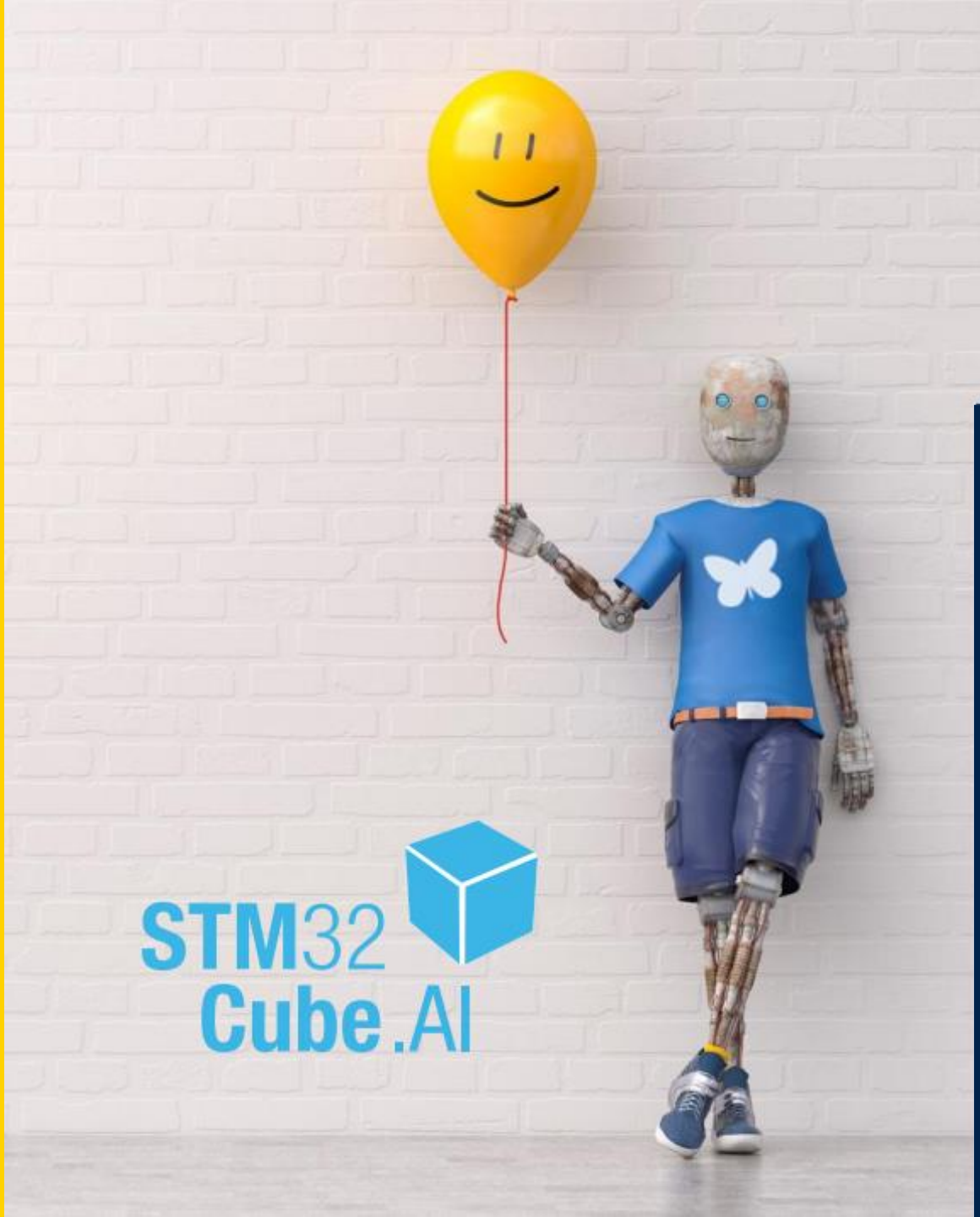


life.augmented

STM32Cube.AI Neural Networks on STM32

Blaine Moon

AI on the news…

Can you easily access it?

What if we talk about bringing AI to your 2020 projects?

life.augmented

ST and AI on the Edge

- **Microcontrollers** are the brain of these smart objects
- AI is a disruptive technology where ST has been investing for many years …
- Local processing to fix limitations :
  - Latency and cost communications
  - Autonomy (Battery-operated devices)
  - Limited networks bandwidth or connectivity loss
  - Data privacy

life.augmented

Distributed AI from Edge to Cloud

Edge

Cloud

Endpoint
(Node/Smart Object)

- Sensing
- Data privacy
- Micro Decisions (AI)
- Real time
- Micro Actuation

Sensor Data (Option)

Event detection

Actions

Gateways

Home

- Data or Event integration
- Data Privacy
- Meso Decisions (AI)
- Meso Actuation

