tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"State of the TinyML today"

Frédéric Pétrot – l'Université Pierre et Marie

Etienne Balit – Neovision

Loic Lietar - GreenWaves Technologies

France Area Group – July 1, 2021

www.tinyML.org

tinyML Talks Sponsors

TINY ML TALKS webcast

arm
tinyML Strategic Partner

Deeplite

EDGE IMPULSE
tinyML Strategic Partner

emza visual sense
tinyML Strategic Partner

GREENWAVES TECHNOLOGIES
tinyML Strategic Partner

LatentAI
Adaptive AI for a Smarter Edge
tinyML Strategic Partner

HOTG
tinyML Strategic Partner

maxim integrated™

Qeexo

Qualcomm
tinyML Strategic Partner

RealityAI®
tinyML Strategic Partner

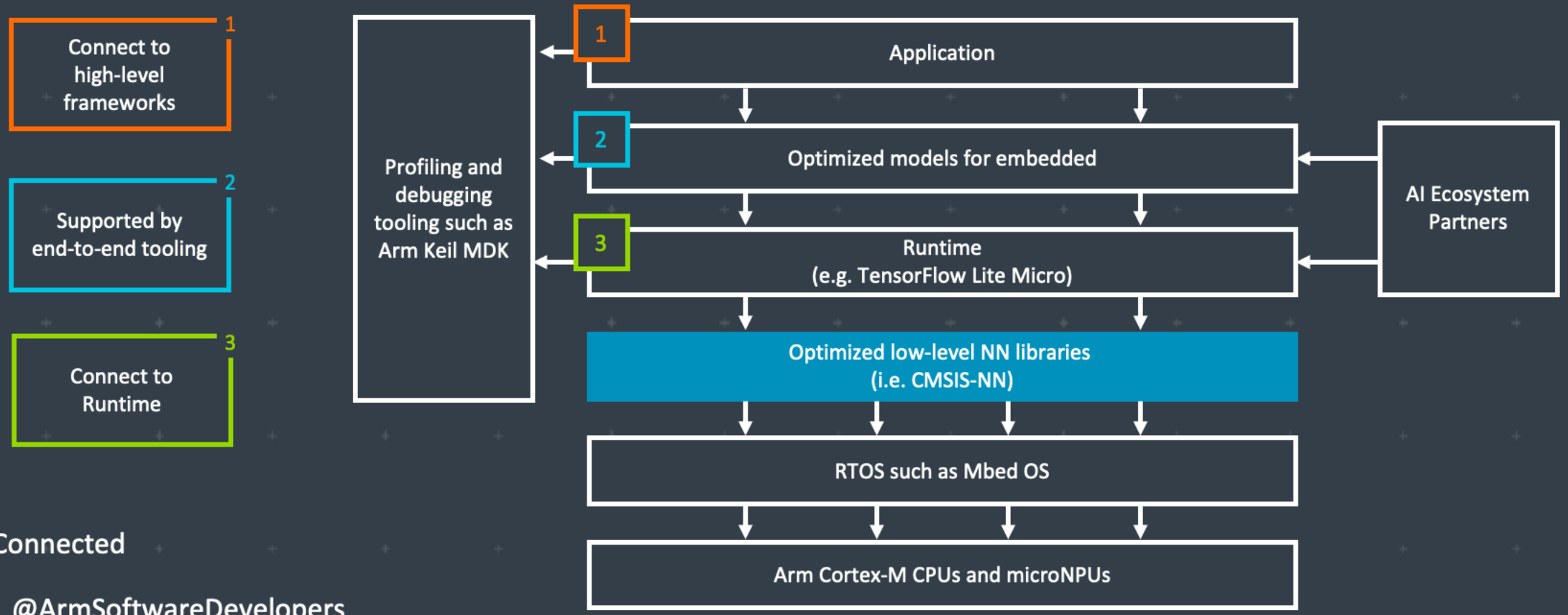
SensiML™
tinyML Strategic Partner

SynSense

SYNTIANT
tinyML Strategic Partner

Additional Sponsorships available – contact Olga@tinyML.org for info

Arm: The Software and Hardware Foundation for tinyML



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

arm

Deeplite

WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

Automatically compress SOTA models like MobileNet to <200KB with **little to no drop in accuracy** for inference on resource-limited MCUs

Reduce model optimization trial & error from weeks to days using Deeplite's **design space exploration**

Deploy more models to your device without sacrificing performance or battery life with our **easy-to-use software**

BECOME BETA USER **bit.ly/testdeeplite**

mobilityXlab arm AI 100 CBINSIGHTS

TinyML for all developers

C++ library

Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device

Real sensors in real time

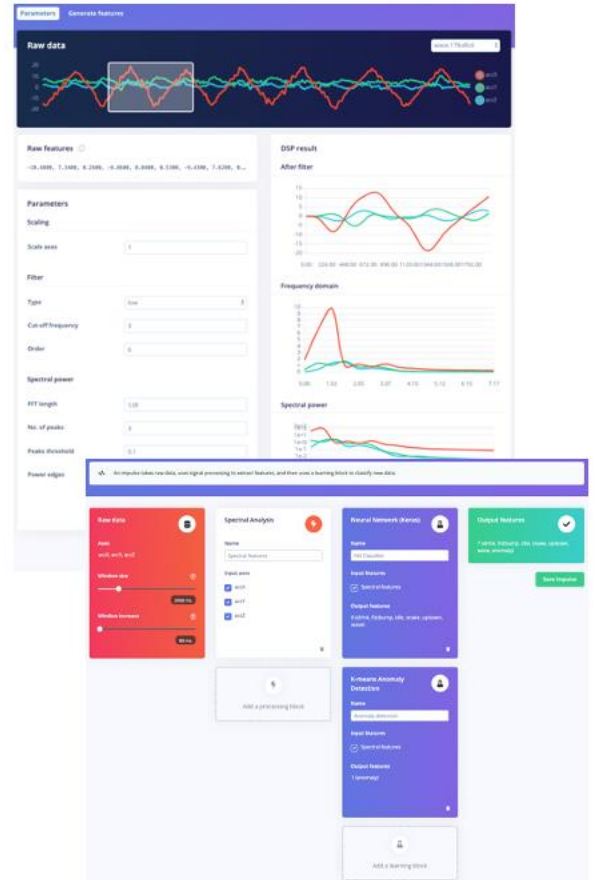
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test

