytes of data a day, an autonomous car will generate 4 terabytes a day, a connected plane will generate 50 terabytes a day.”

Source: Samsung HBM

Bandwidth

Latency

Availability

Security

1 Billion cameras WW (2020)
30B Inference/sec

Communication latency also with 5G or other networks is in the range of hundred of milliseconds

50% of world at less than 8mbps Only 73% 3G/4G availability WW

Data traveling in the network are more vulnerable. Attacks to networks and communication towers

Source: IBM

Since 2015, roughly 2.5 Exabyte of data are being generated per day. Projection shows a 44 Zettabytes of data per day by 2020.
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
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
Overview

1. Introduction
2. System Architecture
3. Neural Net Deploy
4. Experimental Benchmark
5. Demo
System overview Always-on Smart Camera

Self-sustainability for neural network inference is a challenging task:
- A microcontroller has limited resources
  - 100-100KB RAM
  - M-ops or Best 1-10 GOps

Sensing: camera’s low power characteristics are preferable over high resolution
Himax HM01B0:
- Excellent low power capabilities: 3mA average power consumption at 320*240 pixel image
- Small form factor

Sensing: camera’s low power characteristics are preferable over high resolution
Himax HM01B0:
- Excellent low power capabilities: 3mA average power consumption at 320*240 pixel image
- Small form factor

Energy harvesters:
- With Maximum power Point Tracking
- ST VL53L1X for asynchronous wake up

Long range radio: low power preferred over high datarate.
Samtech SX1262:
- Up to 22dBm of transmitting power
- Several km range
- High current scalability: 30mA in TX mode, hundreds of nA in Sleep mode.

Sensing: ToF sensor used for asynchronous wake up
ST VL53L1X
System overview – working prototype

We realized a working prototype of a small always-on system with Long Range communication.

- Processor
- Camera Himax
- Time of Flight Sensor
- Flexible Solar panel
- LoRa Radio SX1262
- Energy Harvesting power management

We realized a working prototype of a small always-on system with Long Range communication.
New Trend is improving operations for Cycle: parallel + accelerators + computational efficient architectures

Evaluate the performance in terms of Latency, Computation power, Energy Efficiency, of below 200mW power platforms.
Overview

Introduction
System Architecture
Neural Net Deploy
Experimental Benchmark
Demo
The proposed Tiny Neural Network

**Goal:**
- design a neural network to support inference with limited resources

**Challenges:**
- Needed to reduce the input size
- Camera shoots greyscale images
- Quantized weights were used to optimize network size and speed up inference

**Expected Results:**
- More than 95% accuracy with 5 classes
- Around 50kB memory footprint
- Less than 30mJ per inference (For self-sustainability)
- less than few hundreds ms per inference
Data augmentation

Data augmentation represented an important step towards a successful training of the neural network:
- Used the open CelebA[1] dataset as a reference
- Only 20 to 30 images per actor are provided

To overcome this shortcoming:
- Images rotation (-20, -10, 0, +10, +20 degrees)
- Exposition alteration (gamma transform coefficient: 0.1, 0.4, 1, 2.5, 5)

Benefit:
- Better simulation of possible working conditions
- Improved generalization towards subject inclined, different light conditions

Confusion matrix – Float vs int8

The resolution of 30*40 has been chosen as the best tradeoff between the model’s memory occupancy and computing time over a very moderate loss of accuracy. The numbers in the confusion matrix represent the 5 different faces.

<table>
<thead>
<tr>
<th>Input size/data type</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>240x320, float (1)</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>240x320, int8 (2)</td>
<td>0.95</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>30x40, float (3)</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>30x40, int8 (4)</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Overview

Introduction
System Architecture
Neural Net Deploy
Experimental Benchmark
Demo
General Comparison: Power vs energy efficiency

Platforms power efficiency
- Single ARM Cortex-M4
- All Cores Active for min latency

Platforms energy consumption
- Single core energy
- Max Freq for min latency

Best
- SAMDS1 @48MHz: 5.34
- Apollo 3 @48MHz: 1.31
- Spresense @156MHz: 8.09
- PULP @250MHz: 0.52
- xCORE.ai @700MHz: 1.26
- MAX78000 @50MHz: 0.09
General comparison: Computational efficiency vs min Latency

Inference time:
- SAM D51 @48MHz
- Apollo 3 @48MHz
- Spresense single core @150MHz
- PULP single core @250MHz
- xCORE-al @250MHz
- MAX78000 @50MHz

Max Freq for min latency:

Inference performance:
- SAM D51
- Apollo 3
- Spresense
- Spresense single core
- PULP
- PULP single core
- xCORE-al
- MAX78000
- Cortex M55
Proof-of-Concept standalone smart camera

Assumptions:
- Trigger time: once per minute.
- Battery capacity: 8.64J.
- Energy per camera image captured: 0.5 mJ.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Energy per inference (mJ)</th>
<th>Battery Lifetime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apollo3</td>
<td>1.31</td>
<td>80h00’</td>
</tr>
<tr>
<td>Spresense</td>
<td>8.09</td>
<td>16h45’</td>
</tr>
<tr>
<td>PULP</td>
<td>0.52</td>
<td>140h15’</td>
</tr>
<tr>
<td>xCORE.ai</td>
<td>1.26</td>
<td>81h50’</td>
</tr>
<tr>
<td>MAX78000</td>
<td>0.09</td>
<td>244h00’</td>
</tr>
</tbody>
</table>
Overview

Introduction → System Architecture → Neural Net Deploy → Experimental Results → Demo
Video Demo (in preparation – just an example)
Conclusions - Battery-less long-range wireless smart camera

- Battery-less design with energy harvesting
- Tiny machine learning on the edge
- Efficient neural network for face ID
- Long range LoRa communication
- Two different implementation has been evaluated

- >95% accuracy over 5 faces
- Proposed neural network model fit in only 115kByte

- Benchmark of novel and promising processors below 100mW

- This work has been accepted in ENSSys workshop: A Battery-Free Long-Range Wireless Smart Camera for Face Detection

Thank you for your attention!
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.

Executive Sponsors
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
2. Supported by end-to-end tooling such as Arm Keil MDK
3. Connect to Runtime

Profiling and debugging tooling such as Arm Keil MDK

- 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

AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers
@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
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
- Embedded and edge compute deployment options

Impulse
- Test impulse with real-time device data flows

Test

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

A platform to scale AI across the industry

- **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

**Perception**
Object detection, speech recognition, contextual fusion

**Reasoning**
Scene understanding, language understanding, behavior prediction

**Action**
Reinforcement learning for decision making

IoT/IoT
Automotive
Mobile
Edge cloud
Cloud

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
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™ 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
Part of your life. Part of tomorrow.

www.infineon.com
Add Advanced Sensing to your Product with Edge AI / TinyML

Reality AI®

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Pre-built 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
Gold Sponsors
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
Silver Sponsors
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