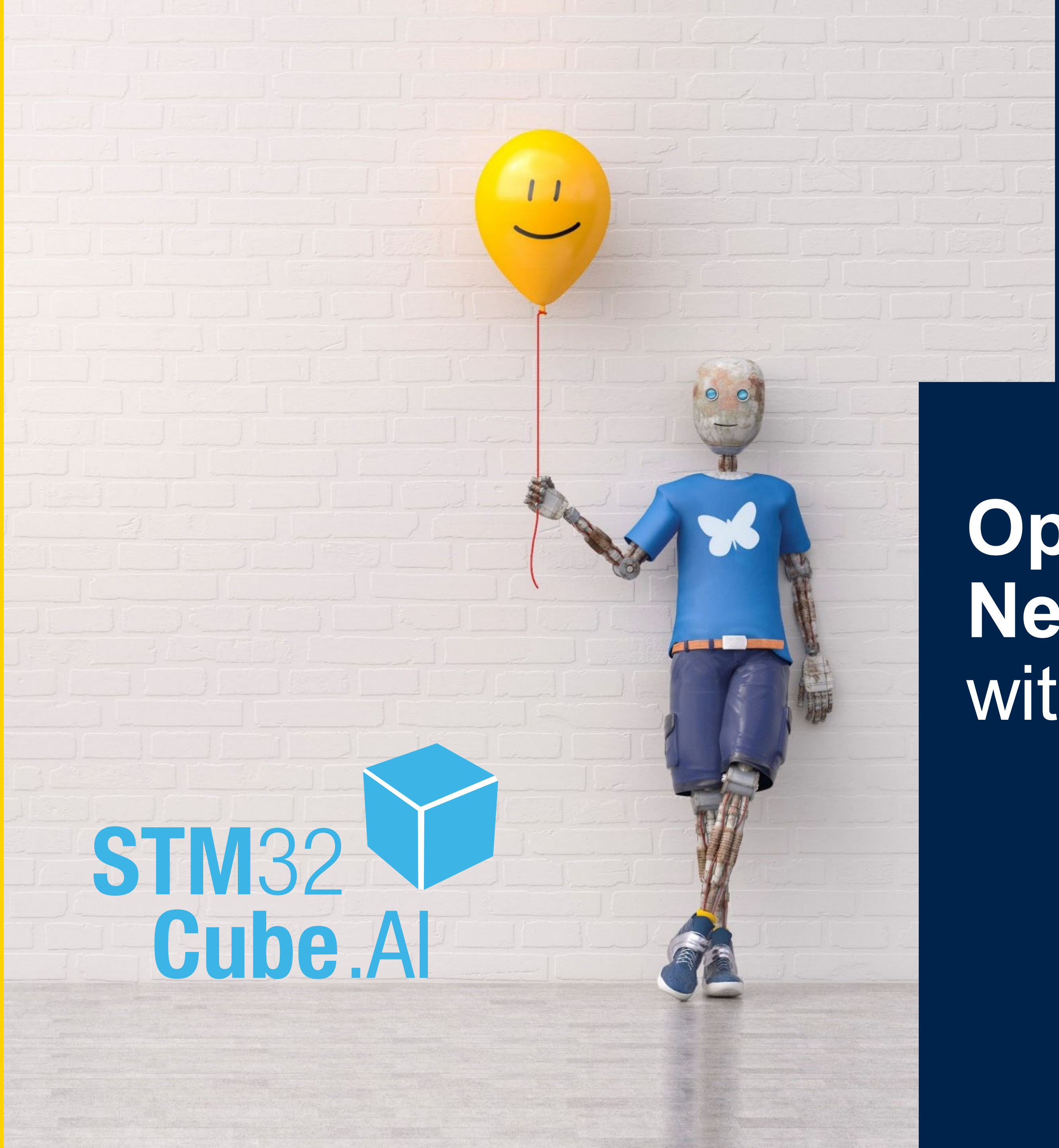


life.augmented

Optimized Neural Networks on STM32 with STM32Cube.AI

STM32

Introduction to Edge AI

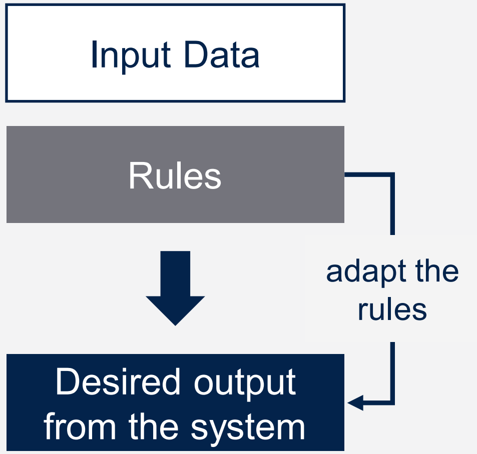
life.augmented

A new way to add environment awareness to your products

From rule-based engineering to data-driven engineering

Standard programming

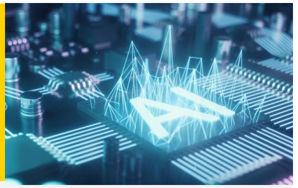
Handcrafted rules based on experience

