AI @ IoT

Didier DONSEZ
Université Grenoble Alpes - Grenoble INP
LIG ERODS-DRAKKAR & Polytech Grenoble

Quick reminder about ML on DL

Supervised

- Training on labelled dataset
- Linear models, Tree-based models, ((Very) Deep) Neural Networks

Unsupervised

- Clustering, Association (Apriori), Matrix Profile (for time-series) …

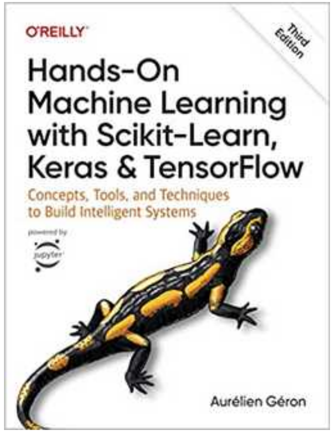
Frameworks

- Keras, TensorFlow, PyTorch, [MindSpore]() …

Studios and MLOps

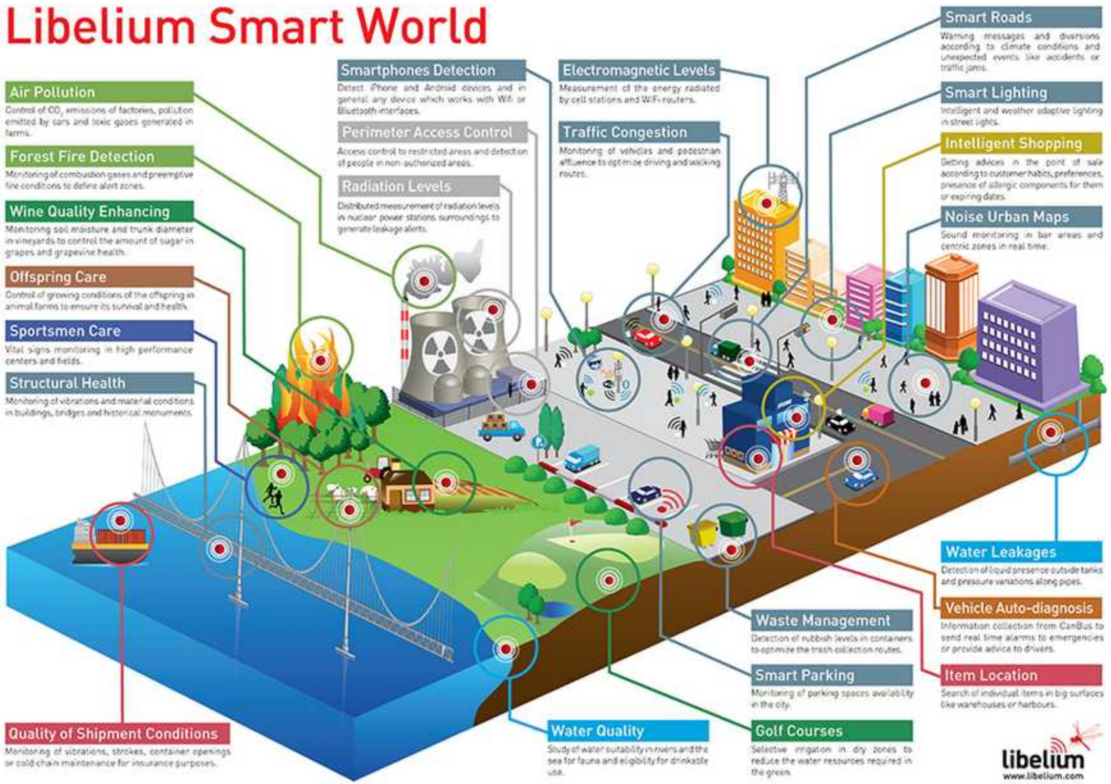
- Jupyter, [Google Colab](), [Edge Impulse]() …

EDGE IMPULSE



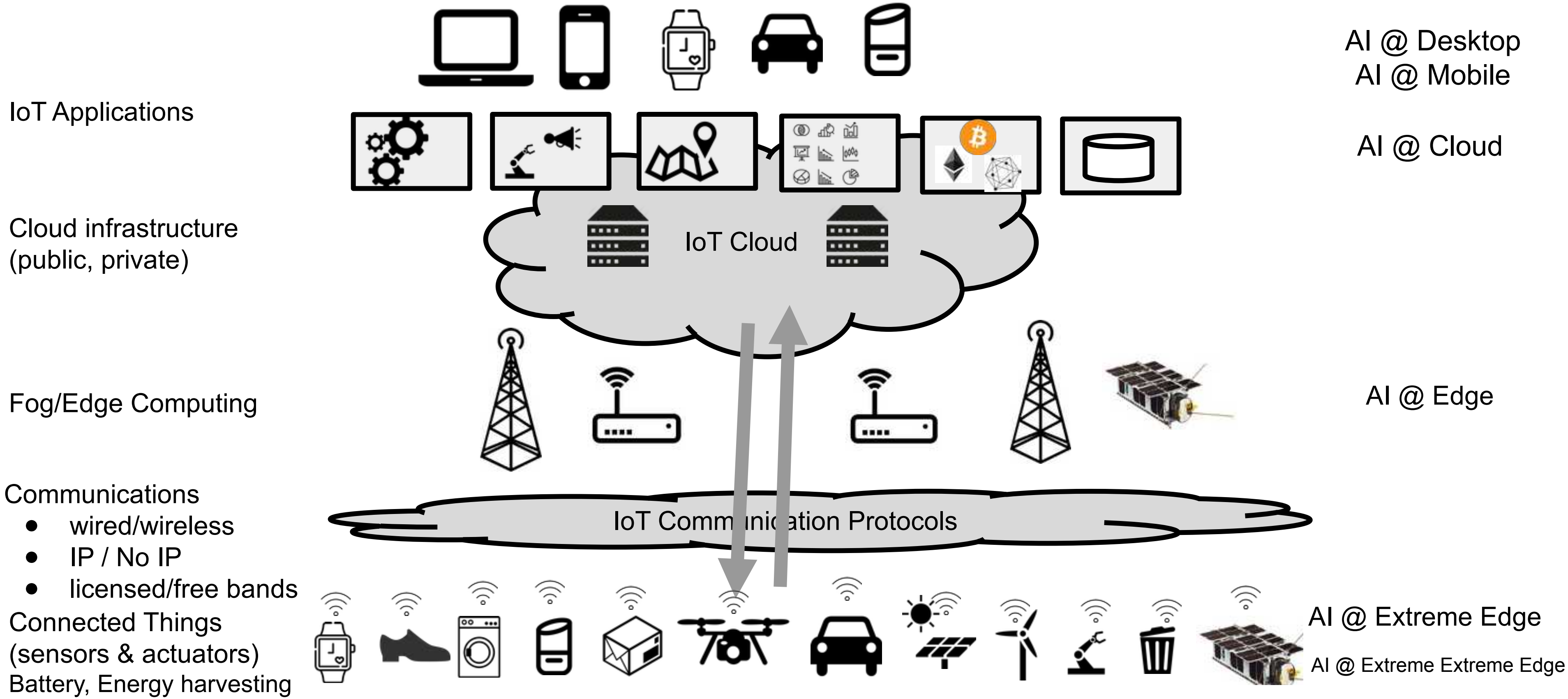
https://github.com/ageron/handson-ml2

Des objets connectés omniprésents (IoT)



https://www.libelium.com/success-stories/

IoT Infrastructure and AI



icons from https://thenounproject.com/, https://www.flaticon.com/free-icon/

IoT Data movement

Energy consumption

Privacy, confidentiality, sovereignty

Availability, Resilience

IoT and energie(s)