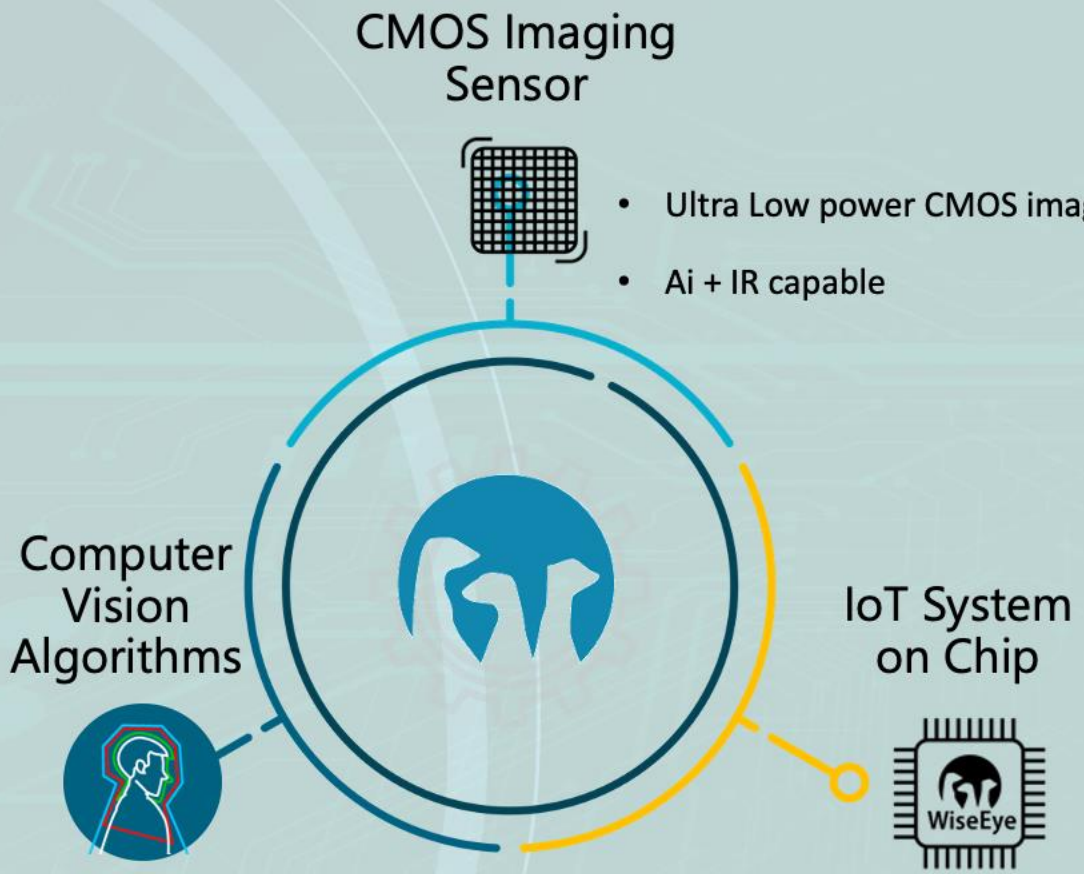


www.edgeimpulse.com

emza
visual sense

The Eye in IoT

Edge AI Visual Sensors



- Ultra Low power CMOS imager
- Ai + IR capable

- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

info@emza-vs.com

Enabling the next generation of Sensor and Hearable products to process rich data with energy efficiency