Industrial

Event detection

Actions

- Macro Decisions (AI)
- Macro Actuation

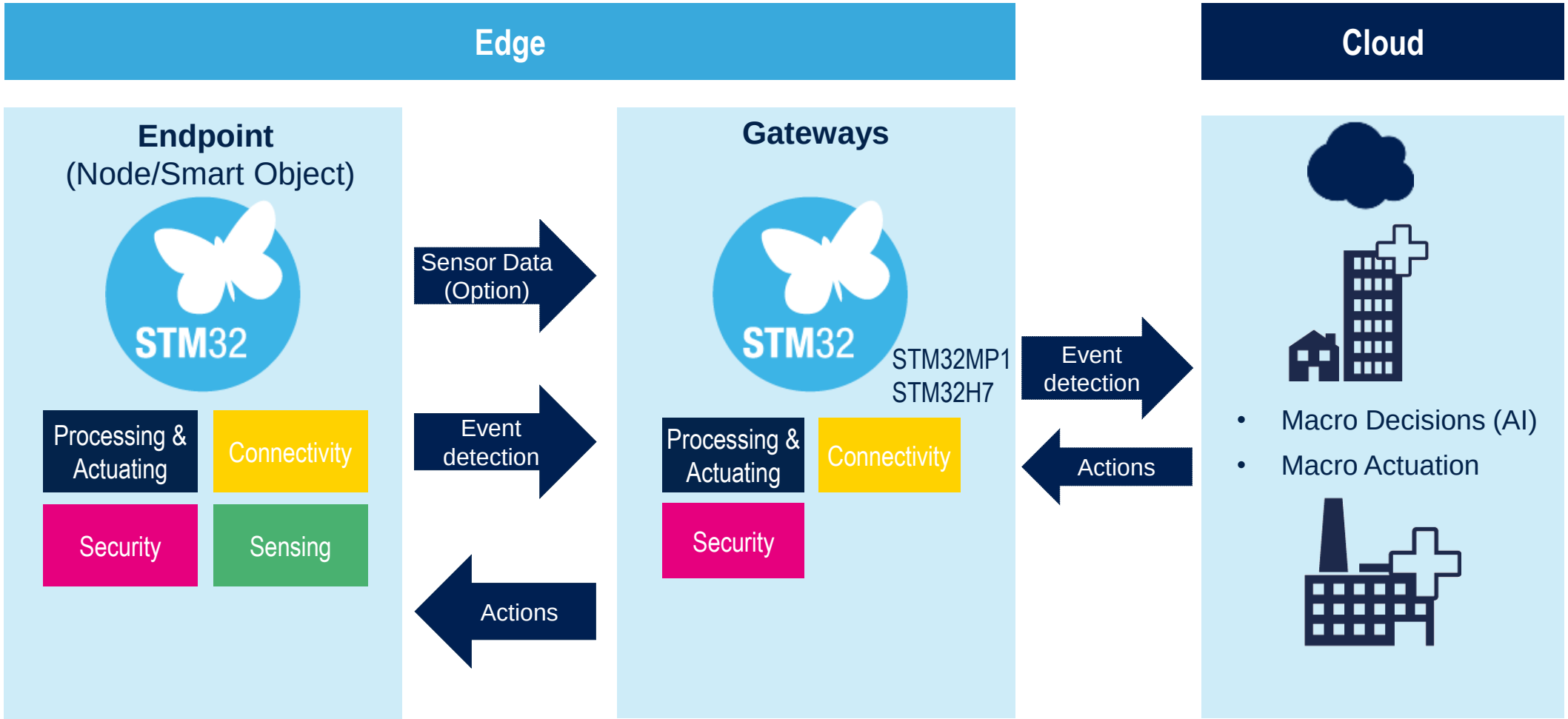
Machine level Decisions

Room or building-level Decisions

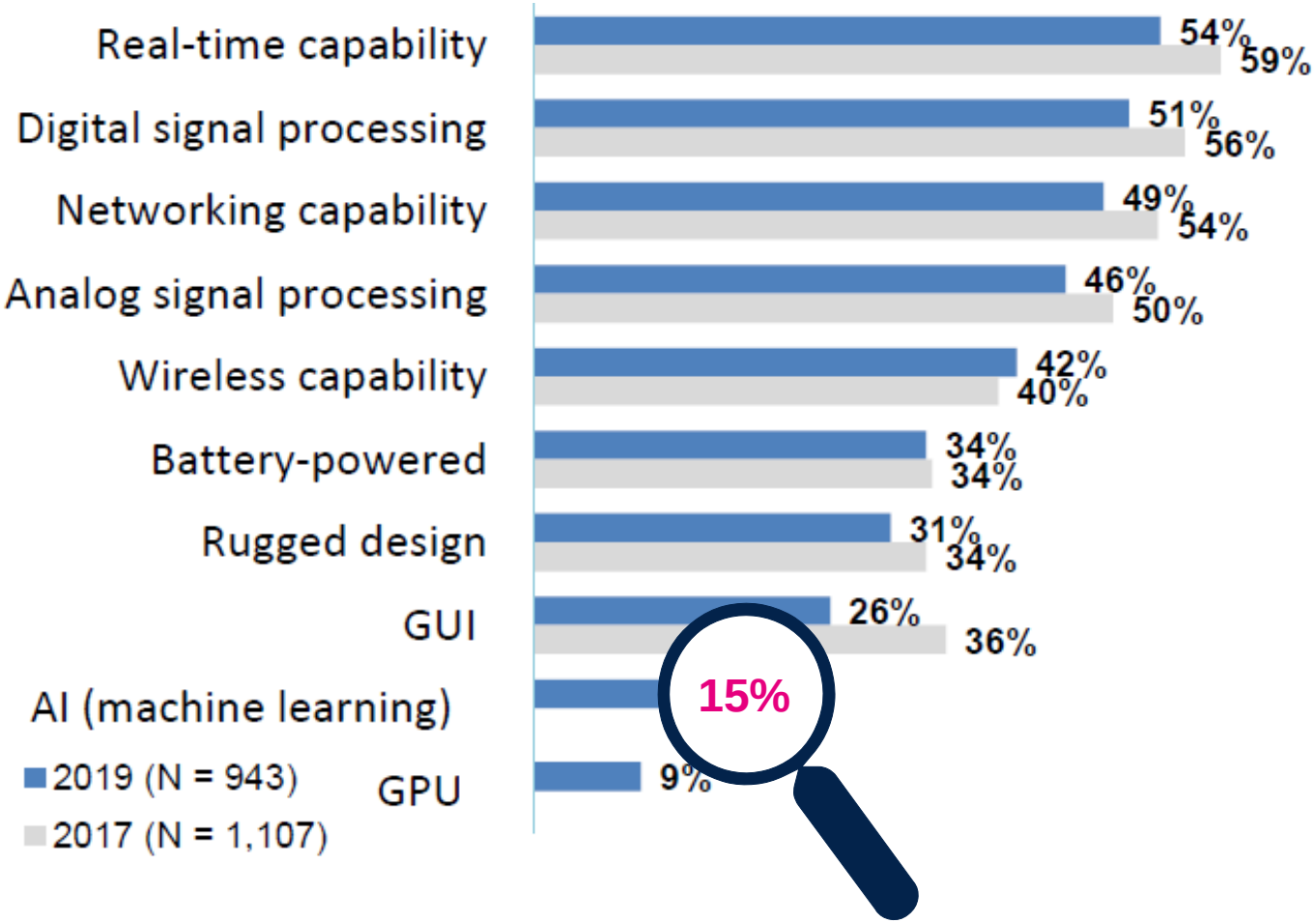
Cities or factory-level decisions

life.augmented

Distributed AI from Edge to Cloud



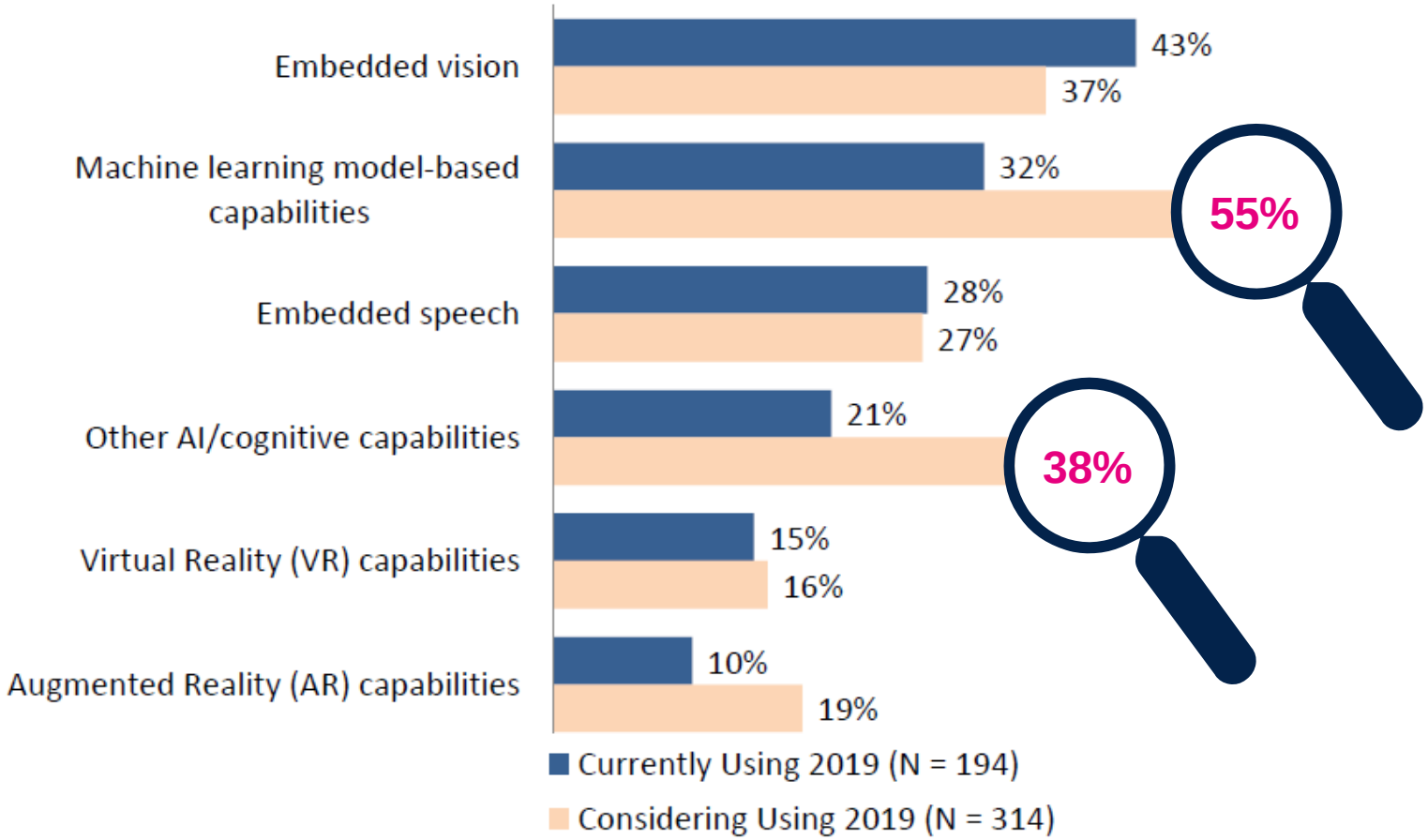
AI moving to the Edge



- 15% of embedded projects already include AI in 2019
- 0% in 2017

Aspencore 2019 embedded markets study

AI moving to the Edge



- MCU pervasion will grow
  - With Machine Learning
  - Other AI cognitive capabilities
- MPU adoption will continue
  - With Embedded vision & Speech

68%
of EMEA users
are considering
using Machine
Learninf

Aspencore 2019 embedded markets study

Neural Networks on STM32

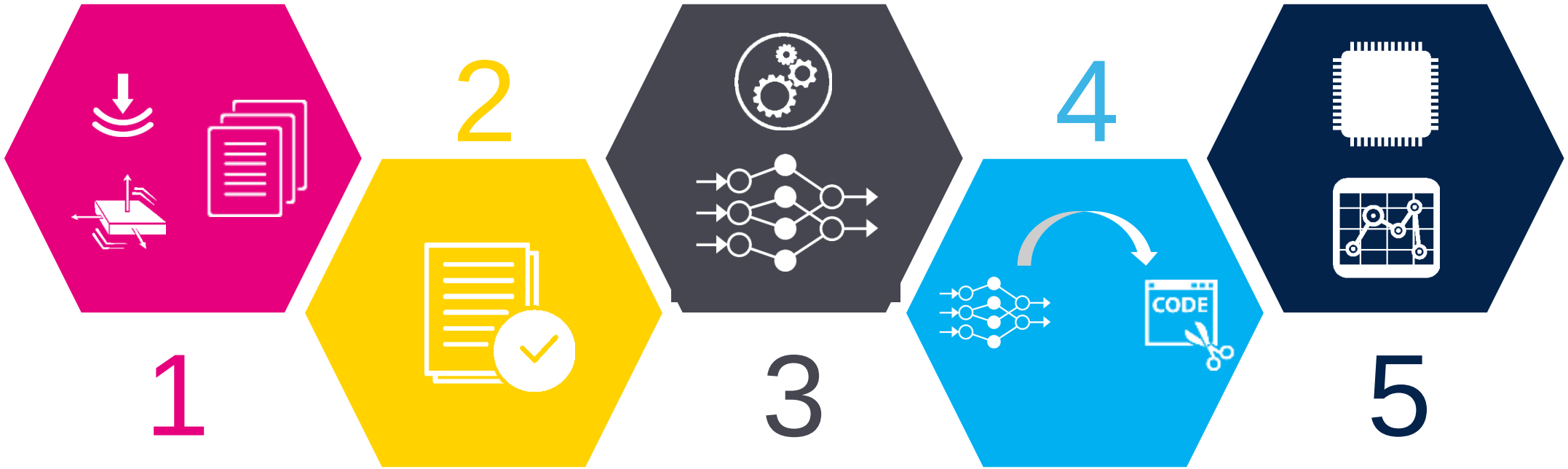
STM32 Cube.AI

life.augmented

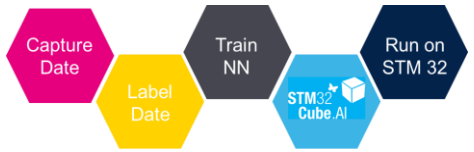
The key steps behind Neural Networks

Neural Network (NN) Model Creation

Operating Mode



life.augmented



ST toolbox for Neural Networks

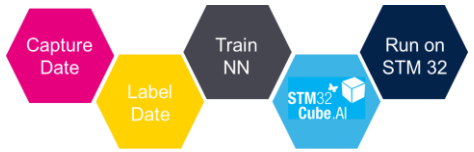
Capture Data

Process & analyze
new data using trained NN

life.augmented

Clean, label Data
Build NN topology

Convert NN into
optimized code for MCU



Collecting data & architecting an NN topology

Services provided by Partners

ST tools to support

Capture Data

Clean, label Data
Build NN topology

ST BLE Sensor mobile phone application

Collect and label data from the SensorTile.

ST BLE Sensor

Selected partners

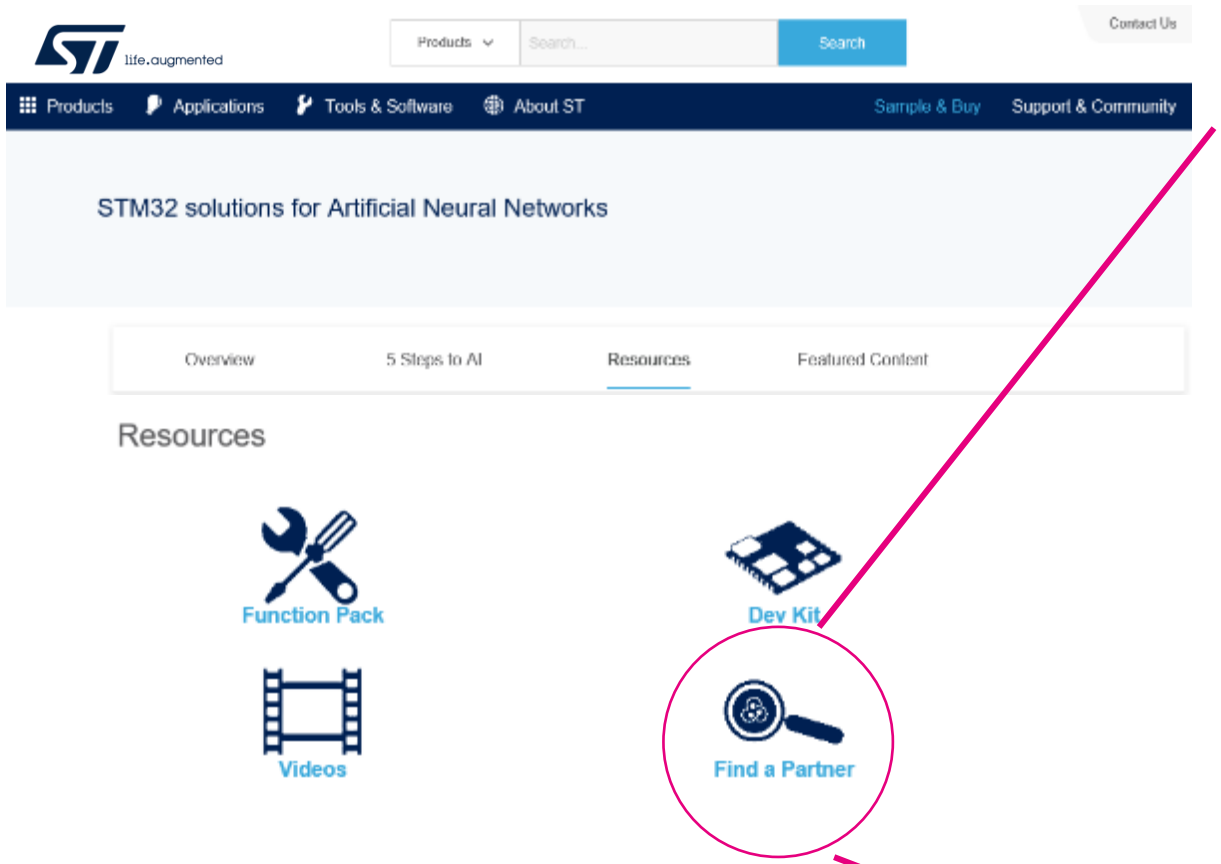
Neural Networks engineering services support.