3.6V 2600mAh

Sample time

600

Seconds

Sensor type

EMS

Select Elsys sensor

Number of batteries

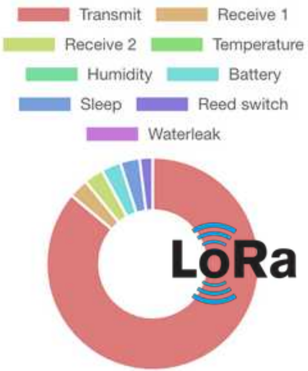
◉ 1 ○ 2

Capacity: 2700 mAh

Spreading factor

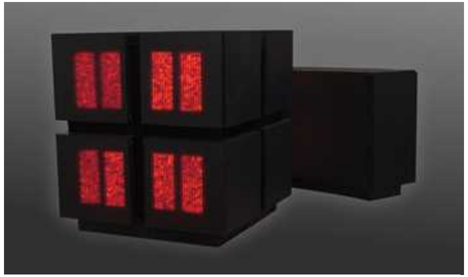
○ SF7 ○ SF8 ○ SF9 ○ SF10 ○ SF11 ◉ SF12

The battery will last for 1.6 years with an average current of 158 uA*.



Name	Time	Per hour	Current	Battery use
Transmit	1 650 ms	6	50 000 uA	87%
Receive 1	244 ms	6	12 000 uA	3%
Receive 2	200 ms	6	12 000 uA	3%
Temperature	5 ms	6	1 500 uA	0%
Humidity	5 ms	6	1 500 uA	0%
Battery		continuously	4 uA	3%
Sleep		continuously	4 uA	3%
Reed switch		continuously	3 uA	2%
Waterleak	100 ms	6	400 uA	0%

Les précurseurs : Thinking Machines
Connection Machines CM1 (1986) and CM2 (1987)



- 65,536 1-bit processors
- 512 MB (CM-2)
- Up to 80 GB with eight dataVaults
- Programming language: [Lisp]

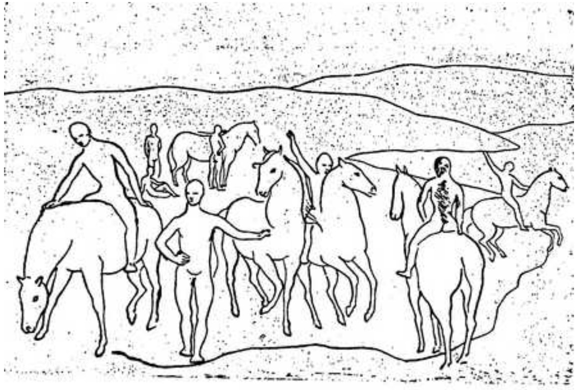


Figure 1.1: *The Watering Place*, Pablo Picasso, 1905

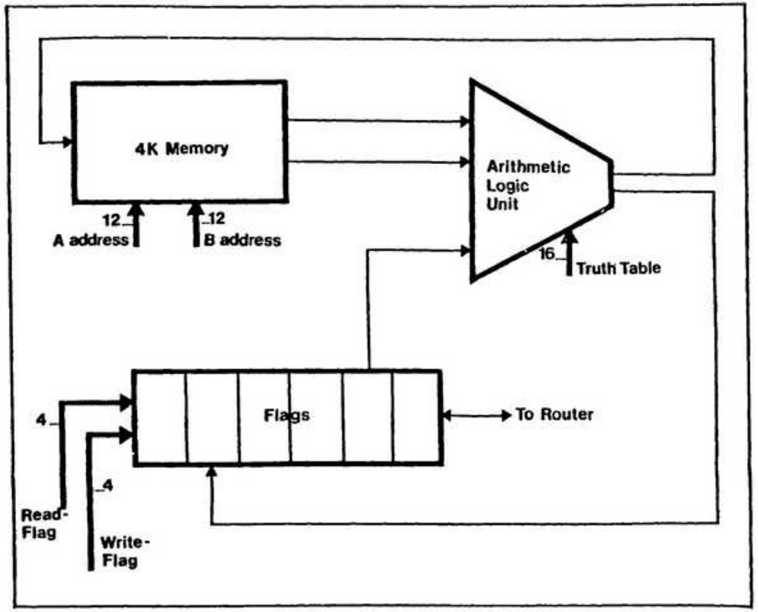
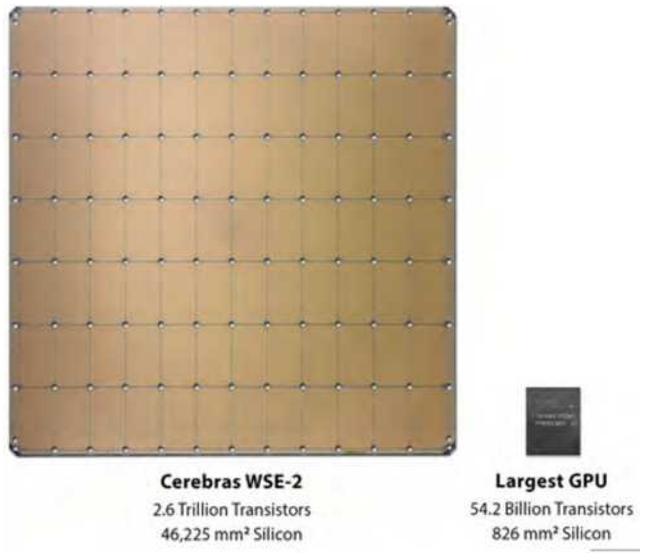


Figure 4.1: Block diagram of a single Connection Machine processing element

The connection machine : Hillis, W. Daniel, PhD MIT 1985

https://dspace.mit.edu/bitstream/handle/1721.1/14719/18524280-MIT.pdf

Hiperf ML & DL



GPU Processors

- NVidia GPU A100 (@ 250-400W), H100/H200 …
- …

AI Processors & Accelerators

- Google TPU v4 (275 TFLOPS FP16 @ 170 W)
- Huawei IA Ascend 910 (320 TFLOPS FP16, 640 TOPS INT8 @ 310 W)
- Amazon Trainium (380 INT8 TOPS, 190 FP16/BF16/cFP8/TF32 TFLOPS, and 47.5 FP32 TFLOP.)
- …
- Celebras Wafer Scale Engine (@ **20kW**) for ~1Meuros