Visible Image

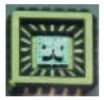
Sound

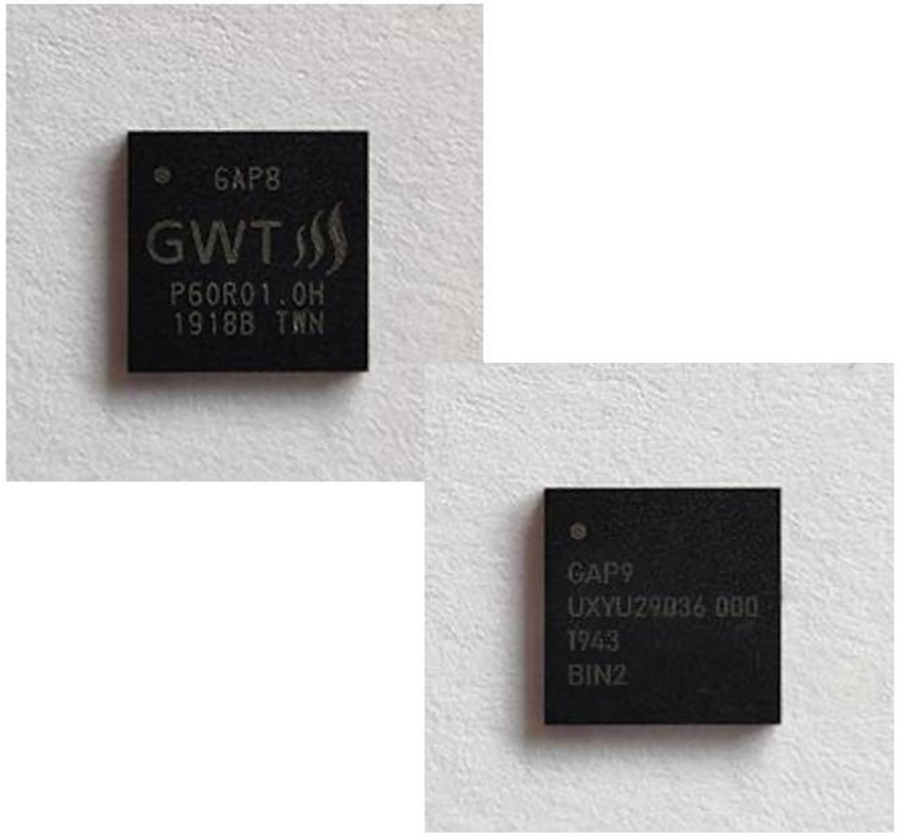
IR Image

Radar

Bio-sensor

Gyro/Accel

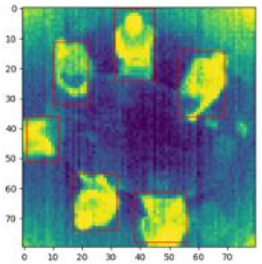




Wearables / Hearables

Battery-powered consumer electronics

IoT Sensors



GREENWAVES TECHNOLOGIES

LatentAI

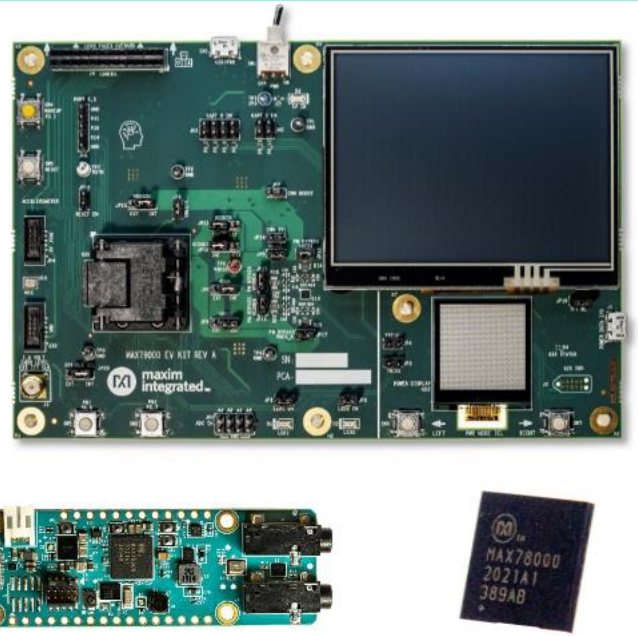
Adaptive AI for the Intelligent Edge

Latentai.com


HOTG

Maxim Integrated: Enabling Edge Intelligence

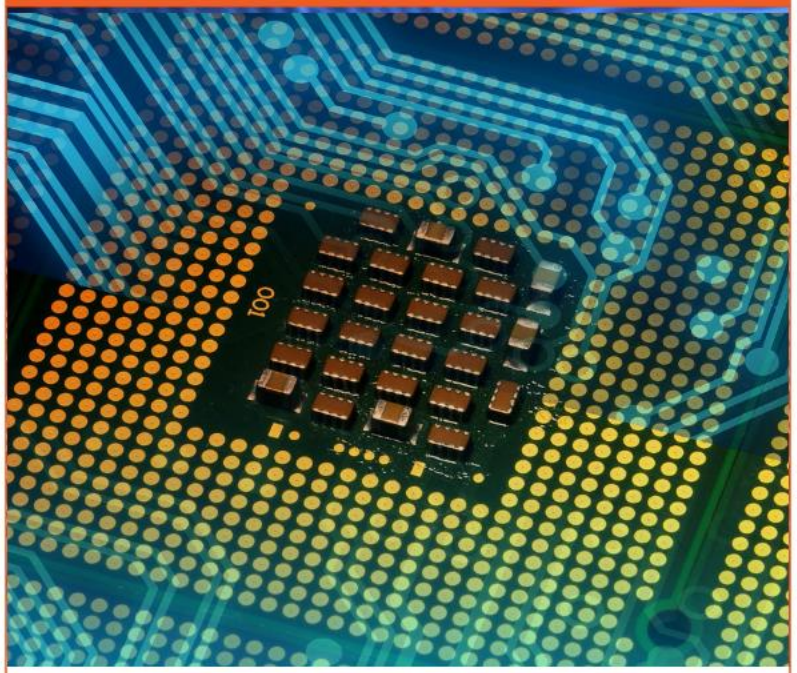
Advanced AI Acceleration IC



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

Low Power Cortex M4 Micros



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

Sensors and Signal Conditioning

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

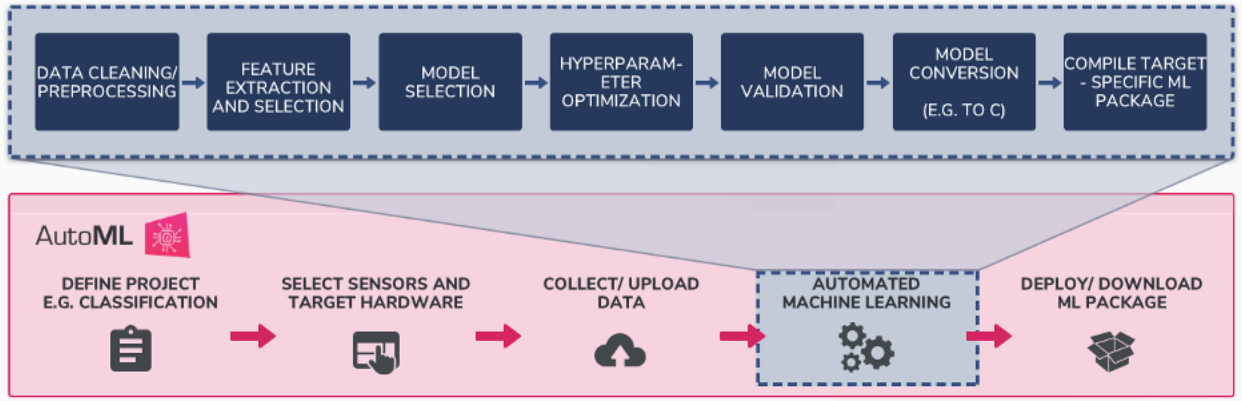
Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform



For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

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

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Reality AI®

Add Advanced Sensing to your Product with Edge AI / TinyML

https://reality.ai info@reality.ai @SensorAI Reality AI

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

SensiML™

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

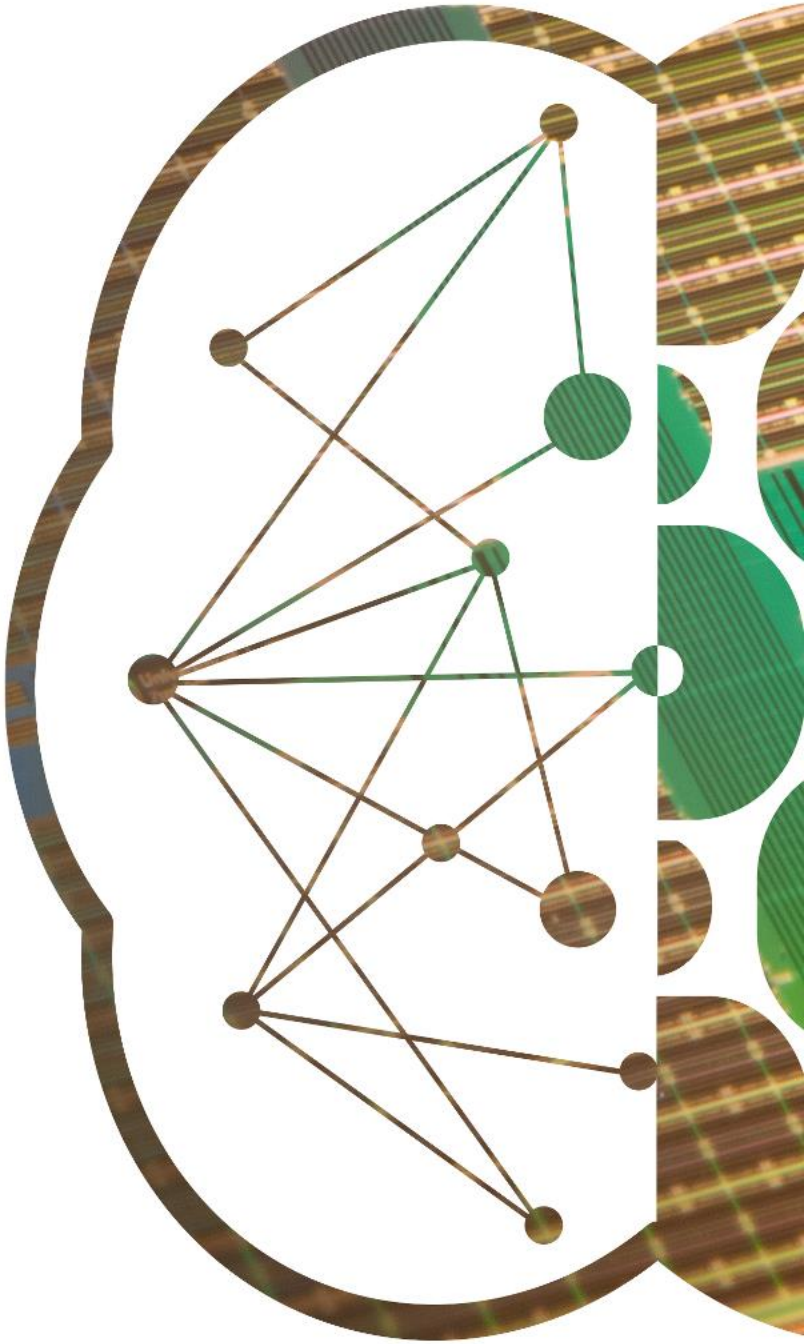
We enable the creation of production-grade smart sensor devices.