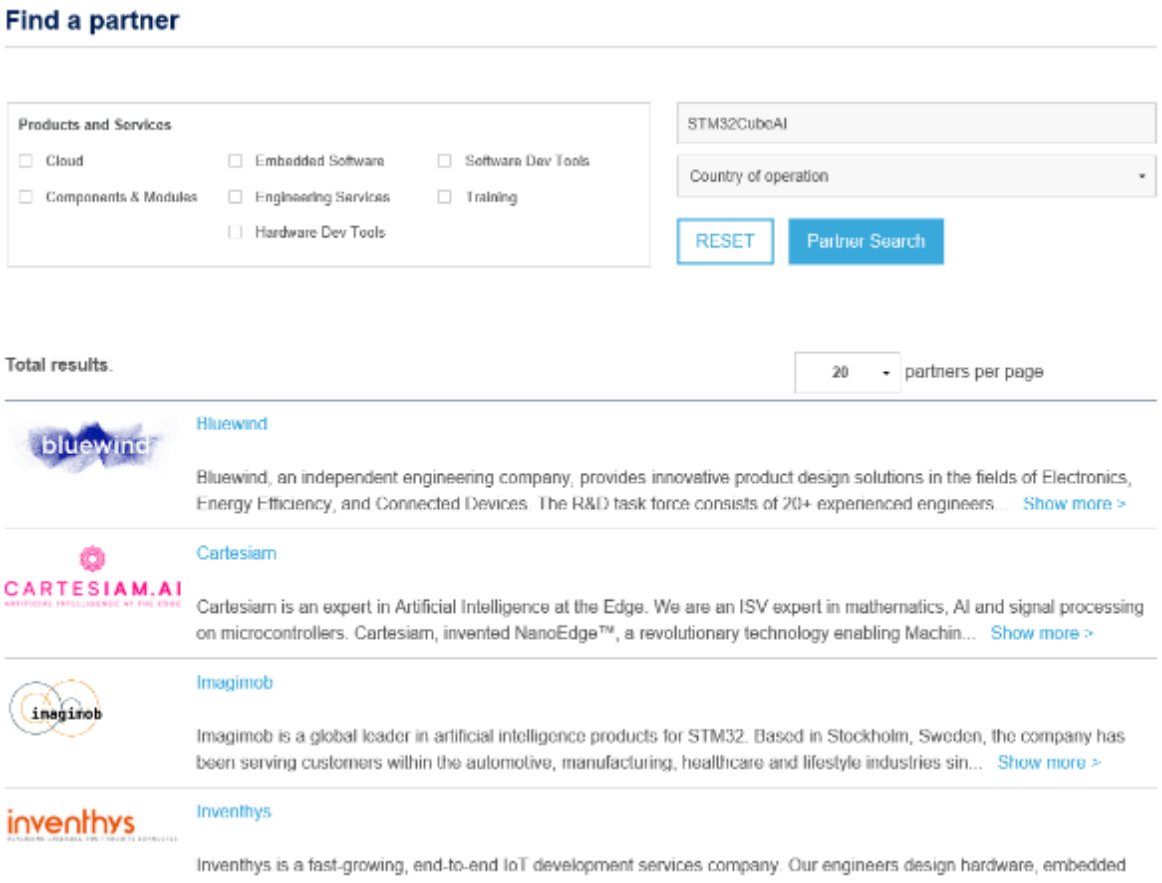
Data scientists and Neural network architects.

life.augmented

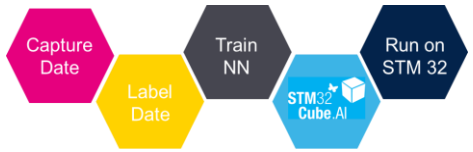
Partner Program

ST AI official partners





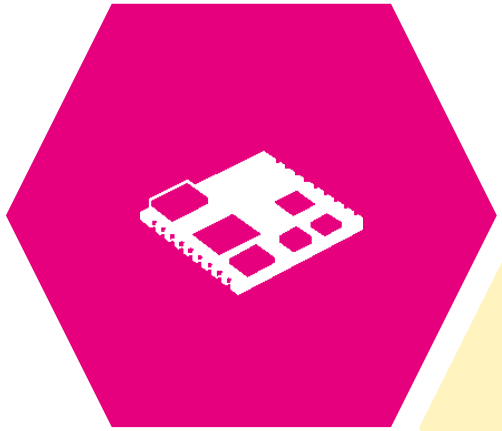
www.st.com/STM32CubeAI

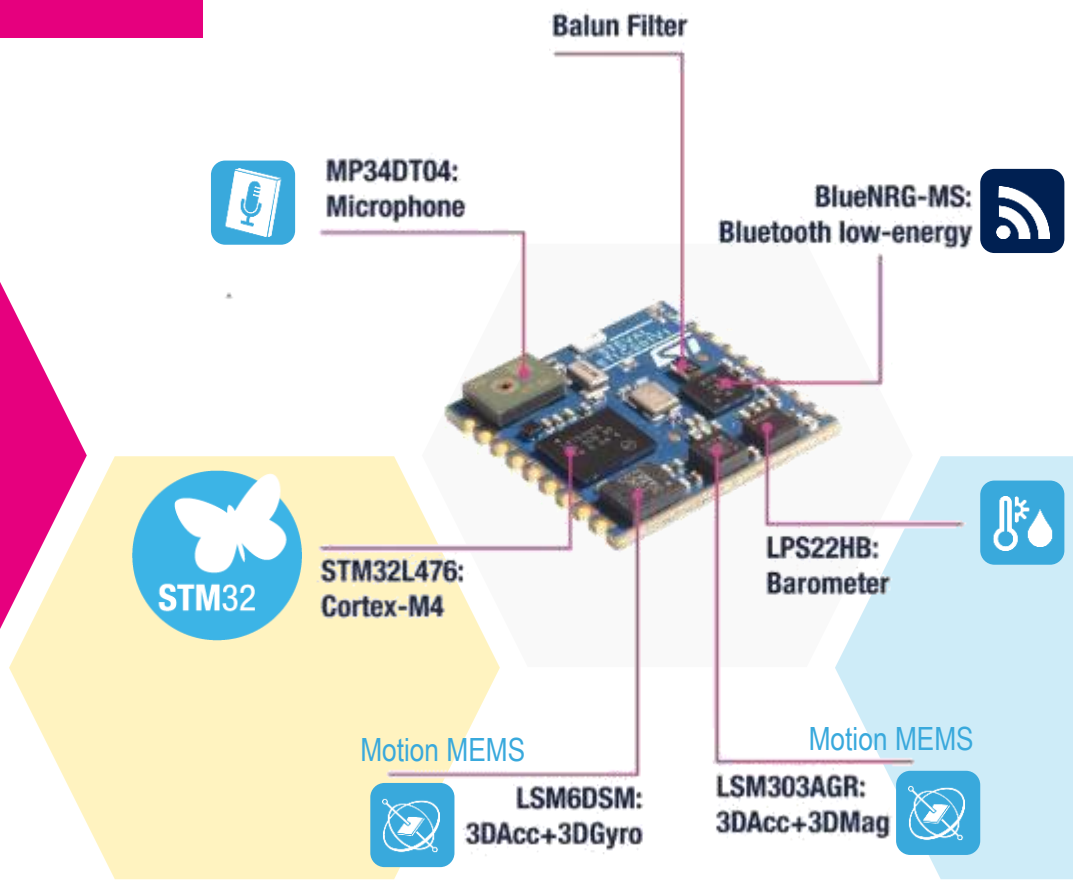


Example form-factor hardware to capture and process data

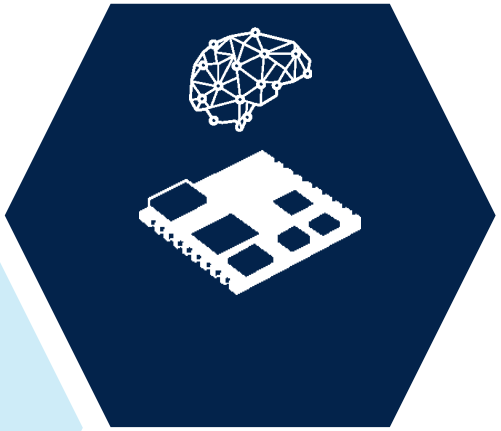
SensorTile

Capture Data



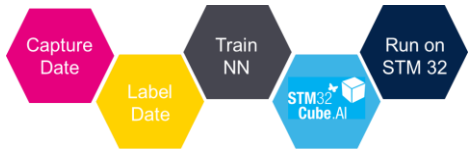


Process & analyze new data using trained NN



www.st.com/SensorTile
www.st.com/SensorTile-edu

life.augmented



Fast go-to-market module to capture data with more accuracy

SensorTile.Box

Capture Data

Process & analyze new data using trained NN

Microsoft IoT Services ready

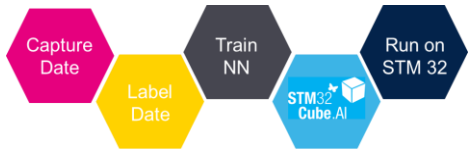
Microsoft Azure

More advanced, high accuracy and low power sensors

- First Inertial module with Machine Learning capabilities.
- Motion (accelerometer and gyroscope, magnetometer) and slow motion (inclinometer)
- Altitude (pressure), environment (pressure, temperature, humidity, compass) and sound (sound and ultrasound analog microphone)
- Microsoft IoT services ready to make available on a web dashboard the result of the embedded processing