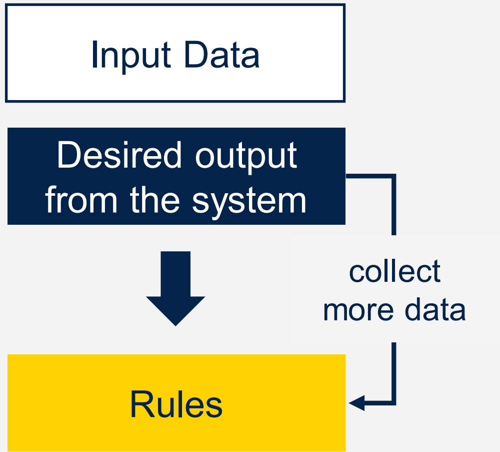


- Requires domain expertise to code
- Need to rewrite if environment evolves

Machine Learning

Rules learnt from real-world data

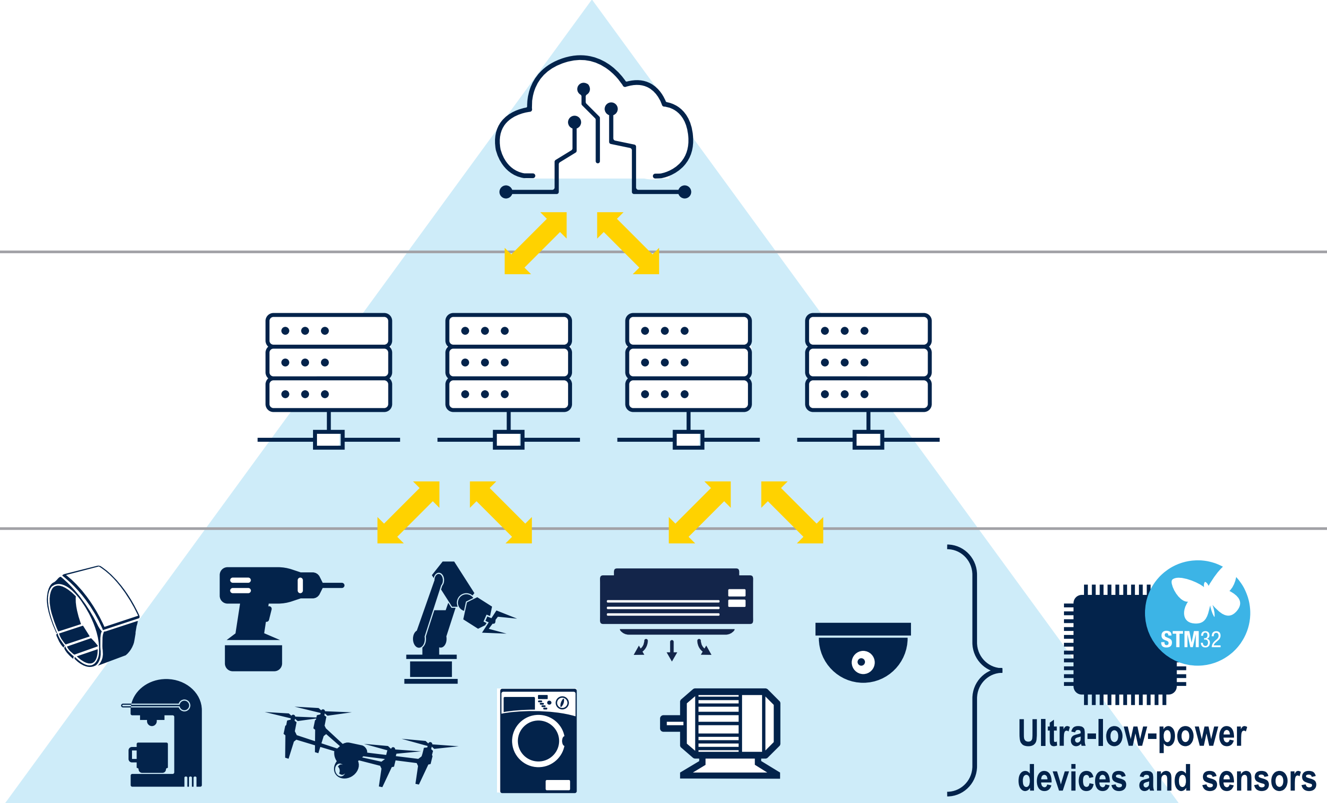




- Generate code from real-world observations
- Re-learn from data if environment evolves

Distributed Artificial Intelligence approach

Leverage billions of devices at the Edge!