Others

- Processing-in-Memory (UpMem) → DRAM + DPU

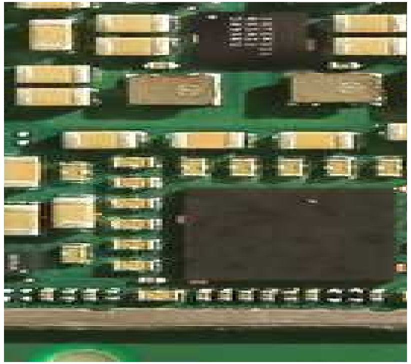
Edge ML & DL

GPU Processors

- Jetson Orin Nano : 40 TOPS @ 5-10W
- ARM Mali GPU in Cortex-A53 (ARM NN)

AI Processors

- Google Coral Edge TPU : 4 TOPS @ 2 W
    - MobileNet V2 model : 400 images per second
    - Asus AI Accelerator PCIe Card (8 Google Coral Edge TPU for 36-52 watts)
- Intel Movidius Myriad X VPU : 4 TOPS @ 1.5 W
- **Qualcomm Networking Pro A7 Elite (NPU)** : 40 TOPS @ ??



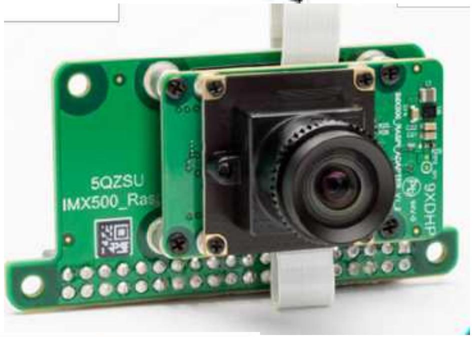
https://ai-benchmark.com/ranking_processors.html
https://ai-benchmark.com/ranking_4_0_3.html
https://ai-benchmark.com/ranking_IoT.html

Extreme extreme edge : l’IA dans les MEMS

ST IMU ISM330BX

accelerometer, gyroscope

sensor fusion low-power (SFLP) algorithm



Sony IMX500

Input tensor size : 64(H)×48(V) to 640(H)×480(V)
int8 or uint8, TensorFlow Lite
8388480 bytes for firmware network weight file, and working memory

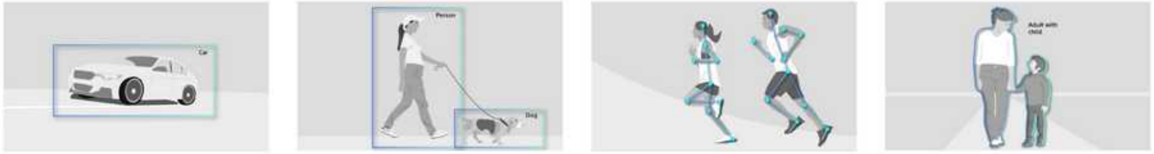


Image classification | Object detection | Pose detection | Semantic image segmentation

Hardware accelerators

Vector processors with ALU for quantized datatypes

Floating point

- 32 and 16-bit Floating Point (FP32 / FP16)
- Tension Float-32 (TF32)
- Brain Floating Point (BFloat16)
- 8-bit Floating point with configurable range and precision
  - cFP8, FP8, ms-FP11, ms-FP8
- 4-bit Floating point (FP4)

Integer

- INT8, INT16, INT32
- Unsigned 8-bit integer (UINT8)

ARM CMSIS-NN

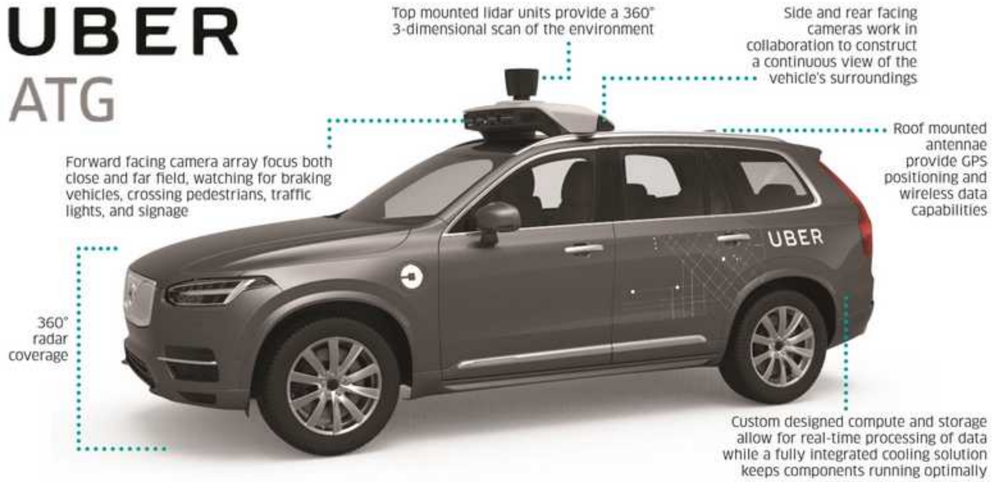
DSP extension, M-profile Vector Extension (MVE)

Operator	C int8	C int16	DSP int8	DSP int16	MVE int8	MVE int16
Conv2D	Yes	Yes	Yes	Yes	Yes	Yes
DepthwiseConv2D	Yes	Yes	Yes	Yes	Yes	Yes
Fully Connected	Yes	Yes	Yes	Yes	Yes	Yes
Add	Yes	Yes	Yes	Yes	Yes	Yes
Mul	Yes	Yes	Yes	Yes	Yes	Yes
MaxPooling	Yes	Yes	Yes	Yes	Yes	Yes
AvgPooling	Yes	Yes	Yes	Yes	Yes	Yes
Softmax	Yes	Yes	Yes	Yes	Yes	No
LSTM	Yes	NA	Yes	NA	Yes	NA

Application : Self Driving Car/Drone

Sensor data fusion

- GNSS (RTK)
- Radar (24 GHz Monolithic Microwave Integrated Circuit (MMIC))
- LiDar
- Thermal cam
- Visible cam



Application : Robotique

Des bergers pour des troupeaux de robots …

DJI + Movidius VPU

VITIROVER https://www.vitirover.fr/
vignes, vergers, voies ferrées,
fermes photovoltaïques

Application : Sécurité des travailleurs (IIoT)

- Detect motions like helmet wear on / take off, falling down, man-down, head impact, etc

Robert Bosch’s IoT Module GCY 500-1

Helmet | Safety Goggles | Safety reflective jacket | Multiple labels

https://dati-plus.com/

https://docs.edgeimpulse.com/experts/worker-safety-monitoring

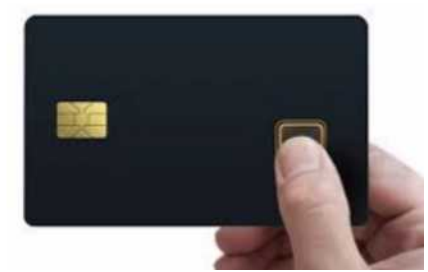
Application : Sport, Santé, Personne fragile