www.st.com/SensorTileBox

life.augmented



Form factor hardware
AI IoT node for more connectivity

IoTNode

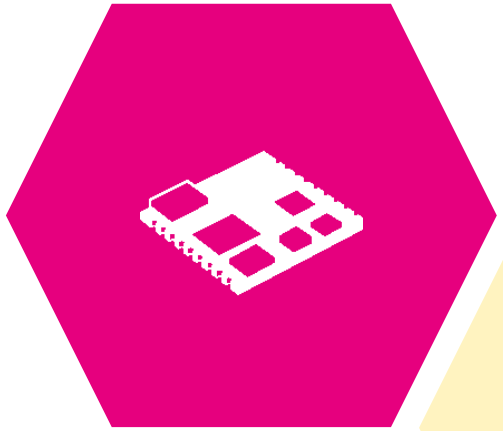
+

Sub 1GHz Sub-1GHz

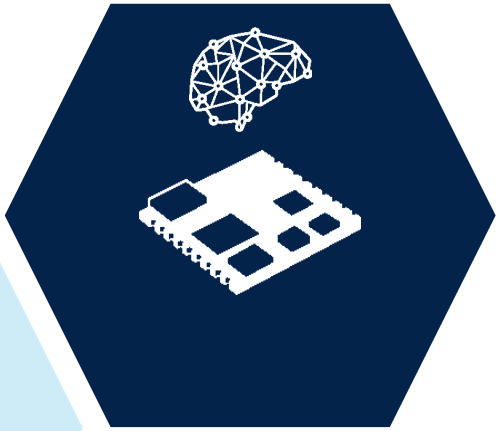
NFC Dynamic NFC Tag

Wi-Fi Wi-Fi

Capture Data



Process & analyze
new data using trained NN

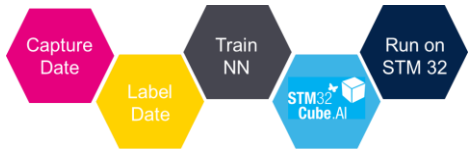


More debug capabilities

- Integrated ST-Link/V2.1
- PMOD extension connector
- Arduino Uno extension connectors

https://www.st.com/IoTnode

life.augmented



OpenMV integration
Fast machine vision prototyping

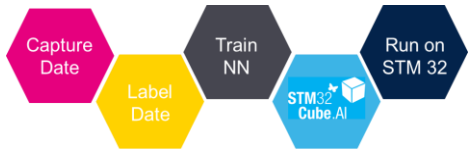
OpenMV CAM
Running MicroPython on STM32

Configure Machine Vision in real-time over USB in Python

Run and validate optimized Neural Network

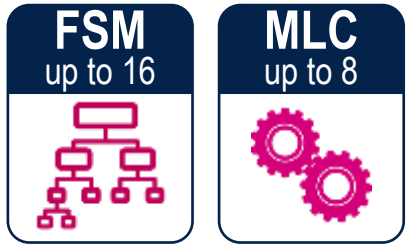
STM32 Cube.AI

https://github.com/openmv/openmv



Distributed AI: Sensor + STM32
Optimize performance & power consumption

Smart Sensor
with Machine Learning Core



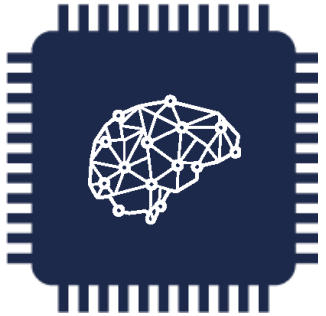
Inertial Sensor
New LSM6DSOX

Raw Data

Event Decision

FSM and MLC
Re-configuration

Smart STM32
second level of AI processing



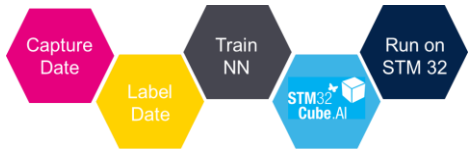
Deep Learning
Neural Networks
Machine Learning

- Best ultra-low-power sensing at high performance
  - 550µA (gyroscope and accelerometer)
    -25% power compared to competition
  - 20~40µA (Accelerometer only for HAR)
- Efficient Finite State Machines: 2µA
- Configurable Machine Learning Core: 4~8µA

- More advanced and complex Neural Networks
- Input can be data from multiple sensor data and/or sensor Machine Learning decisions
- Possible to run multiple Neural Networks
- Actuation & communication

life.augmented

https://www.st.com/content/st_com/en/products/mems-and-sensors/inemo-inertial-modules/lsm6dsox.html



STM32CubeMX extension AI conversion tool

Input your framework-dependent, pre-trained Neural Network into the **STM32Cube**.AI conversion tool

Automatic and fast generation of an STM32-optimized library

STM32Cube.AI offers interoperability with state-of-the-art Deep Learning design frameworks

Train NN Model

ONNX

Caffe

TensorFlow Lite

Lasagne

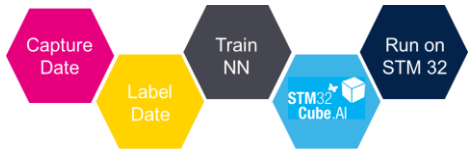
STM32 Cube.AI

Process & analyze new data using trained NN

STM32

Convert NN into optimized code for MCU
Validate code directly on target

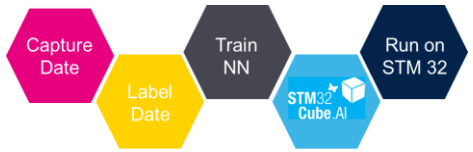
life.augmented



Import Neural Networks
Open Neural Network (ONNX) exchange format

- ONNX open format enables models to be trained in one framework and transferred to another for inference
- Common import/export format of many frameworks
- STM32 Cube.AI hardware optimization is available for any tools exporting ONNX models

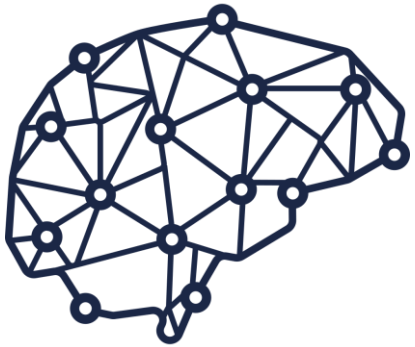
life.augmented



STM32 solutions for AI
More than just the STM32Cube.AI

An extensive toolbox to support easy creation of your AI application

STM32 Cube.AI



AI extension for STM32CubeMX
to **map pre-trained Neural Networks**

Software examples for Quick prototyping
Audio, Motion and Vision Function packs
On **ST development Hardware**

STM32 **Community** with dedicated
Neural Networks topic

Trainings, hands-on exercises, MOOCs
and partners **videos**

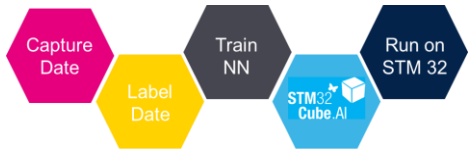
STM32 AI Partner Program
with dedicated Partners providing
Machine or Deep Learning engineering services

life.augmented

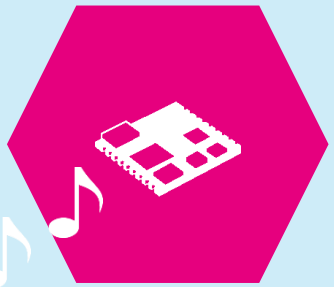
Function Packs
(Software examples)

STM32 Cube.AI

life.augmented



Audio Scene Classification (ASC)
Audio example in FP-AI-SENSING1 package



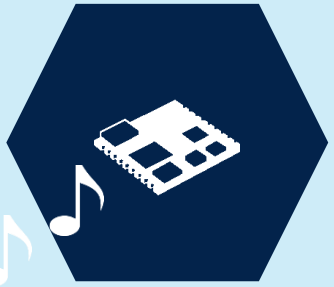
3 classes

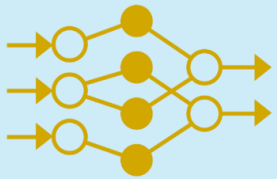
Audio **Data capture**

Labeling controlled
by smartphone application

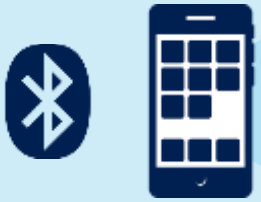
Data stored on the device
SD card for future **learning**

Indoor, Outdoor, In vehicle
labeling





NN & example
dataset provided

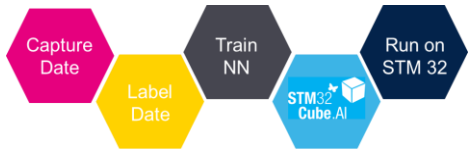


Embedded **audio**
pre-processing

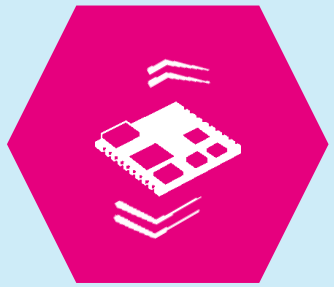
Inferences running
on the microcontroller

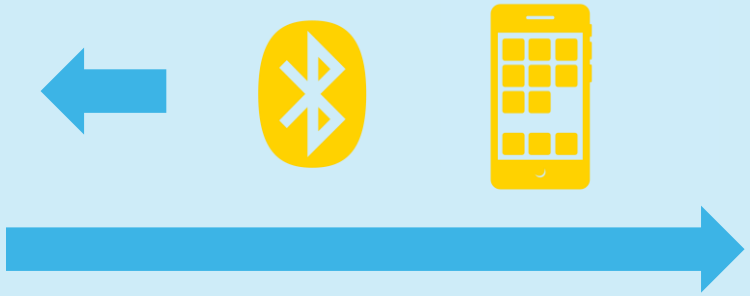
Inference result
displayed on mobile app

life.augmented



Human Activity Recognition (HAR)
Motion example in FP-AI-SENSING1 package



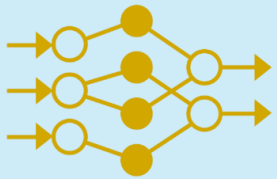


Motion **Data Capture**

Labeling controlled
by smartphone application

Data stored on the device
SD card for future **learning**

Stationary, walking, running,
biking, driving **labeling**



NN & example
dataset provided

Embedded **motion**
pre-processing

Inferences running
on the microcontroller

Inference result
displayed on mobile app

life.augmented

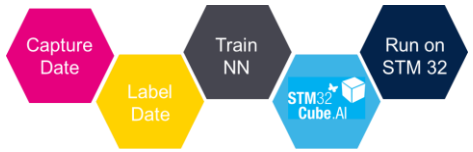


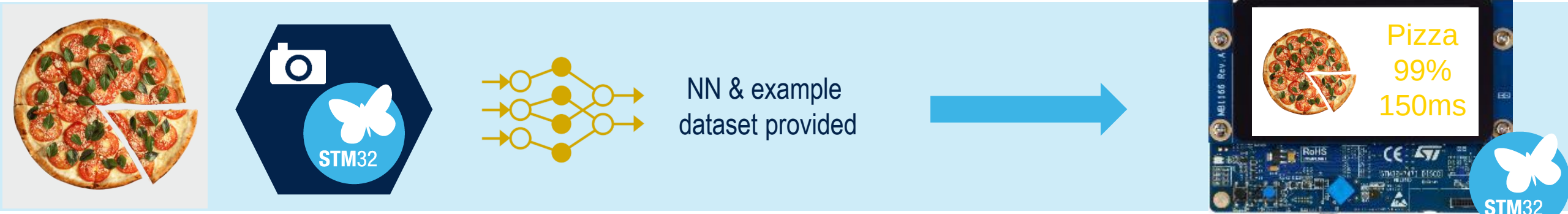
Image Classification
Vision example in FP-AI-VISION1 package