sensiml.com

SynSense

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



SYNTIANT

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

TINY ML FOUNDATION

collaboration with hackster.io

AN AVNET COMMUNITY

Focus on:
(i) developing new use cases/apps for tinyML vision; and (ii) promoting tinyML tech & companies in the developer community

Submissions accepted until August 15th, 2021
Winners announced on September 1, 2021 ($6k value)
Sponsorships available: *sponsorships@tinyML.org*

https://www.hackster.io/contests/tinyml-vision

Next tinyML Talks

Date	Presenter	Topic / Title
Tuesday, July 6	**Shivy Yohanandan,** Xailient	Cracking a 600 million year old secret to fit computer vision on the edge

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting

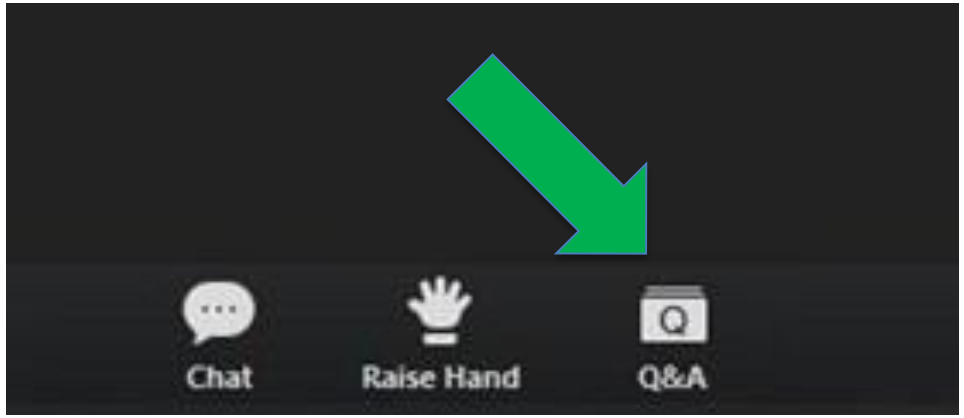
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions



TINY ML TALKS webcast

Frédéric Pétrot

Frédéric Pétrot holds a doctorate in computer science from Pierre et Marie University, Paris, France, since 1994. He joined Grenoble INP / Ensimag in 2004 as a professor. His work focuses on the design and implementation of integrated digital systems for general or specialized purposes, for example for the acceleration of artificial intelligence.

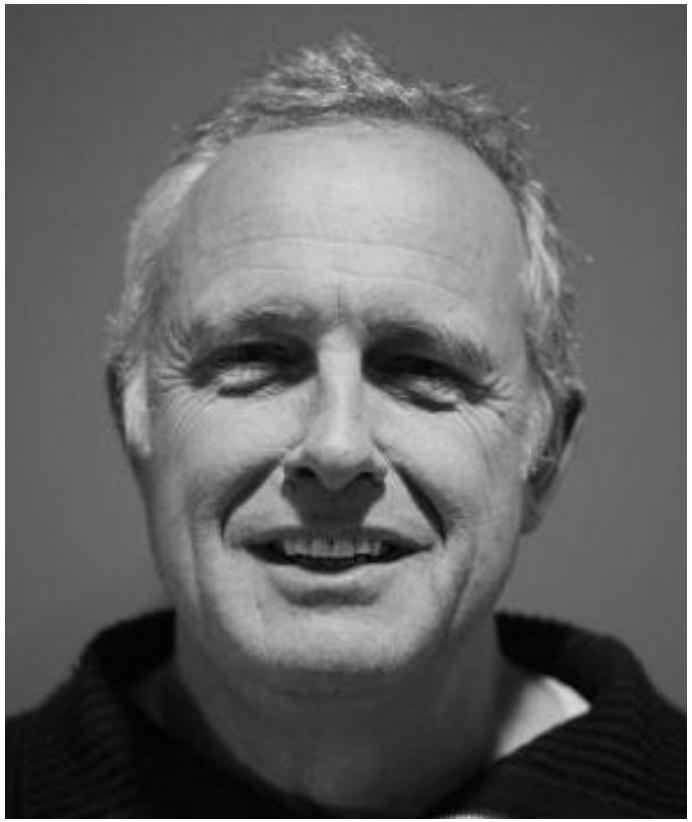
TINY ML TALKS webcast

Etienne Balit

Etienne Balit is R&D Director at Neovision. He obtained a master's degree in modelling for cognitive sciences at Ecole Normale Supérieure. He completed a thesis in computer science at Inria, where he developed advanced knowledge in artificial intelligence and particularly in neural networks. At Neovision, he coordinates the R&D team and monitors the latest scientific and technological advances to drive high-performance and state-of-the-art innovations.

TINY ML TALKS webcast

Loic Lietar



Loïc is a co-founder and the CEO of GreenWaves Technologies. GreenWaves Technologies, a fabless semiconductor company, designs disruptive ultra-low power embedded solutions for interpreting and transforming rich data sources such as images, sounds, radar signatures and vibrations using AI and signal processing in highly power-constrained devices such as hearables, wearables and IoT sensors. Prior to this, Loic worked 25 years for ST where he led several product divisions, has been the Chief Strategy Officer and co-founded and managed ST's corporate venture fund. Loïc has been an active business angel for the last 8 years. He has also been president of Minalogic, the French cluster for semiconductors.

TinyML

For which uses cases, today and tomorrow?