Amazon Halo Band
https://foundation.mozilla.org/fr/privacynotincluded/amazon-halo-band/

Semtech LR1110 tracker
(pet, cattle, wild animals …)

Arduino Nicla Sense
https://sites.arduino.cc/k-way-project

Roche glucometer and insuline pump
4 millions de diabétiques en France

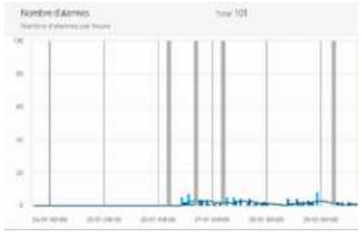


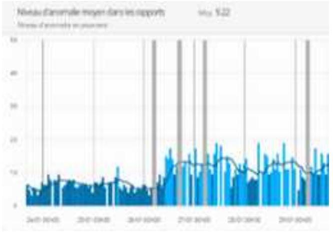
Application : Maintenance Préventive (IIoT)

- Surveillance
  - Convoyeurs, Moteurs, Canalisations, Ventilations ...
  - Vibration, ultrasons, température, pression, niveau ...
- Prévention de pannes
  - Remplacement des pièces avant interruption de service de l'équipement

nKe Bob sur un convoyeur

Adeunis
Delta P





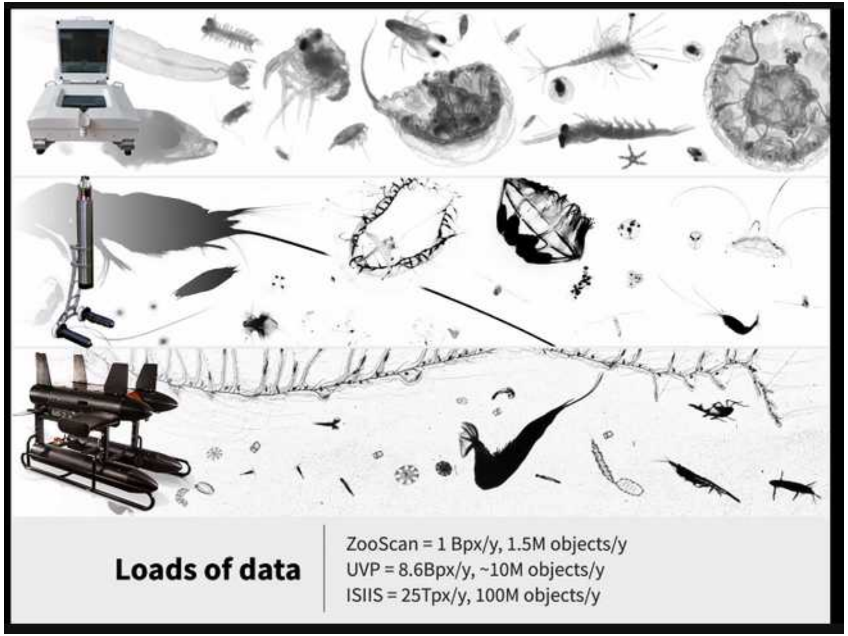
vibrations enregistrées sur 2 moteurs d'agitation dans une station d'épuration d'eau

32000 convoyeurs @ CDG

https://www.berger-levrault.com/fr/parole-d-expert/berger-levrault-editeur-integrateur-de-solutions-disruptives/
https://www.carl-software.fr/alstef-maintenance-sites-aeroportuaires/

Application : Ecologie

Bio loggers (video, photo, audio, motion …)

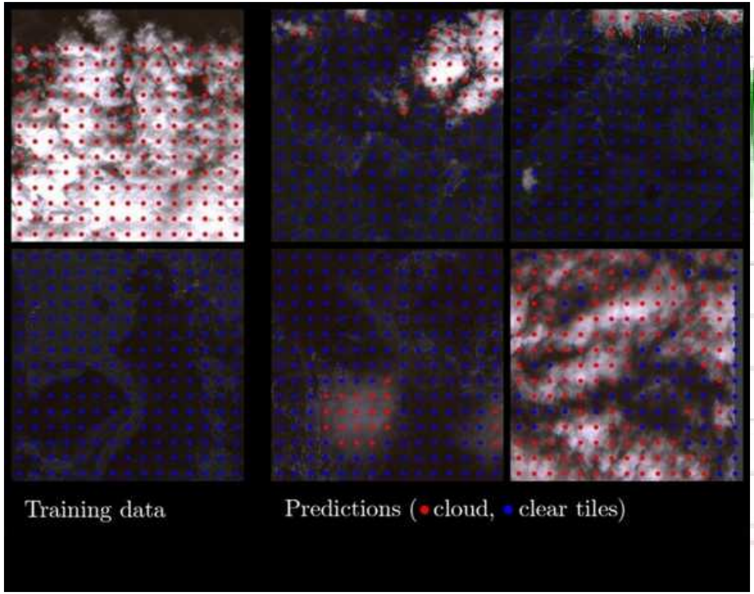


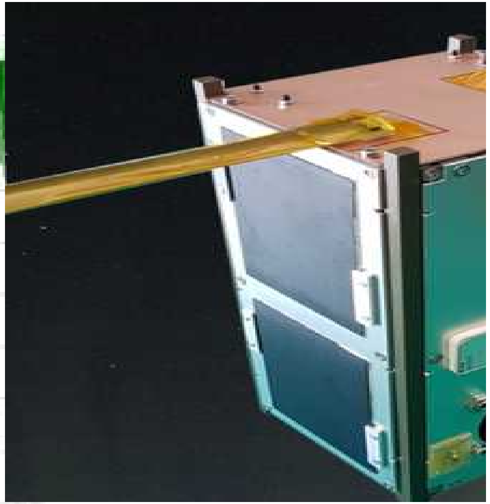
Semtech LR1110 tracker
(pet, cattle, wild animals …)

Application : Domaine Spatial

Space imaging (Visible, IR, X-Ray …), Satellite Outlier detection, …

Exemple: CSUG's QIverSat, ION SCV004, ITU.dk Discosat , CNES AeroSat …





source: https://www.ox.ac.uk/news/2023-07-28-researchers-successfully-train-machine-learning-model-outer-space-first-time, https://arxiv.org/pdf/2307.08700.pdf