Enjoy the **food classification demo**

- Default demo based on 18 classes (224x224 RGB pictures)
- Several camera image output size possible

Full **end-to-end optimized software example**

- from camera acquisition to image pre-processing before feeding the NN
- Multiple memory mapping possibilities to optimize and test impact on performances
- Retrain this NN with your own dataset
- Quantize your trained network to optimized inference time and memory usage



Embedded **image** pre-processing (SW) on the STM32H747

Inferences running on the microcontroller

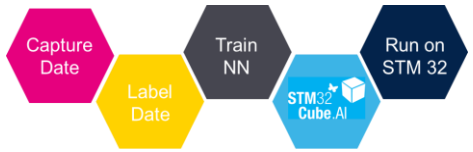
Inference result displayed on STM32H747 Discovery board LCD display

life.augmented

AI solutions for STM32MP1

STM32 Cube.AI

life.augmented



STM32MP1 microprocessor Augmented intelligence

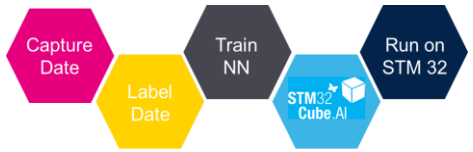
STM32MP1 = Cortex-M4 Realtime OS STM32CubeMX + Dual Cortex-A7 OpenSTLinux Distribution STM32CubeMx

STM32Cube.AI

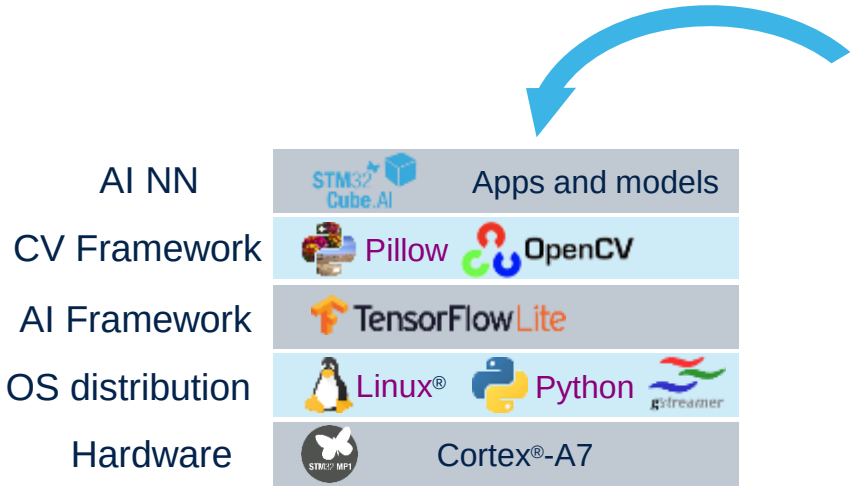
X-LINUX-AI-CV support for TensorFlowLite & OpenCV

- STM32Cube.AI to convert pre-trained NNs for the Cortex-M4 core
- TensorFlow Lite STM32MP1 support up streamed for native NN inferences support on the dual Cortex-A side

life.augmented



X-Linux-AI-CV package for STM32MP1 computer vision application

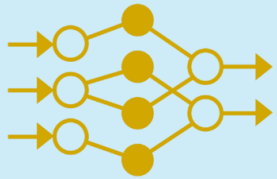


Application examples in C/C++ and Python

- Image classification: 1000 objects classified
- Multiple object detection: 90 classes

Includes code for camera acquisition and image pre-processing

STM32MP1



AI, CV frameworks & application examples provided

USB camera or built-in camera module

Inferences running on the microprocessor in 80ms for image classification

Displayed on STM32MP1-DK2, STM32MP1-EV1 and Avenger96 board

life.augmented

Vision on MCU – A reality now diving into STM32Cube.AI

STM32 Cube.AI

life.augmented

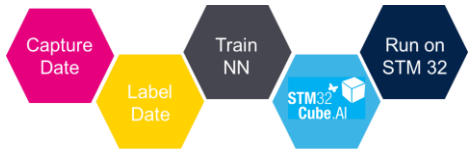
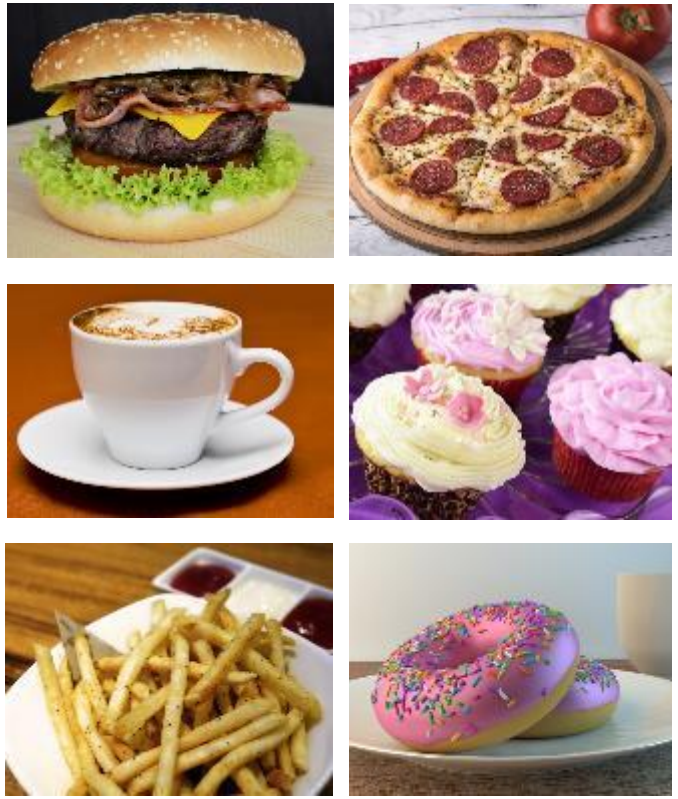
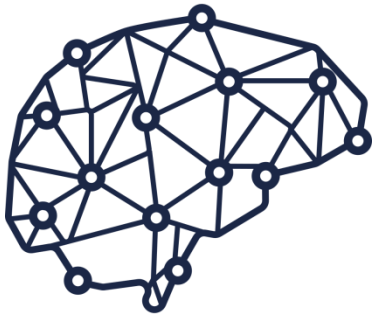


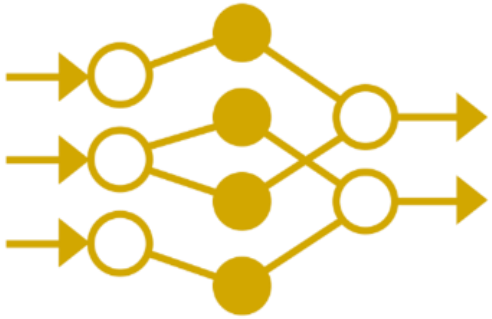
Image Classification
Vision example in FP-AI-VISION1 package



Live video stream



18 classes of food recognized



Recognition: Hamburger
Confidence: 85%

Embedded image pre-processing (SW) on STM32H747

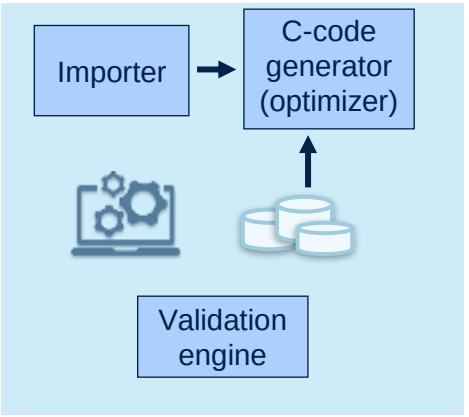
Inferences running on the microcontroller

Inference result displayed on STM32H747 Discovery board LCD

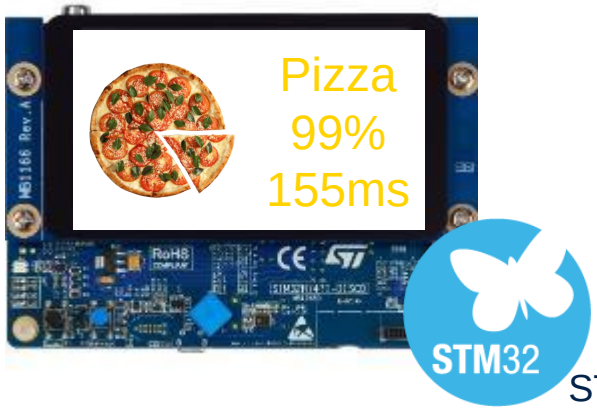
life.augmented

The generation steps

1. Design the Neural Network
2. Optimize the execution thanks to Cube.AI
3. Validate the optimization with Cube.AI
4. Integrate in a complete application thanks to function pack example