Etienne Balit
Head of R&D

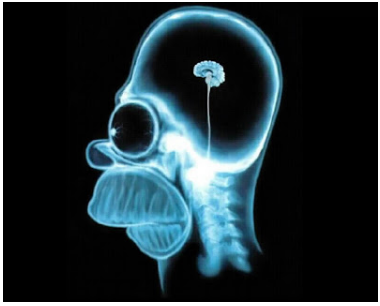
Some thoughts about low-power neural networks

Frédéric Pétrot
Univ. Grenoble Alpes, CNRS, Grenoble INP*,
TIMA, F-38000 Grenoble, France
tima.imag.fr/sls/people/petrot
frederic.petrot@univ-grenoble-alpes.fr
*Institute of Engineering Univ. Grenoble Alpes

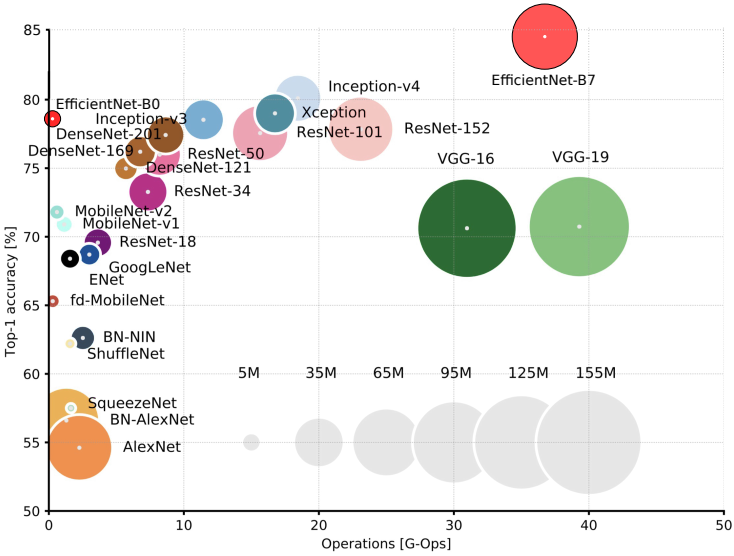
Multidisciplinary Institute In Artificial Intelligence

What AI do we really need?

What AI do we really need?



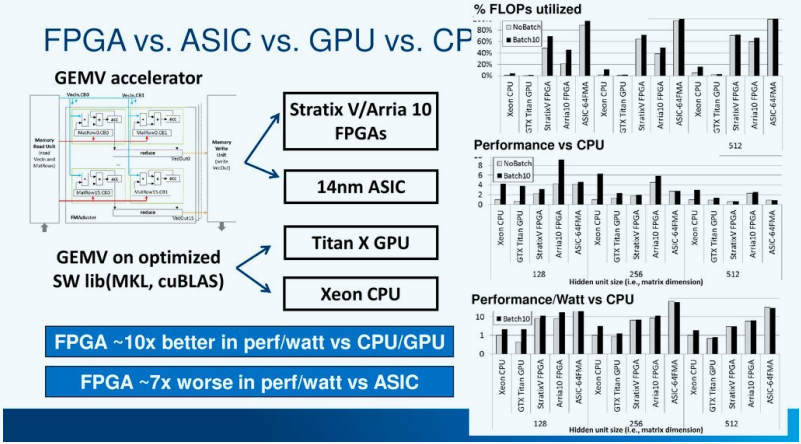
CNN Models: Accuracy, Operations and Weights



A. Canziani, E. Culurciello, A. Paszke, *"An Analysis of Deep Neural Network Models for Practical Applications"*, 2017 (EfficientNet-B0/B7 added by myself)

Hardware Accelerated Neural Network

What's the interest?



E. Nurvitadhi, J. Sim, D. Sheffield, A. Mishra, S. Krishnan, D. Marr, Intel Labs

Hardware Accelerated Neural Network

What's the interest?

Optimize server side AI

- Energy
  Minimize TCO for AI workloads
  Greener AI for social acceptance
- Throughput
  Enhance job throughput at constant energy budget

Local computation possible!

- Energy
  No router, cloud server, ...
  $\Rightarrow$ Huge constraint in Edge Computing
  $\Rightarrow$ Worse in IoT
  $\Rightarrow$ Transmitting data costs energy
- Latency
  Immediate response, no dead zone, no network reliability issue, ...
- Privacy/security
  No storage in someone else's servers
  Neither wire nor wireless sniffing possible

Hardware Accelerated Neural Network

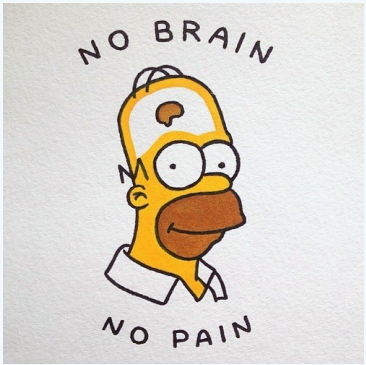
What are the constraints?

- Accuracy needs depend on the application
- Silicon resources:
  $\Rightarrow$ Computations to perform
  $\Rightarrow$ Weights storage and access
- Energy efficiency
  Typical constraints :
  - 10-100 uW for wearables,
  - 10-100 mW for phones,
  - 1-10 W for plugged edge devices
  - 100-1000 W for plugged cloud devices

Hardware Accelerated Neural Network

What are the constraints?

- Accuracy needs depend on the application
- Silicon resources:
  $\Rightarrow$ Computations to perform
  $\Rightarrow$ Weights storage and access
- Energy efficiency
  Typical constraints :
  - 10-100 uW for wearables,
  - 10-100 mW for phones,
  - 1-10 W for plugged edge devices
  - 100-1000 W for plugged cloud devices



Computation Demanding

Inference involves a lot of computation...

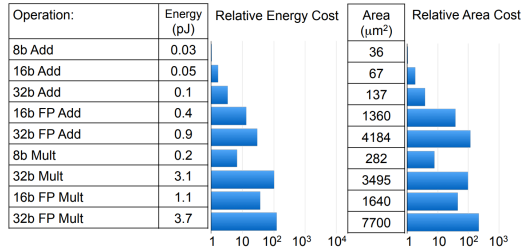
- Elevated number of floating point (FP) operations

  $0.5G \leq$ Nb of FLOPs $\leq 40G$

- Floating point operations are energy and area costly

(My 4 core-i7 PC ~120 GFLOPs $\Rightarrow$ 30 GFLOPs/core.)

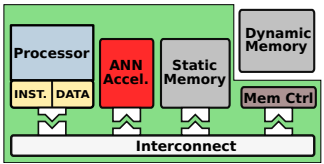
Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700



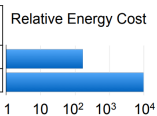
"Hardware Architectures for Deep Neural Networks", ISCA Tutorial, 2017

Memory demanding

Inference involves a lot of memory access...