TinyML

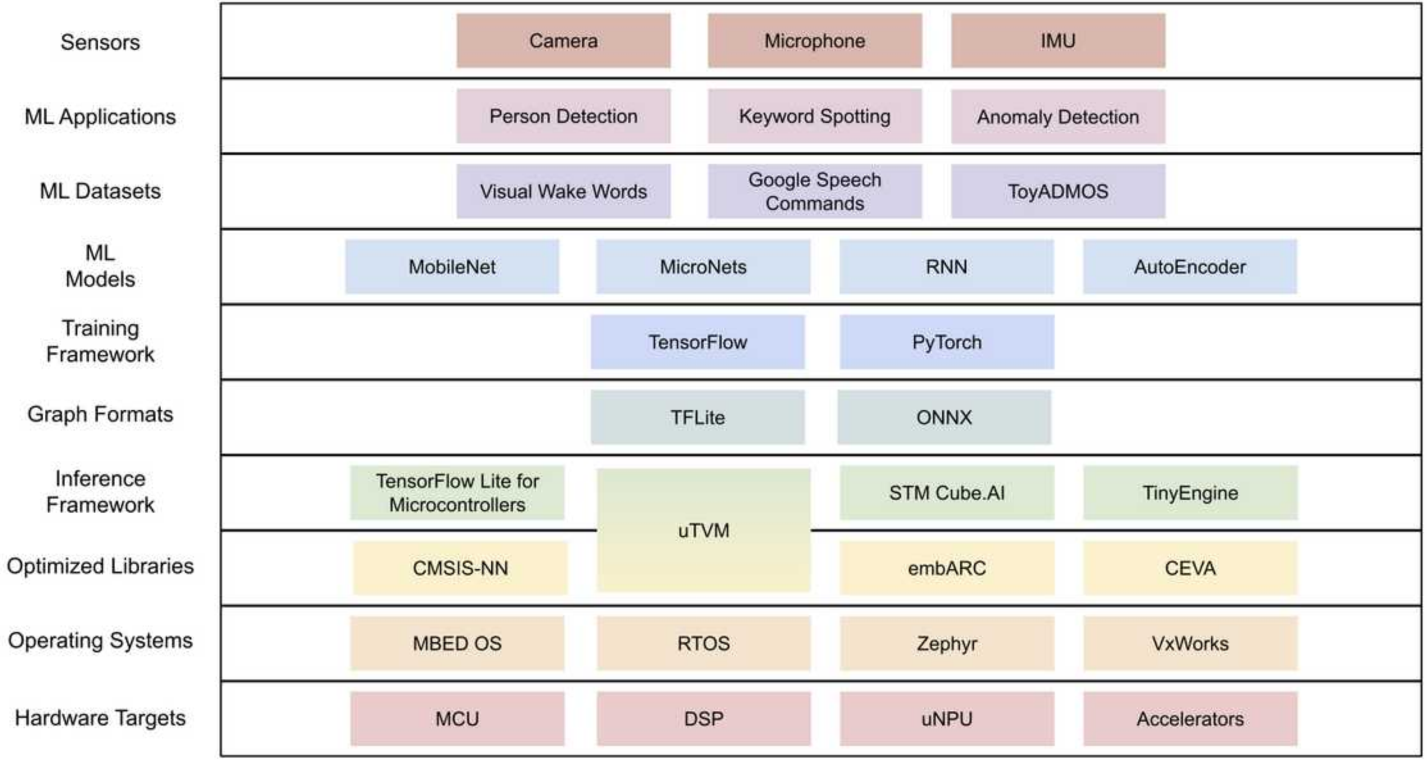
ML and DL on low-power (~1 mW) MCUs, DSP, FPGA, AI accelerators

Challenges for inference and for On Device Learning (ODL).

- Fragmented MCU market (heterogeneity)
  - ISAs (ARM Cortex M, RISC-V, ESP32, x86 …)
  - w/o extensions (DSP, ARM CMSIS-NN, ESP-NN, RISC-V NN …)
    - required specific optimizations
- SRAM (64KB to 1.5MB), FlashRAM (128KB to 8MB), w/o FPU, w/o File System
- Cost by unit (< 10 USD)
- standard tools and frameworks for portability
- benchmarks for comparison
- …

https://arxiv.org/abs/2010.08678

Tiny ML stack



https://arxiv.org/pdf/2106.07597

TinyML Applications

- Predictive maintenance (outlier detection …)
- Wake word (Hey Google ! Alexa !)
- Activity detection (parkinson, alzemier, cattle, pet …)
- Privacy-friendly security camera
- Traffic counting (vehicle, pedestrian, animals …)
- Person/Worker Safety (Medical mask/Hardhat detection)
- Biologger (video, photo, audio, motion …)
- …

https://arxiv.org/abs/2010.08678

Tensorflow Lite Micro (aka TF Micro)

Tensorflow DNN for tiny MCU and DSP
- *low power* CPU, w/o FPU, few RAM …

Design
- 130 operations instead of 1400 for TF
  - default implementations
  - platform-optimized operator implementations (ie CMSIS-NN)
    - operator implementations can exploit multiple cores
- list of TFL operations (no DAG)
  - operations are interpreted at runtime
- optimization for memory and latency
  - quantization (float32 → int8)
  - memory management (bin-packed arena)
- multi-tenancy (*multiple models*)
- thread-safe

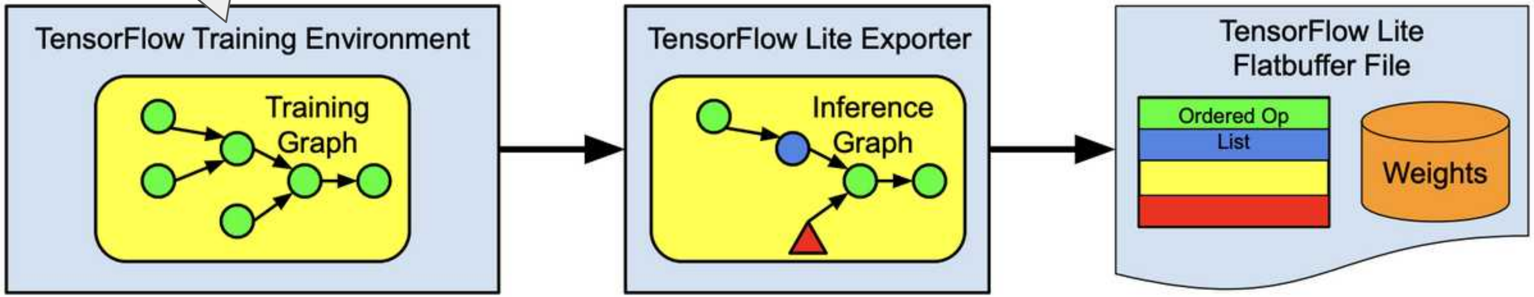
https://arxiv.org/abs/2010.08678

TF Micro workflow

TensorFlow

Transfert learning
(pretrained models)

lib C (weight + code)



ie jupyter notebook on GPUs, collab, edge impulse …

optimization
f32->i8 quantization,
pruning, dag → list …

runtime: list interpretation

AOT compilation

https://arxiv.org/abs/2010.08678

Optimization : Quantization and Pruning

Goal : reduce the number of parameters and the memory size, as well as the computational complexity of the network

Quantization

- float32 → bool, int8, int16, int32, fp16, bfp16, fp8, fp4 …
- logarithmic, half-wave gaussian, Power-of-Two (PoT for FPGA)

Pruning

- zeroes very small weights

Compression techniques available out-of-the-box in Edge Impulse include quantization and activations (Jacob et al., 2017) and operator fusion (Goo, 2022)

Quantization in practice

Post-training Quantization (PTQ)

- train the model using float32 weights and inputs, then quantize the weights. Its main advantage that it is simple to apply.
- Downside is, it can result in accuracy loss.
- transfert learning from float32 trained model

Quantization-aware training (QAT)

- quantize the weights during training. Here, even the gradients are calculated for the quantized weights.
- When applying int8 quantization, this has the best result, but it is more involved than the other option.
- No transfert learning from float32 trained model ?