STM32 Cube.AI



Function Pack

STM32H7

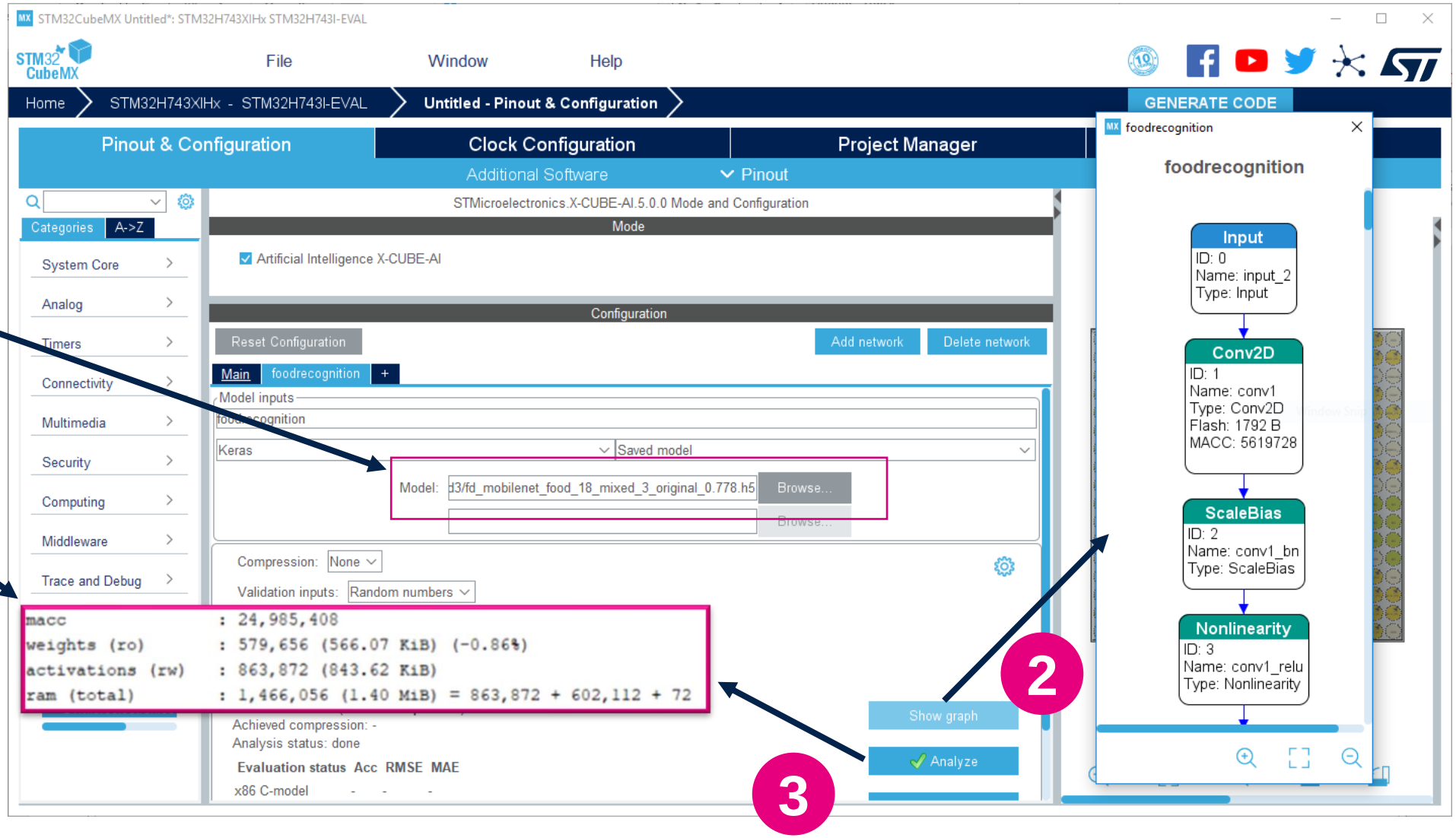
life.augmented

STM32Cube.AI code generation

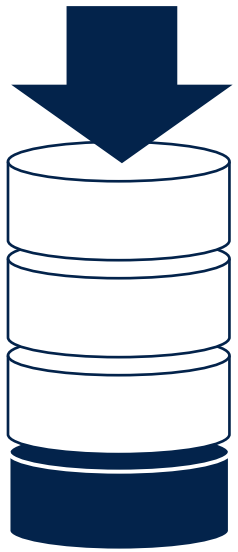
1 Enter the model: here float Keras h5 model

Memory footprint after analysis:
> 1 MB of RAM

How to reduce?



Quantization objective



Computer Vision Use Cases need a lot of memory
➔ Critical to reduce the memory for efficiency

Convert floating-point (32 bits) model to quantized (8 bits) model

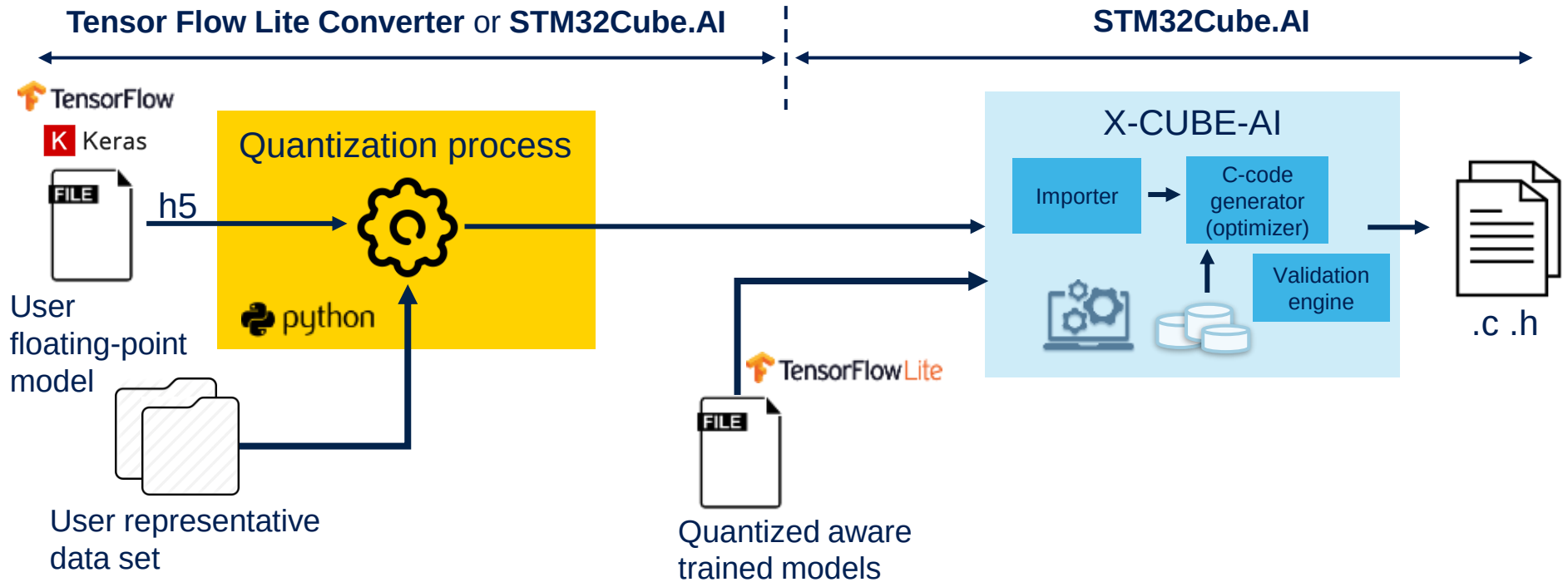
- **Reduce model size** (size of the memory to store the weights, in Flash). Up to x4
- **Reduce memory usage** (size of the activations buffer, in RAM). Up to x4
- **Improve latency.** Consequently **power consumption is improved**. Up to x3.
- With **minimal loss of accuracy**. Network size/complexity dependent

Reference: **<STM32CubeMX_localpath>\Packs\STMicroelectronics\X-CUBE-AI\5.0.0\Documentation**

life.augmented

STM32Cube.AI solution

- Post-training quantization algorithm
  - Based on an already-trained floating-point model
  - Parameters and activations are quantized
- Integrated code generation
  - Including adapted validation process
  - Support of TFL quantized aware trained models



Validation flow for quantized model

.[npz|npy|csv]

Validation data

1

Reset Configuration

Main | network | +

Model inputs

network

Keras

Model: ...hition\FoodReco_Mobilel

Quantization file: ...tion\FoodReco_MobileN

Compression: None

Validation inputs: ...food_validation\x_val.npy

Validation outputs: ...food_validation\y_val.npy

Complexity: 12071092 MACC
Flash occupation: 128.82 KBytes
RAM: 125.58 KBytes
Achieved compression: -
Analysis status: -

Evaluation status	Acc	RMSE	MAE
x86 C-model	-	-	-
stm32 C-model	64.0%	0.158075	0.043342
original model	73.0%	0.144605	0.037849
X-cross	86.0%	0.053415	0.015071

3

Please wait...

Validation on target

```
NOTE: Computed against the provided ground truth values

acc=64.00%, rmse=0.1581, mae=0.0433

18 classes (100 samples)
```

	c0	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14
c0	2	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	.
c1	.	2	.	.	.	.	.	.	2	1	.	.	.	.	.	.	.	.
c2	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
c3	.	.	.	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.
c4	1	.	.	.	5	.	.	.	1	.	.	1	2	.	.	.	.	.
c5	.	.	1	.	.	5	.	.	.	.	.	.	.	.	.	.	.	1
c6	.	.	.	.	.	.	6	.	1	.	.	.	.	.	.	.	.	.
c7	.	.	.	.	.	1	1	1	.	.	.	.	.	.	.	.	.	.
c8	.	1	.	1	.	.	.	.	2	.	.	.	2	.	.	.	.	.
c9	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	.	.	.
c10	.	.	.	.	.	.	.	.	.	.	4	.	.	.	.	.	.	1
c11	.	.	.	.	.	.	.	.	.	.	.	2	.	.	1	.	.	.
c12	.	.	.	.	.	.	.	.	.	1	.	.	4	.	1	.	.	.
c13	.	.	.	.	.	.	.	.	.	.	.	.	1	0	3	.	3	.
c14	.	.	.	.	.	.	.	.	.	.	.	.	.	1	6	.	1	.

OK

3

Show graph

Analyze

2

Validate on desktop

Validate on target

Results quantized vs float

Values measured on FP-AI-VISION1 1.1.0

Using FD-MobileNet
(optimized 18-classes)

Input: 224x224x3
MACC: 24 M
Parameters: 145 K
Nb of layers: 12
Main kernel: DS-Conv, quantized 8 bits

KPI	Float model	Quantized model Cube.AI 4.0.0	Quantized model Cube.AI 5.0.0
Flash	566 kB	**148 kB**	**148 kB**
RAM	844 kB	**393 kB**	**212 kB**
Accuracy	77.8 %	77.1 %	77.1 %
Inference time	420 ms(1)	**155 ms**	**153 ms**
Frame Per Second(2)	2.2	**5.8**	**5.8**

Inference times on the STM32H7 @400MHz, quantization done by Cube.AI (Qmn, greedy options)
(1) External memory: activation buffer & I/O buffers in external SDRAM
(2) FPS takes into account the capture, preprocessing and inference times.

life.augmented

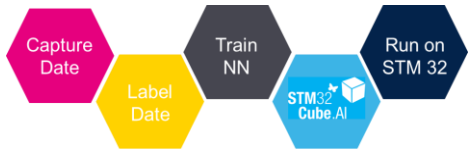
Results through STM32 portfolio Performance measurement

Using FD-MobileNet
(optimized 18-classes)

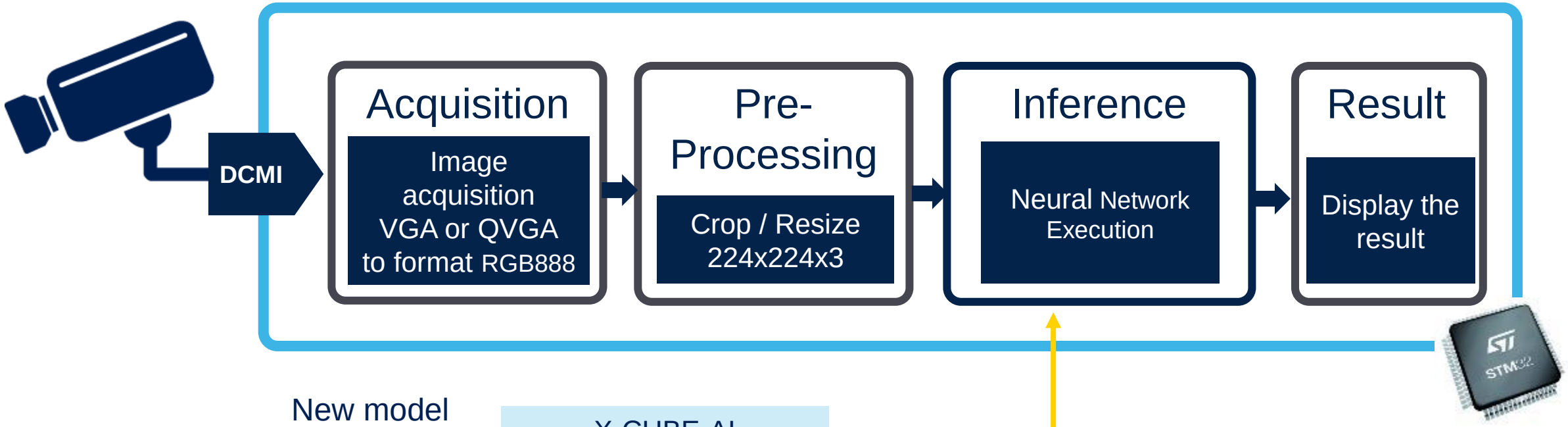
Input: 224x224x3
MACC: 24 M
Parameters: 145 K
Nb of layers: 12
Main kernel: DS-Conv, quantized 8 bits