Ultra-low-power devices and sensors

Thousands

Millions

100 Billions

Data Center Cloud

Analytics, storage, compute

Edge Nodes

IoT gateways, micro datacenters

Edge Things

Real time, local processing

life.augmented

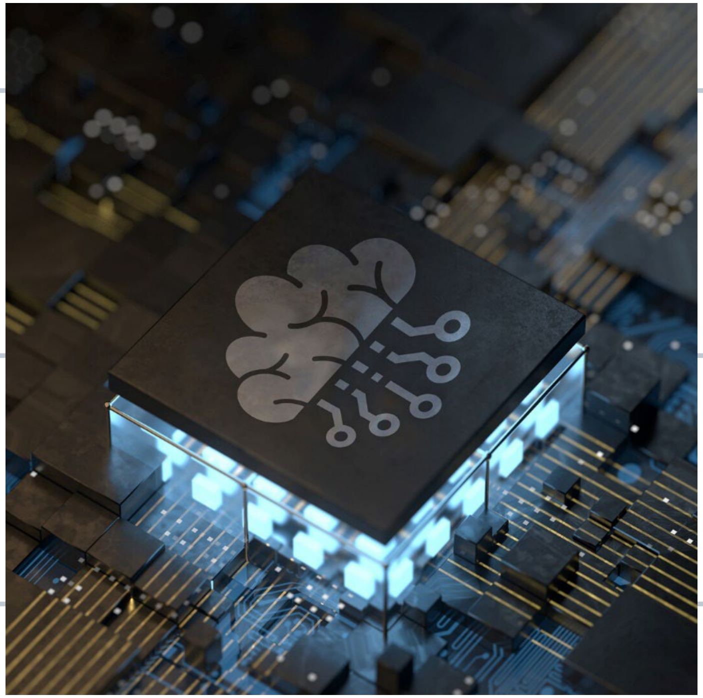
Artificial intelligence at the Edge

Moving part of Artificial Intelligence closer to the data acquisition brings several benefits