Quantization in TF Lite

Model	Top-1 Accuracy (Original)	Top-1 Accuracy (Post Training Quantized)	Top-1 Accuracy (Quantization Aware Training)	Latency (Original) (ms)	Latency (Post Training Quantized) (ms)	Latency (Quantization Aware Training) (ms)	Size (Original) (MB)	Size (Optimized) (MB)
Mobilenet-v1-1-224	0.709	0.657	0.70	124	112	64	16.9	4.3
Mobilenet-v2-1-224	0.719	0.637	0.709	89	98	54	14	3.6
Inception_v3	0.78	0.772	0.775	1130	845	543	95.7	23.9
Resnet_v2_101	0.770	0.768	N/A	3973	2868	N/A	178.3	44.9

Comparison of quantization methods in TensorFlow Lite for several convolutional network architectures. Source: TensorFlow Lite documentation

Quantization in TF Lite

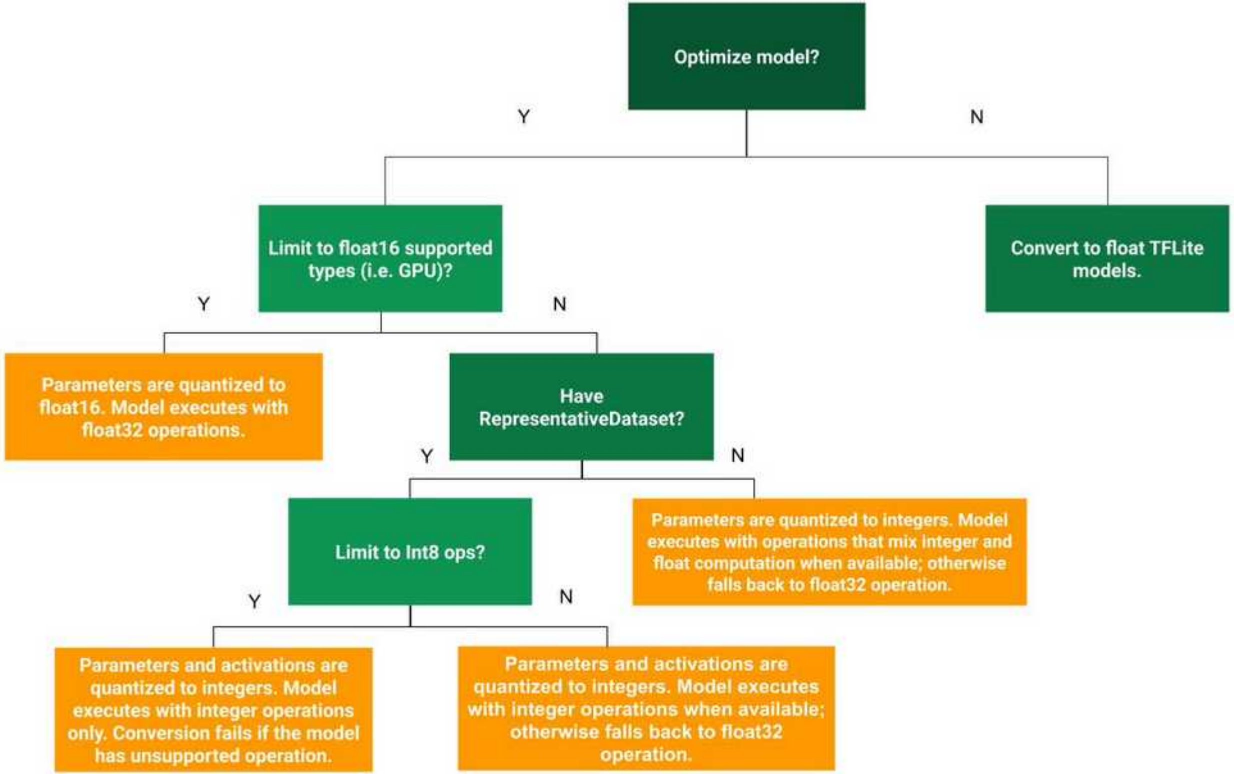
Technique	Benefits	Hardware
Dynamic range quantization	4x smaller, 2x-3x speedup	CPU
Full integer quantization	4x smaller, 3x+ speedup	CPU, Edge TPU, Microcontrollers
Float16 quantization	2x smaller, GPU acceleration	CPU, GPU

Image classification with tools

Model	Non-quantized Top-1 Accuracy	8-bit Quantized Accuracy
MobilenetV1 224	71.03%	71.06%
Resnet v1 50	76.3%	76.1%
MobilenetV2 224	70.77%	70.01%

https://www.tensorflow.org/lite/performance/post_training_quantization

Quantization decision tree



https://www.tensorflow.org/lite/performance/post_training_quantization

RAM and ROM Optimizations

Operator implementations pruning (`#define`, `#ifdef`, `#endif`)

Remove (with macro) the operators implementations unused during the interpretation of the model(s)

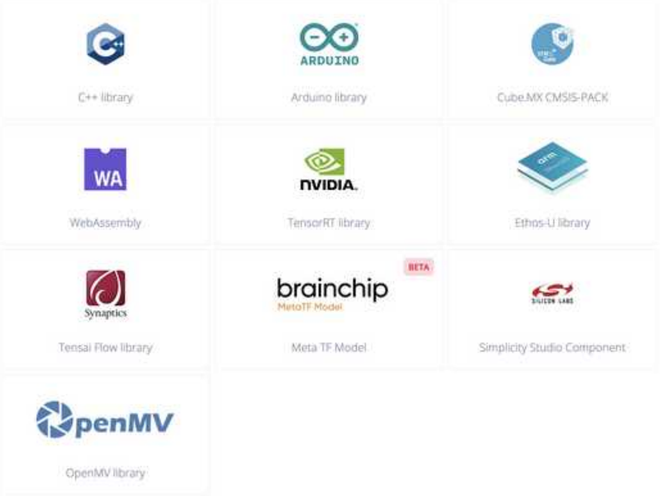
Interpreter-less Code Generation (Ahead-of-Time compilation)

- Edge Impulse's EON compiler
- cpetig/tflite_micro_compiler

TFLite model (secure) update over the air/space → U-TOE https://docs.edgeimpulse.com/docs/

Online **(Tiny)MLOps** platform for (re)training and tuning models fitting TinyML contraints (Low power/low RAM MCU or DSP)