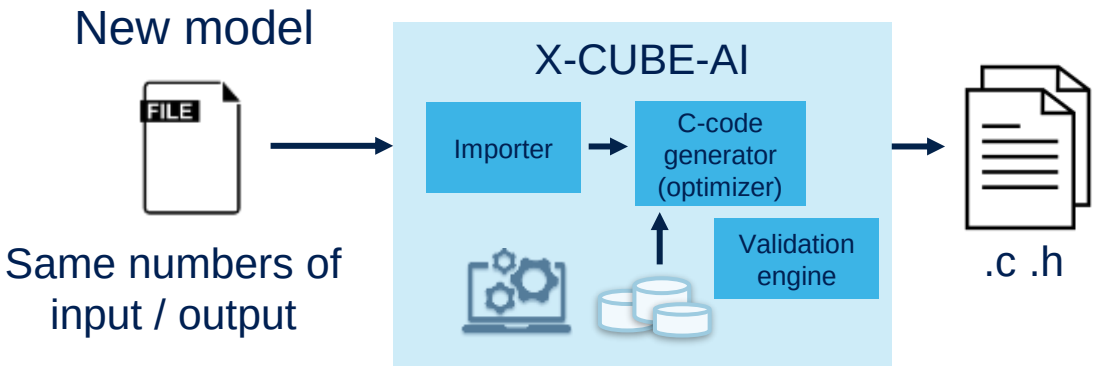
KPI	STM32L4R	STM32H7	STM32MP1
Flash	148 kB	148 kB	191kB
RAM	212 kB	212 kB	1MB
Inference Time	1.062 ms	153 ms	27.6 ms
Frame per Second	0.9 fps	5.8 fps	36.2 fps
Processor	L4R DK M4@120 MHz int RAM, int Flash,	H747 M7@400 MHz int RAM, int Flash,	MP1 DK2 2xA7@650 MHz TFL 2.0 C++
Cube.AI	Cube.AI 5.0.0, 8 bits Ua/Ua quantized model *input buffer allocated in activation buffer		

life.augmented



FP-AI-VISION1 Model update

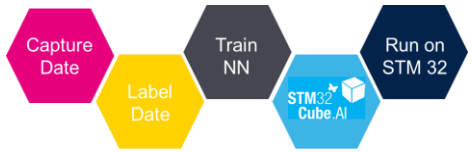




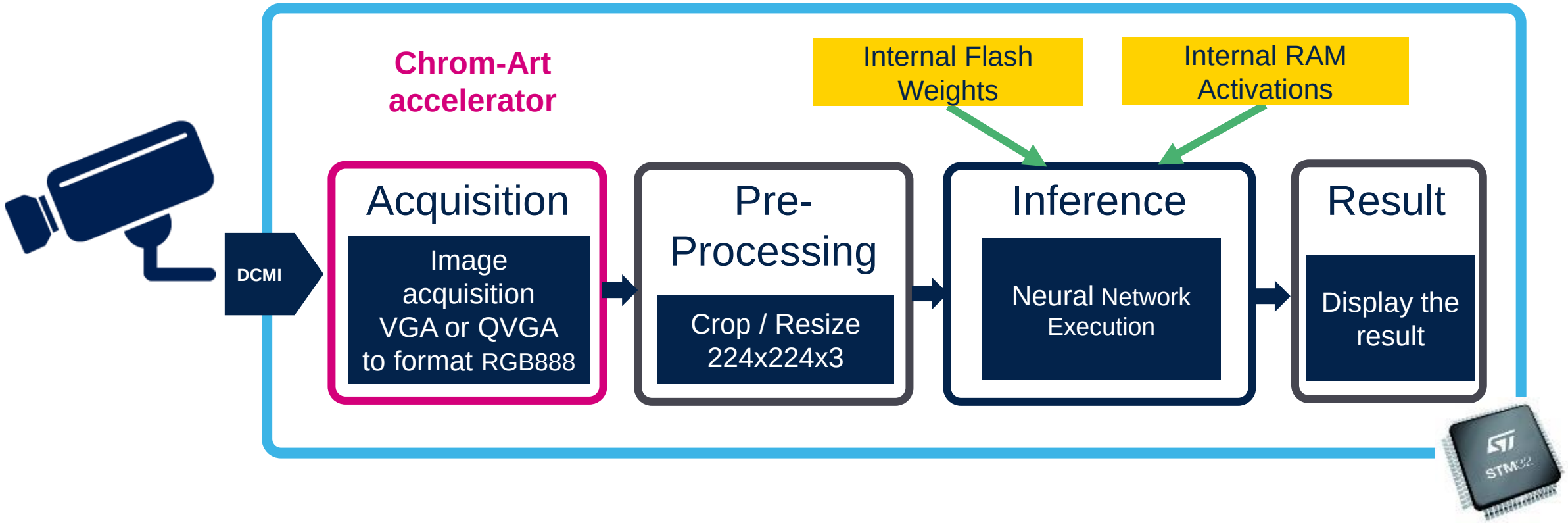
Update only 4 files:

`network.h`	public API
`network.c`	network topology
`network_data.h`	parameters
`network_data.c`	weights