Ultra-low latency
Real-time applications

More reliability

Security of data
No sharing in the cloud



Privacy by design
GDPR compliant

Sustainable on energy
Low-power consumption

Better user experience

life.augmented

STM32 Cube.AI

life.augmented

A tool to seamlessly integrate AI in your projects

Machine Learning

scikit learn

ONNX

Deep Learning

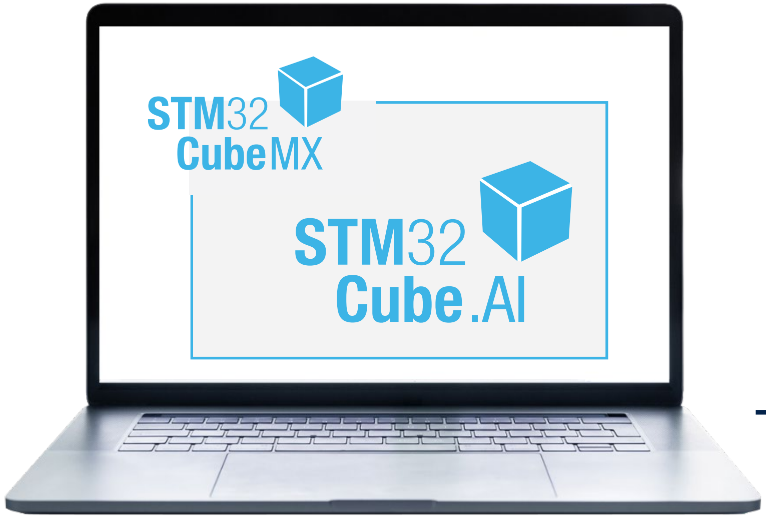
MATLAB

PyTorch

TensorFlow Lite

Keras

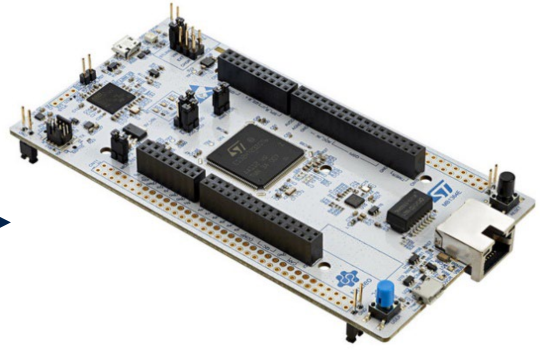
ONNX



Select MCU & upload your model

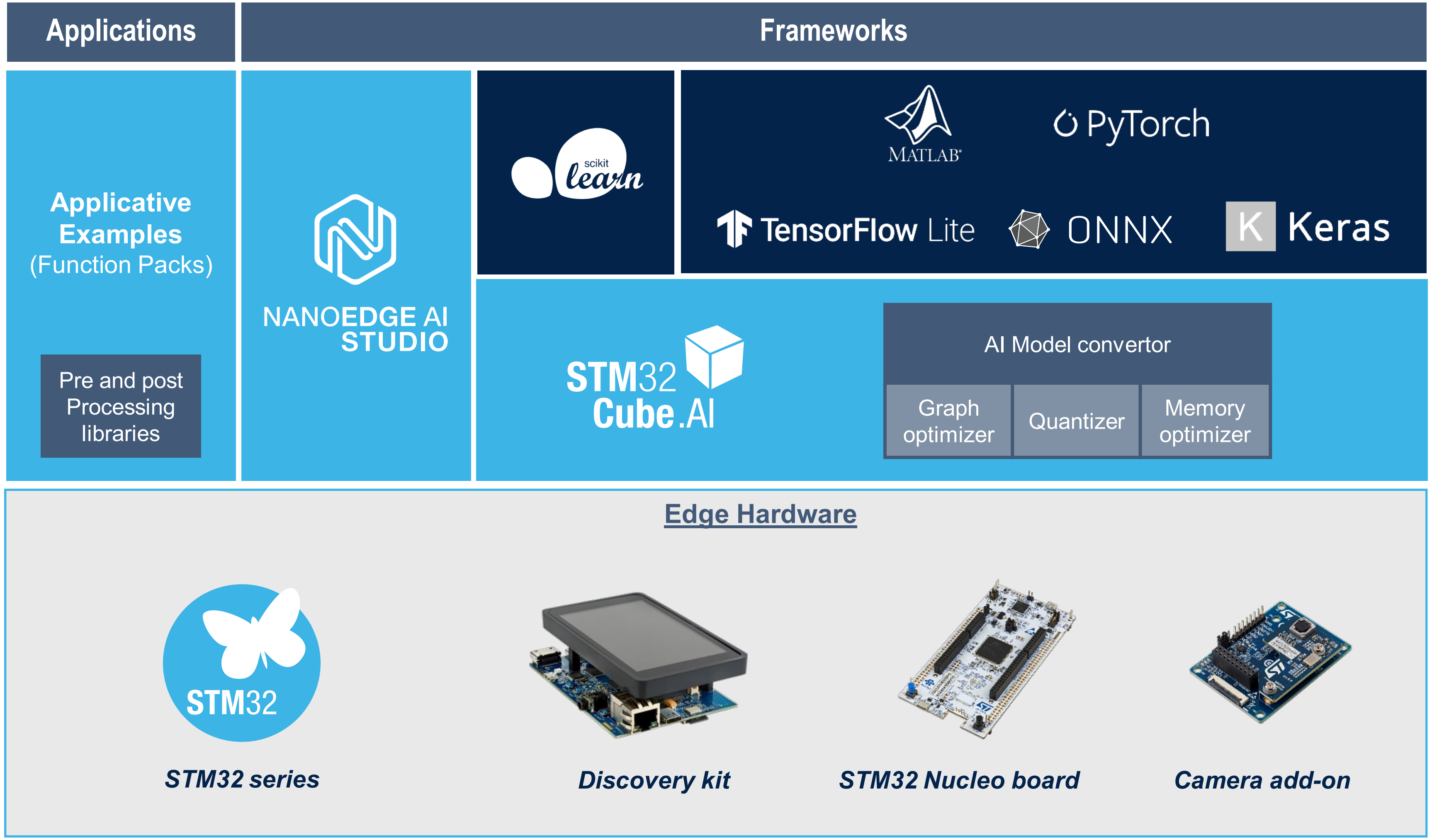
Optimize and validate

Generate project and deploy



HIGH PERFORMANCE EDGE AI PRODUCT

STM32 comprehensive AI ecosystem



Applications

Frameworks

Applicative Examples
(Function Packs)

Pre and post Processing libraries

NANOEDGE AI STUDIO