Generate (C/C++/WASM) TF Micro libs (AOT) for most common MCU/DSP and eval boards (STM32, Sony, ESP32, Jetson …)



https://arxiv.org/pdf/2212.03332.pdf

Platforms/DevKits for TinyML

Brand new MCUs/DSP for Embedded (Very Low Power) AI

- M5 Stack / ESP32 v3 Cam, Maix Speed
- STM32 (such as Arduino Nicla Sense/Vision (H7))
- RPI Pico, Sony SPresense, Greenwaves
- Brainchip, DSP

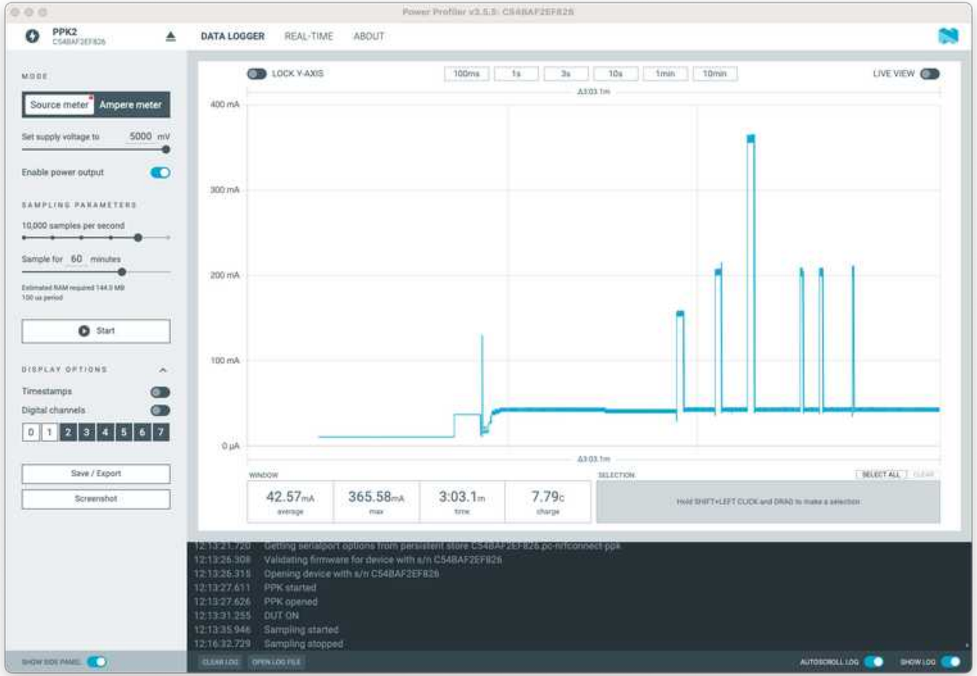
ISA Extensions

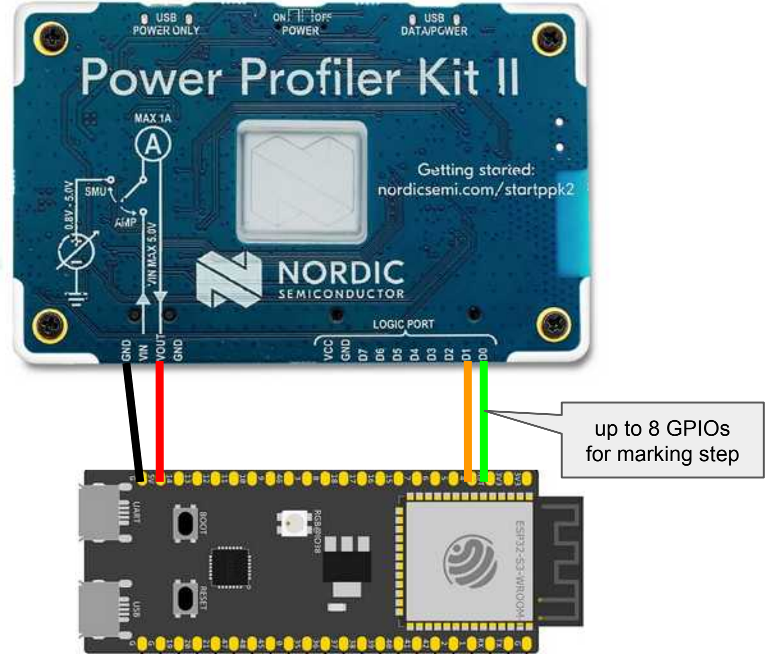
- ARM CMSIS-NN, ESP NN …


SenseCAP-A1101




Setup for monitoring energy consumption @ OD training / inference





https://github.com/CampusIoT/tutorial/tree/master/nrf-ppk2#readme
https://github.com/christophe-cerin/OnlineML_ESP32

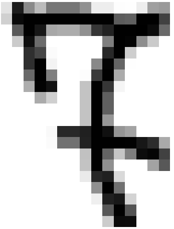
Demo TFLite (MINST) / RIOT OS / Nucleo F446RE

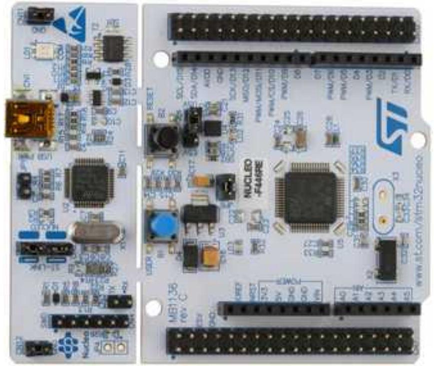
180 MHz, 225 DMIPS (Dhrystone 2.1)
128 KB SRAM 512 KB Flash

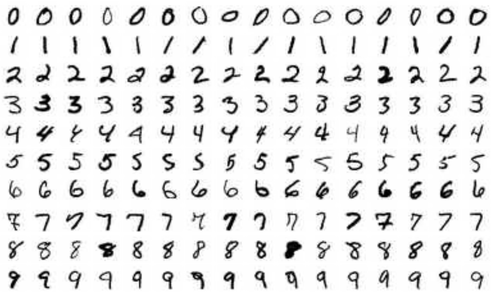
```
cd ~/github/RIOT-OS/RIOT
cd tests/pkg/tflite-micro
gmake BOARD=nucleo-f446re

ls -l external_modules/mnist/digit
784 external_modules/mnist/digit
ls -l external_modules/mnist/*.tflite
52920 model.tflite
ls -l bin/nucleo-f446re/*.bin
113292 tests_tflite-micro.bin
gmake BOARD=nucleo-f446re flash-only term

Help: Press s to start test, r to print it is ready
main(): This is RIOT! (Version: 2025.01-devel-8-g00e25)
Digit prediction: 7 (duration: 7008 usec - 1.755 DMIP)
```