Operation:	Energy (pJ)
32b SRAM Read (8KB)	5
32b DRAM Read	640



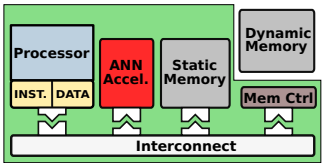
"Hardware Architectures for DNN", ISCA Tutorial, 2017

- Memory stores millions of (64-bit) weights
  $\Rightarrow$ 4M (GoogLeNet), 60M (AlexNet), 130M (VGG)

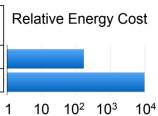
- Memory access becomes the bottleneck
  Each op needs 2 operands and produces a result

Memory demanding

Inference involves a lot of memory access...



Operation:	Energy (pJ)
32b SRAM Read (8KB)	5
32b DRAM Read	640



"Hardware Architectures for DNN", ISCA Tutorial, 2017

- Memory stores millions of (64-bit) weights
  $\Rightarrow$ 4M (GoogLeNet), 60M (AlexNet), 130M (VGG)

- Memory access becomes the bottleneck
  Each op needs 2 operands and produces a result

Coping with GFLOPs and GBytes

Alternatives: trade FLOPs for (some) accuracy

Simplify the operations

- Avoid sigmoid, batch normalization and stuff
- FP arithmetic is not HW friendly
  $\Rightarrow$ Do not use 16/32/64-bit floats

Alternatives: trade bytes for (some) accuracy

- Use “small” data types: {-1, 1}, {-1, 0, 1}, ...

Alternatives: re-architect the “system”

- Integrate many memory cuts with processing elements and use them wisely
- Integrate computation into the memory itself

Frameworks

- Tensorflow: tflite, qkeras
- Pytorch: quantization API
- Larq: binary only
- ...

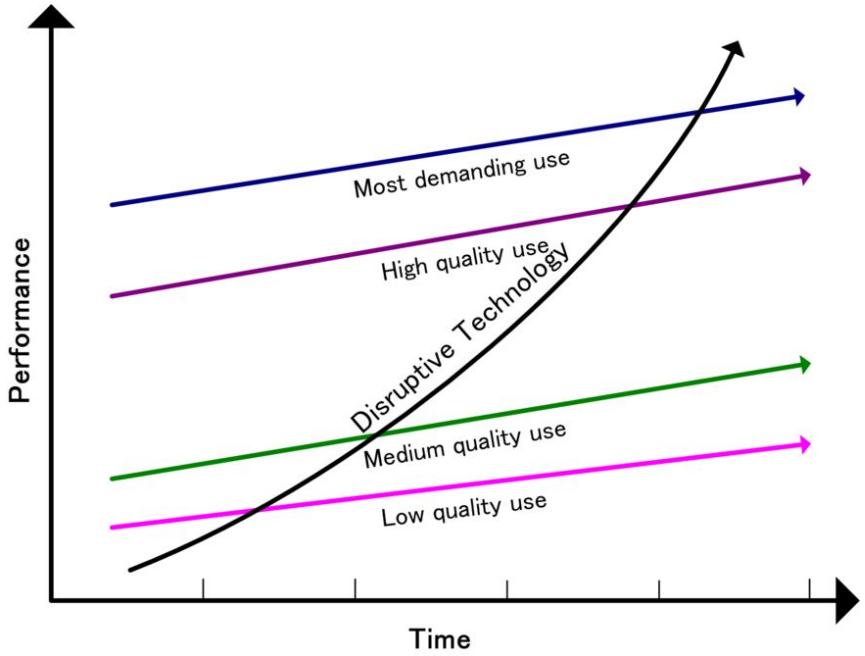
Quantization levels and accuracy...

Kees Vissers, *"A Framework for Reduced Precision Neural Networks on FPGAs"*, MPSOC, 2017

Some commercial use cases of TinyML today

Toys and gadgets?

A disruptive technology?



How to know if a use case is doable on device?

Possible

Trade-off zone

Experimentation!
“This is the way”

Impossible

With some examples

Impossible

- **Real-time translation**
- **Full language understanding**
- **Pedestrian detection & tracking (ADAS)**
- **Medical diagnosis**
- **…**

→ Edge, On premise or Cloud deployments

neoVISION

With some examples

Possible

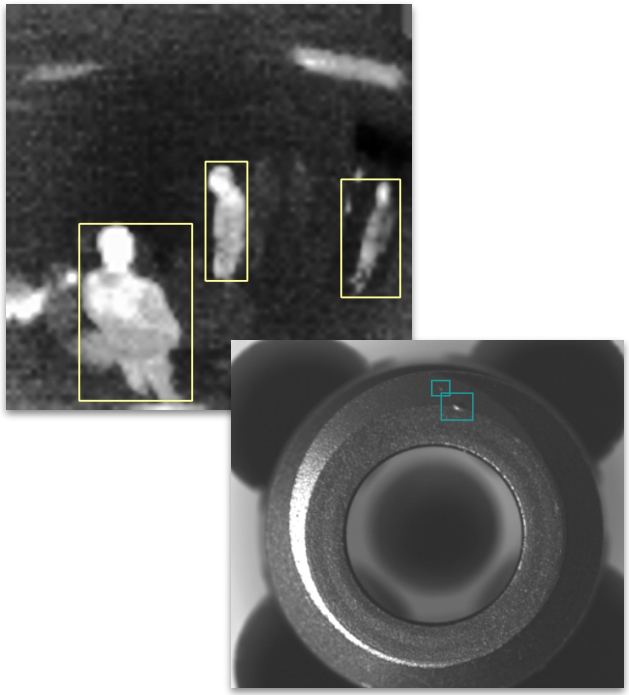
- Activity recognition
- Face detection
- Wake word (ie. “Alexa”)
- Presence detection
- …