scikit learn

MATLAB

PyTorch

TensorFlow Lite

ONNX

Keras

STM32 Cube.AI

AI Model convertor

Graph optimizer

Quantizer

Memory optimizer

Edge Hardware

STM32 series

Discovery kit

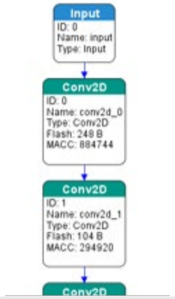
STM32 Nucleo board

Camera add-on

The 3 pillars of STM32Cube.AI

Graph optimizer

Automatically improve performance through graph simplifications & optimizations that benefit STM32 target HW architectures



Input
ID: 0
Name: input
Type: Input

Conv2D
ID: 0
Name: conv2d_0
Type: Conv2D
Flash: 248 B
MACC: 884744

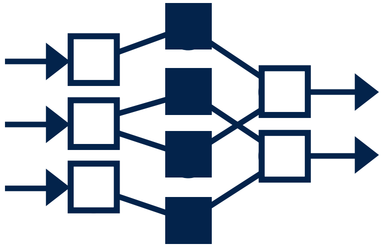
Conv2D
ID: 1
Name: conv2d_1
Type: Conv2D
Flash: 104 B
MACC: 294920

Conv2D

- Auto graph rewrite
- Node/operator fusion
- Layout optimization
- Constant-folding…
- Operator-level info to fine-tune memory footprint and computation