References & Bibliography

TinyML https://www.tinyml.org

Repositories

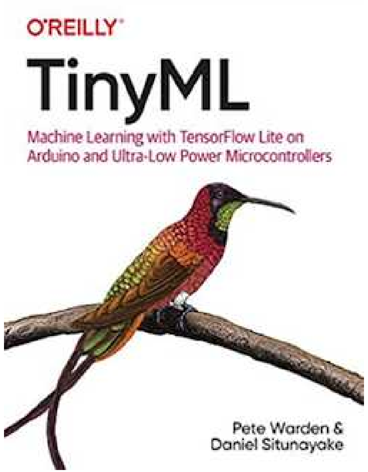
- https://github.com/tensorflow/tflite-micro
- https://github.com/tensorflow/tflite-micro-arduino-examples
- https://github.com/mlcommons/tiny

TinyML book https://tinymlbook.com/

TF Micro design https://arxiv.org/abs/2010.08678

A. Géron, Hands-On ML …

https://mastering-tinyml.github.io/



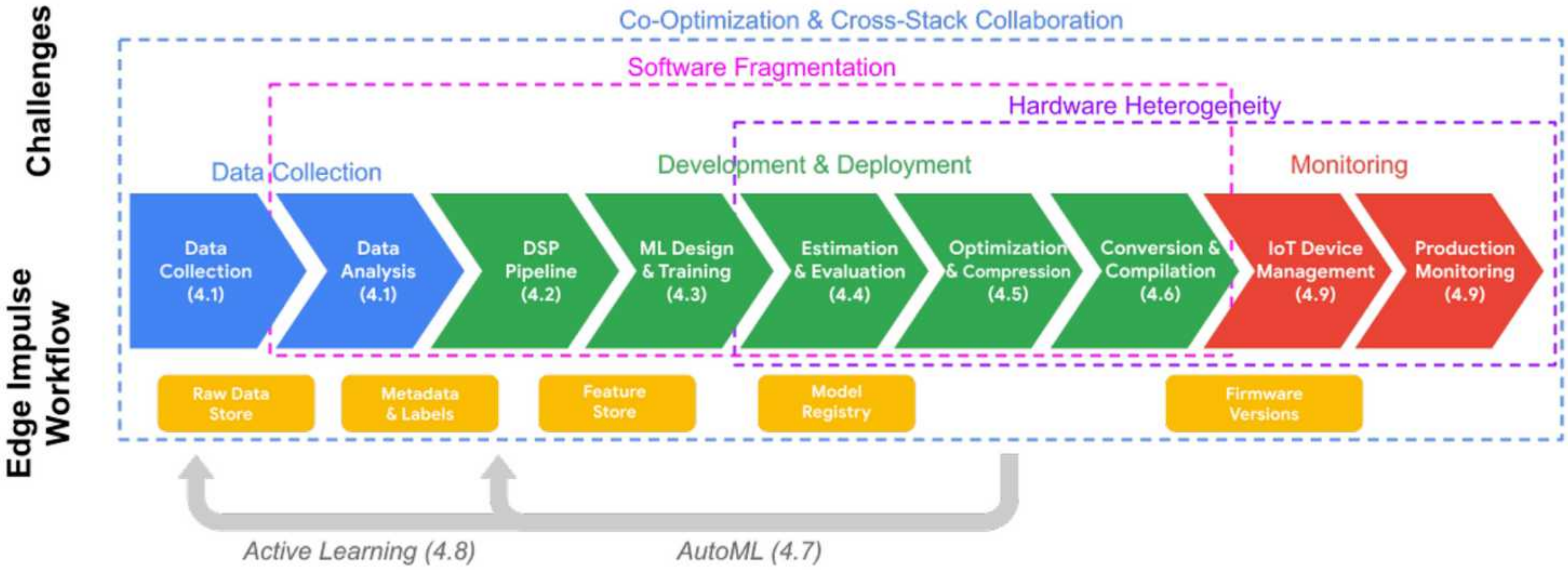
Autres

https://edgeimpulse.com/

https://github.com/Seeed-Studio/CodeCraft

https://github.com/mlcommons/tiny

TinyMLOps


Challenges
Edge Impulse Workflow
Co-Optimization & Cross-Stack Collaboration
Software Fragmentation
Hardware Heterogeneity
Data Collection
Development & Deployment
Monitoring
Data Collection (4.1)
Data Analysis (4.1)
DSP Pipeline (4.2)
ML Design & Training (4.3)
Estimation & Evaluation (4.4)
Optimization & Compression (4.5)
Conversion & Compilation (4.6)
IoT Device Management (4.9)
Production Monitoring (4.9)
Raw Data Store
Metadata & Labels
Feature Store
Model Registry
Firmware Versions
Active Learning (4.8)
AutoML (4.7)

ST X-CUBE-AI

TF Micro + STM32Cube.AI for STM32 MCU

	Model Stats.	TFLite Micro Runtime	STM32Cube.AI Runtime
NN Model	MACs: 81.8M Param: 0.74M Act: 333KB	RAM: 498KB Flash: 994KB	RAM: 321KB Flash: 738KB
Hardware Deployment	**STM32L4R9I** RAM: 640KB Flash: 2MB	**Latency: 7255ms**	**Latency: 3309ms**

Quantization

TABLE I: Accuracy [%] and Recall [%] comparison of the baseline not-quantized model against models quantized with different quantization parameters. Quantization operations were inserted on inputs to ops of type Conv2D and Add; and on outputs of activation operations.

Accuracy	Recall	Quantized	Quant. op	Rounding mode	Bitwidth	Input preprocessing	Output signedness	Locations of quantization op	Range given	Quantize delay [% steps]	Theoratical Inference Model Size
91.14	99.42	×	-	-	-	-	-	-	-	-	89.6 MB
52.90	94.51	✓	QDQv2	half_to_even	4	vgg	signed	Conv, Activation	×	80%	11.42 MB
86.27	99.22	✓	FakeQuant	half_to_even	8	vgg	signed	Conv, Activation	×	80%	22.95 MB
86.60	99.29	✓	QDQv2	half_to_even	8	inception	signed	Conv, Activation	×	80%	22.95 MB
88.06	98.99	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	✓	80%	22.95 MB
88.65	99.10	✓	QDQv2	half_to_even	8	vgg	unsigned	Conv, Activation	×	80%	22.95 MB
88.69	99.61	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	60%	22.95 MB
88.82	99.63	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	90%	22.95 MB
88.86	99.01	✓	QDQv2	half_up	8	vgg	signed	Conv, Activation	×	80%	22.95 MB
88.87	99.28	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation, Add	×	80%	22.95 MB
89.06	99.64	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	70%	22.95 MB
89.37	99.65	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	50%	22.95 MB
89.69	99.66	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	80%	22.95 MB
90.04	99.67	✓	QDQv2	half_to_even	8	vgg	signed	Conv, Activation	×	40%	22.95 MB
90.71	99.25	✓	QDQv2	half_to_even	32	vgg	signed	Conv, Activation	×	80%	89.6 MB
91.26	99.34	✓	QDQv2	half_to_even	16	vgg	signed	Conv, Activation	×	80%	45.90 MB

Deep Learning Optimization for Edge Devices:Analysis of Training Quantization Parameters
https://mostwiedzy.pl/pl/publication/download/1/deep-learning-optimization-for-edge-devices-analysis-of-training-quantization-parameters_40106.pdf