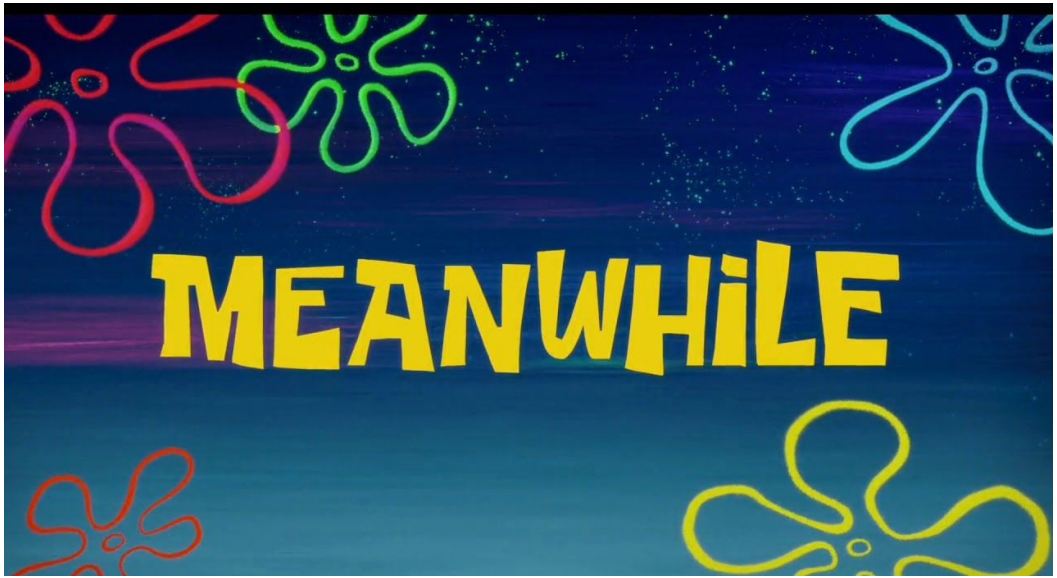
neoVISION

With some examples

Trade-off

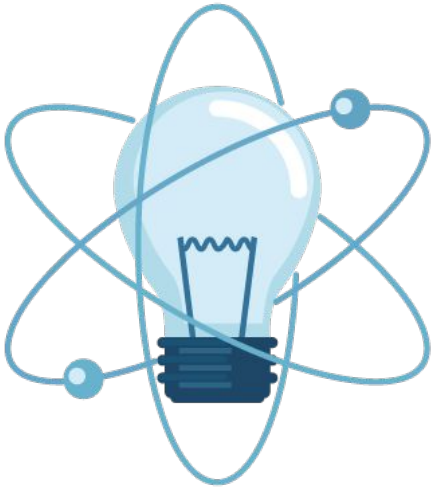
- People detection & tracking
- Face recognition
- Object recognition
- Defect and damage detection
- Optical Character Recognition
- Gaze tracking (AR/VR)
- …




MEANWHILE

“**Smart**” light bulb need...

an internet connection !



Thanks for your attention

Etienne Balit

Head of R&D

+33 6 87 32 32 32

etienne.balit@neovision.fr

For more information about AI, follow us!

@Neovision SAS

@neovisionSAS

@NeovisionSAS

http://neovision.fr

TinyML
Etat de l'art et enjeux

La vue d'un vendeur de processeurs

Loïc Liétar, co-founder & CEO

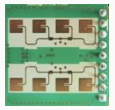
Un système TinyML typique

Capteurs “riches”

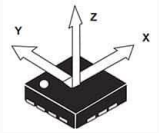
Camera lumière visible

Microphone

Camera infra-rouge

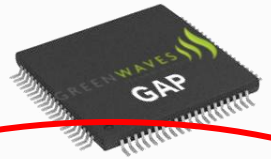


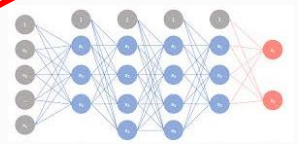
Radar



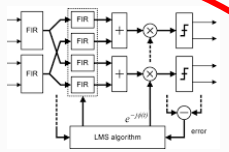
Microphone à conduction osseusse (accéleromètre 3D)

Processeur





Neural Network Inferencing



Digital Signal Processing

Communicatin

Low Power Radio
- BLE, BT
- LoRa …

années

semaines

journée

Pile type AA
5 ans d’autonomie = 300uW

Pile bouton
12h d’autonomie = 15mW

GREENWAVES TECHNOLOGIES

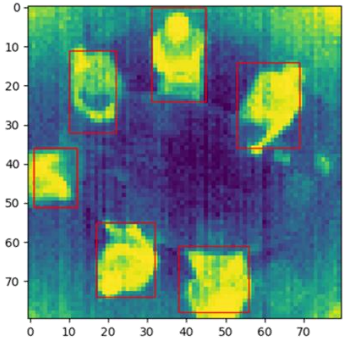
Exemples d’usage réels, sur pile

Commande vocale

Caméra intelligente

Verrou avec identification faciale

Comptage de personnes dans une salle de réunion



Reconnaissance de geste avec une signature radar

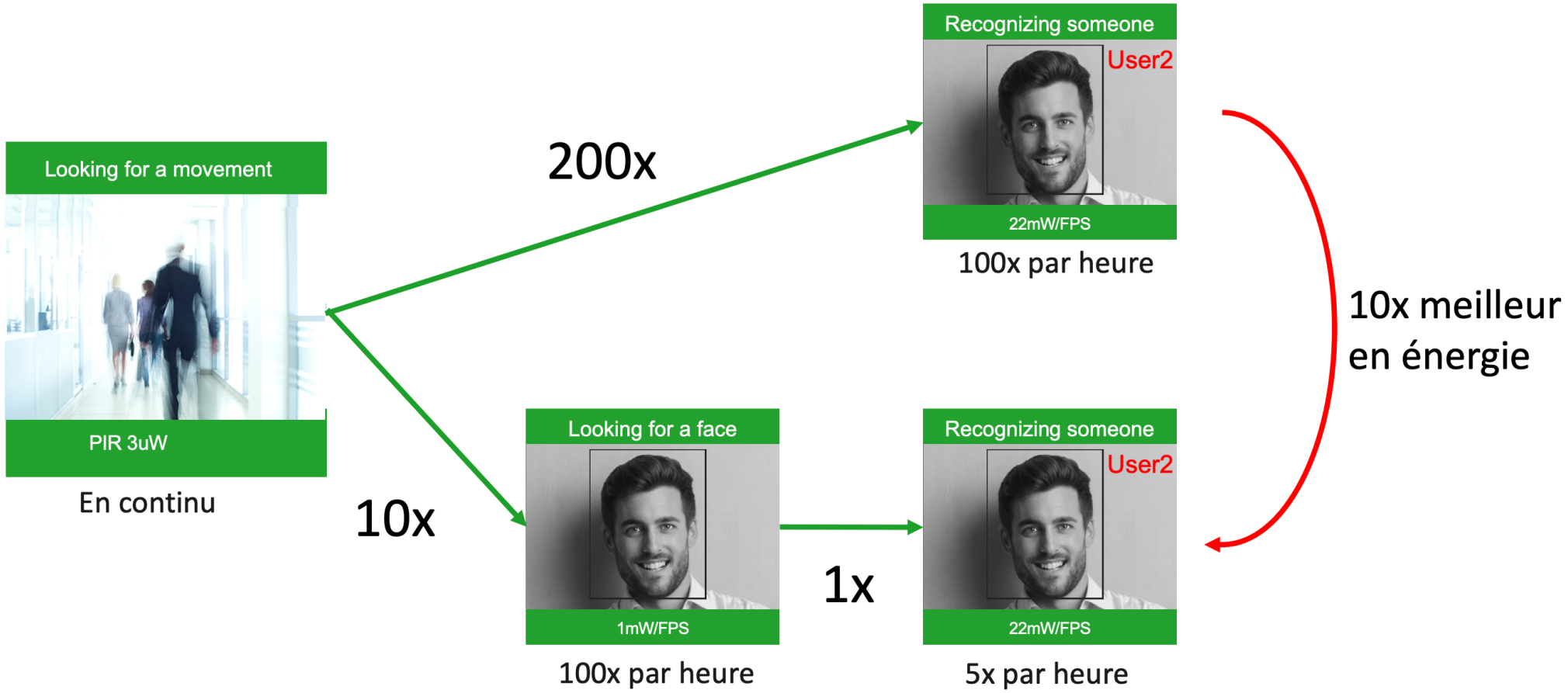
Débruitage aidé par la lecture des lèvres