life.augmented

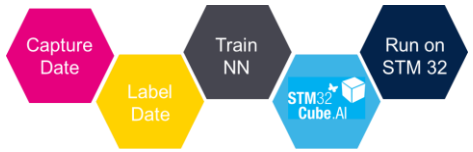


FP-AI-VISION1 Memory update

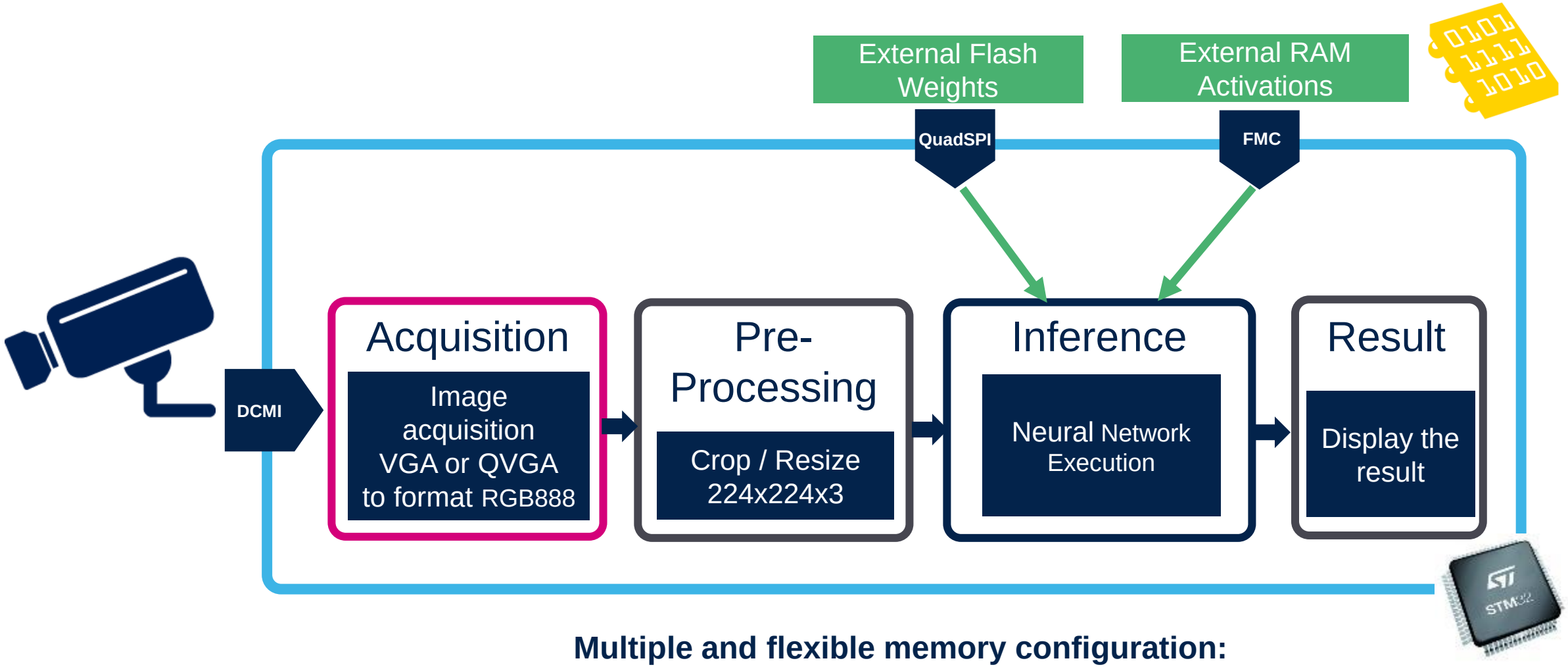


Optimized quantized model is fitting in internal memory: 153 ms

life.augmented

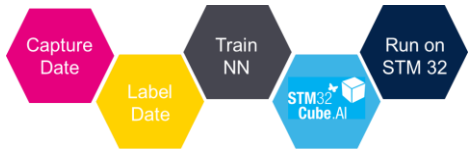


FP-AI-VISION1 Memory flexibility



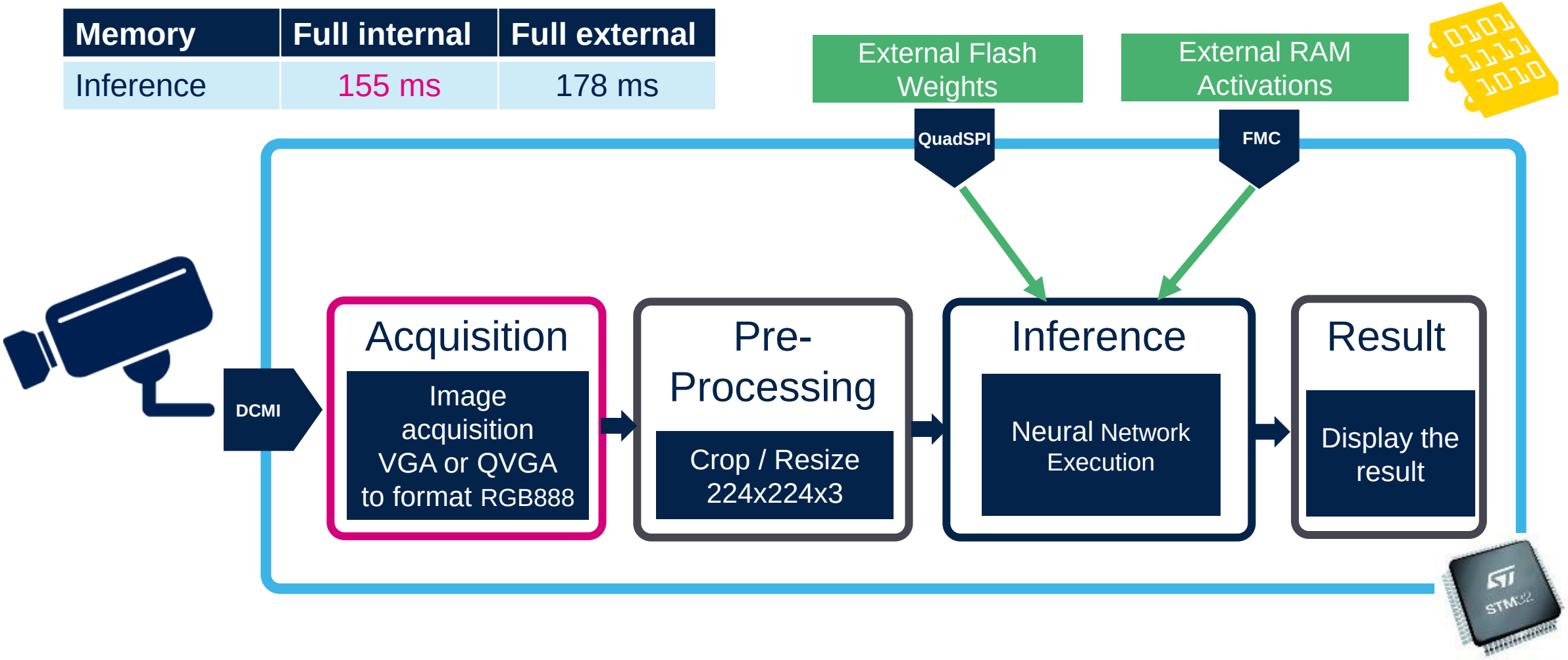
Multiple and flexible memory configuration: full internal, mixed, copy-before-run…

life.augmented



FP-AI-VISION1 Memory flexibility

Memory	Full internal	Full external
Inference	155 ms	178 ms

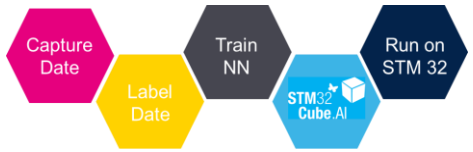


Values measured on FP-AI-VISION1 1.0.0 with the optimized MobileNet Derivative model
Inference times on the STM32H7@ 400MHz, Dual-QSPI DDR @ 50MHz, SDRAM @ 100MHz

Demo

STM32 Cube.AI

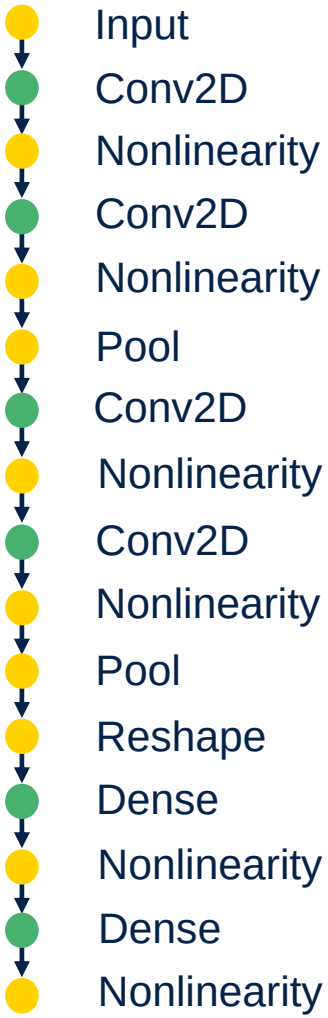
life.augmented



Demo
Handwriting character recognition

Neural Network
- ST CNN
- EMNIST dataset (36 classes)

Implementation
- Exploits touch screen captured as image of size 32x32
- 36 classes: numbers and capital letters



STM32 Cube.AI NN
- Computational complexity 73k MACC
- Memory footprint: 26 KB RAM, 291 KB Flash

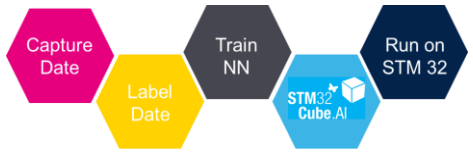
Performance on STM32L562
- 1 inference per image
- STM32L562 110MHz Cortex-M33

life.augmented

Conclusion

STM32 Cube.AI

life.augmented

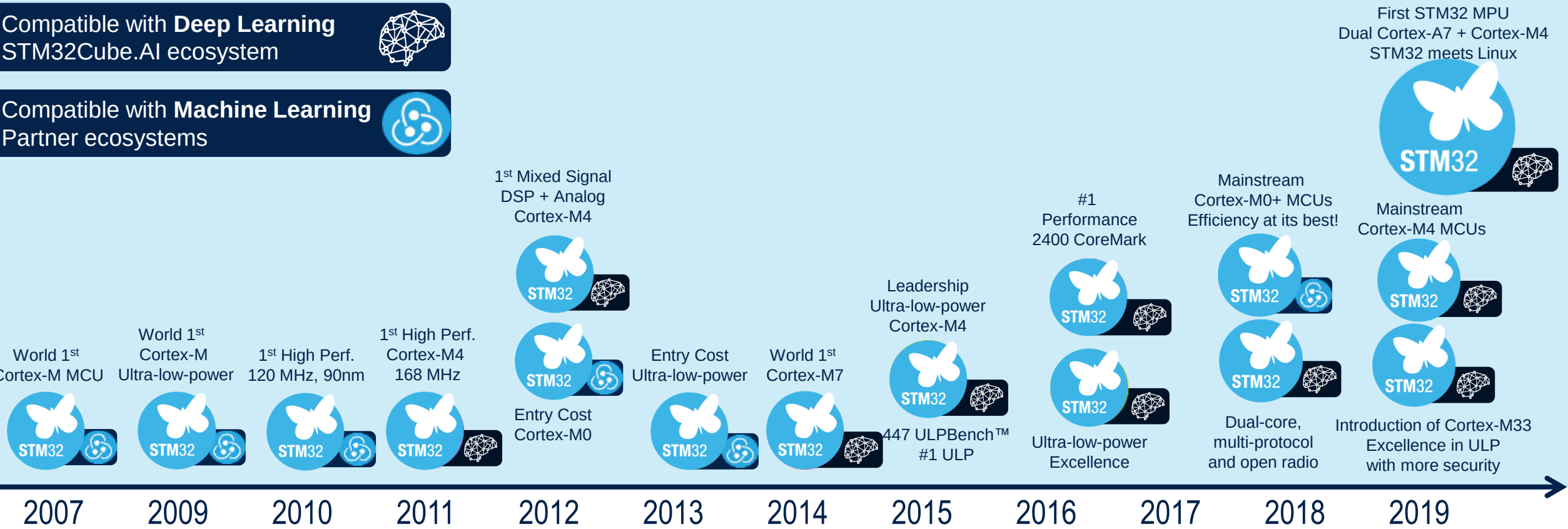


Making AI accessible now

Leader in Arm® Cortex®-M 32-bit General Purpose MCU

Compatible with **Deep Learning** STM32Cube.AI ecosystem

Compatible with **Machine Learning** Partner ecosystems

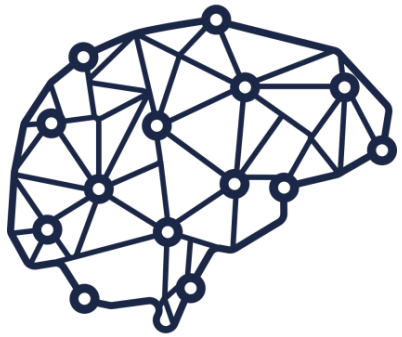


More than 40,000 customers

Over 4 Billion STM32 shipped since 2007

life.augmented

Access the STM32Cube.AI ecosystem



Ecosystem Partner Use cases

- Smart Industry
- Internet of Things
- Smart Things
- Smart Home and City

Support for Deep Learning frameworks

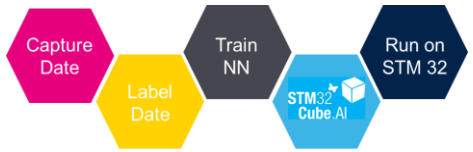
- ONNX
- PyTorch
- mxnet
- Microsoft Cognitive Toolkit
- Lasagne
- MATLAB
- Caffe2
- TensorFlowLite
- Caffe
- K

ST Development Hardware

- SensorTile
- IoTNode
- SensorTile.Box
- STM32H747 Discovery kit
- OpenMV H7 CAM
- STM32MP1 Discovery kit

life.augmented

STM32 Releasing your creativity



ST toolbox for Neural Networks
More than just a conversion tool

- STM32Cube.AI to convert an NN in **optimized code**
- **Interoperability** with state-of-the-art Deep Learning design frameworks
- Support of **quantization**
  - Decrease memory requirement up to ÷4
  - Decrease latency and power consumption up to ÷3

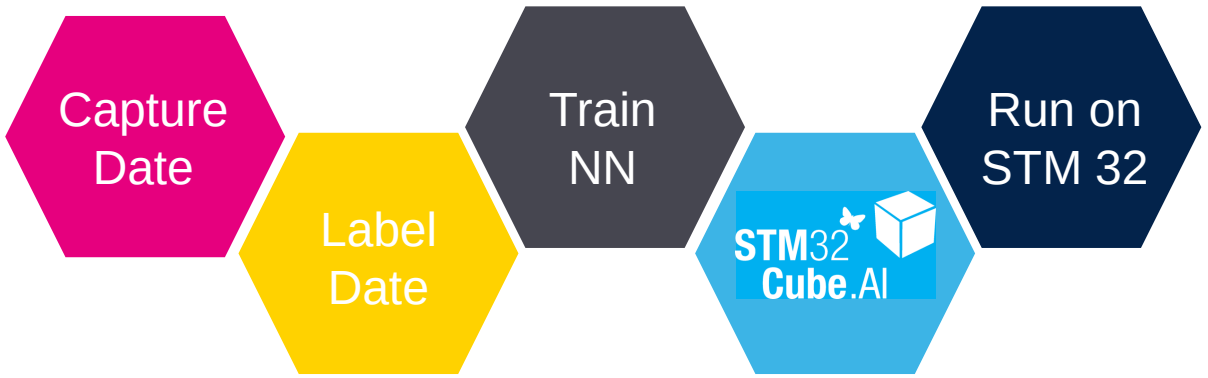
- Software examples for Quick prototyping
- **Audio, Motion and Vision** Function packs **on ST development hardware**
- **STM32 Community with dedicated Neural Networks topic**

- **AI Partner Program**
- Expertise in **Machine Learning and STM32 solutions**

life.augmented

For more information

www.st.com/STM32CubeAI



Thank you

life.augmented