Suppression du bruit aidé par la classification de l’environnement sonore

Débruitage aidé par l’identification du locuteur

GREENWAVES TECHNOLOGIES

Un problème est plus simple si on le hiérarchise

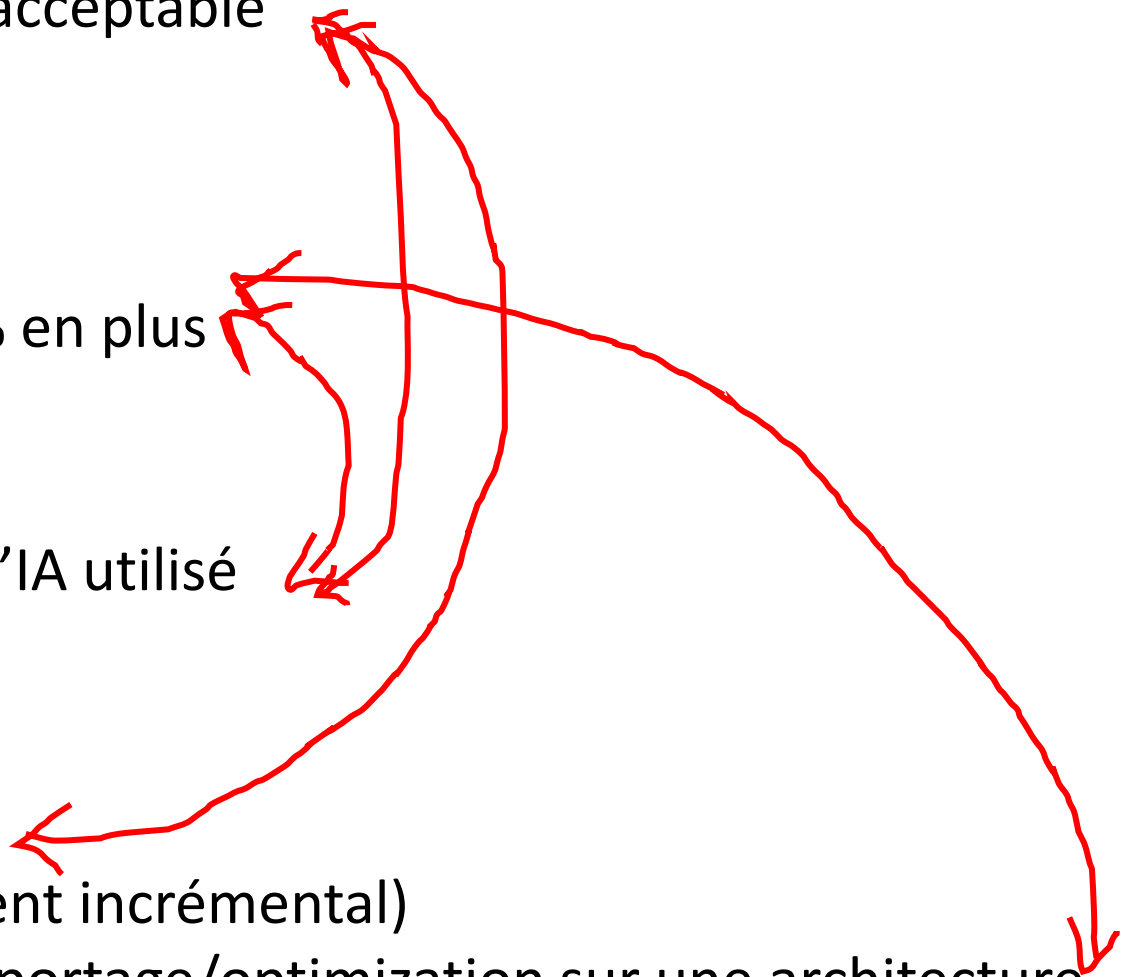


et si l'on utilise plusieurs types de données pour faire de la fusion (e.g. image et voix ici)

GREENWAVES TECHNOLOGIES

Dimensions d'un système TinyML

- Précision
  - Peu de retour encore sur ce qui n'est pas acceptable
- Autonomie
  - C'est l'ordre de grandeur qui compte
  - Personne ne change de solution pour 20% en plus
- Coût unitaire de production
  - Largement défini par la taille du modèle d'IA utilisé
- Coût de développement
  - Production des données d'apprentissage
  - Coût de l'apprentissage (même si seulement incrémental)
  - Réduction de la complexité du modèle et portage/optimization sur une architecture



Rien de tout cela n’est vraiment maitrisé à ce jour

- Les avancées en IA sont (encore) largement faites hors de toute considération matérielle.

- L’ open source semble être de rigueur, en apparence.

- Les techniques de réduction de complexité des modèles sont encore très empiriques, avec des d’outils peu performants et peu prédictibles.

- L’état de l’art en IA évolue plus vite que le temps de cycle de développement des architectures de processeurs.
- Ces architectures doivent donc être spécialisées pour être suffisamment efficientes, mais pas trop pour ne pas être obsolètes avant d’avoir été utilisées.

- Au delà des données “classiques”, images et voix en langue “majeures” qui sont assez ouvertes, la production (ou a minima, l’enrichissement) de données est à faire au cas par cas de façon très peu productive et empiriques.

- Les applications sont très fragmentées.

Le TinyML ouvre la porte de l'IA
à de nombreuses applications sur micro-contrôleur

La boîte à outils en est vraiment encore à ses débuts

Un des marchés les plus propices à l'innovation multi-dimensionnelle pour les années à venir

Merci

GREENWAVES TECHNOLOGIES
28 Cours Jean Jaurès
38000 Grenoble
France
www.greenwaves-technologies.com

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyML.org