Quantized model support

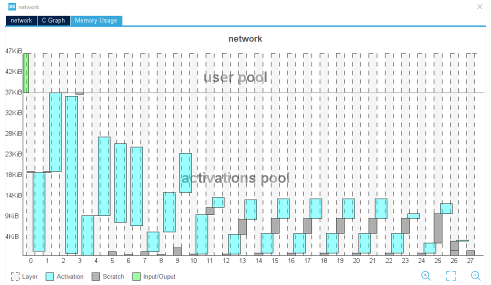
Import your quantized ANN to be compatible with STM32 embedded architectures while keeping their performance



- From FP32 to Int8
- Minimum loss of accuracy
- Code validation on target
  - Latency
  - Accuracy
  - Memory usage

Memory optimizer

Optimize memory allocation to get the best performance while respecting the constraints of your embedded design



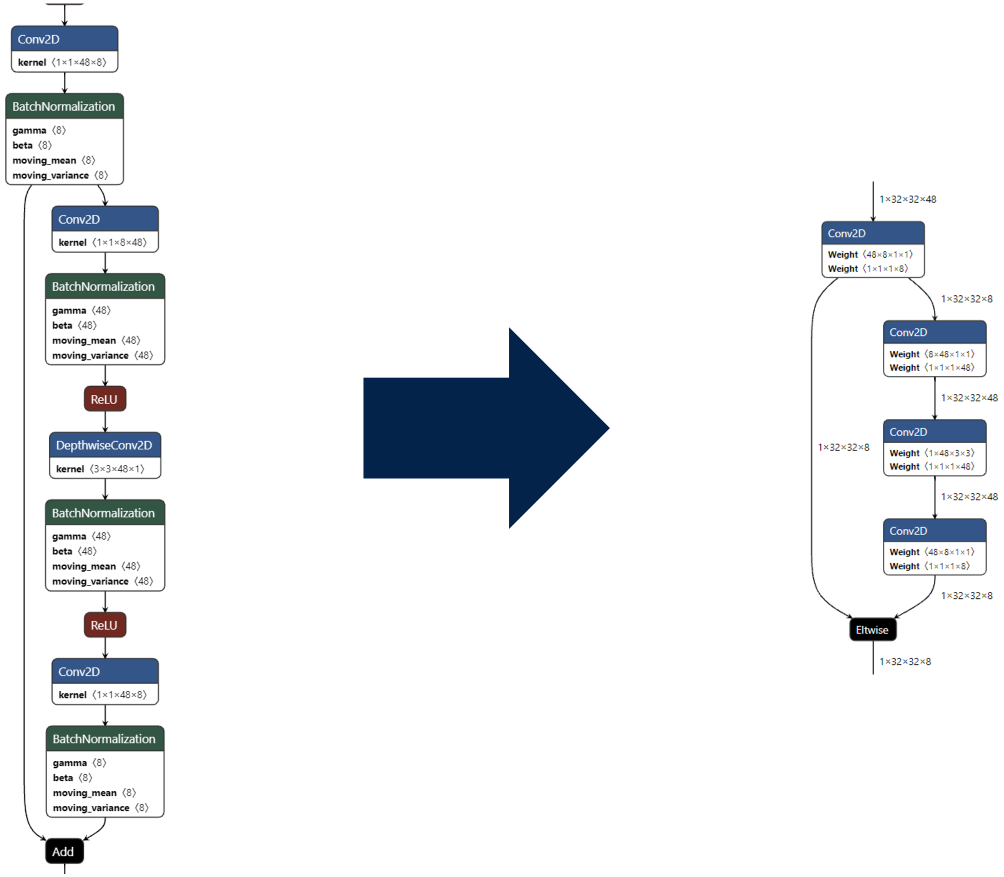
- Memory allocation
- Internal/external memory repartition
- Model-only update option

STM32Cube.AI is **free of charge**, available both in graphical interface and in command line.

life.augmented

Graph optimizer

Squeeze your graph to fit into an MCU!



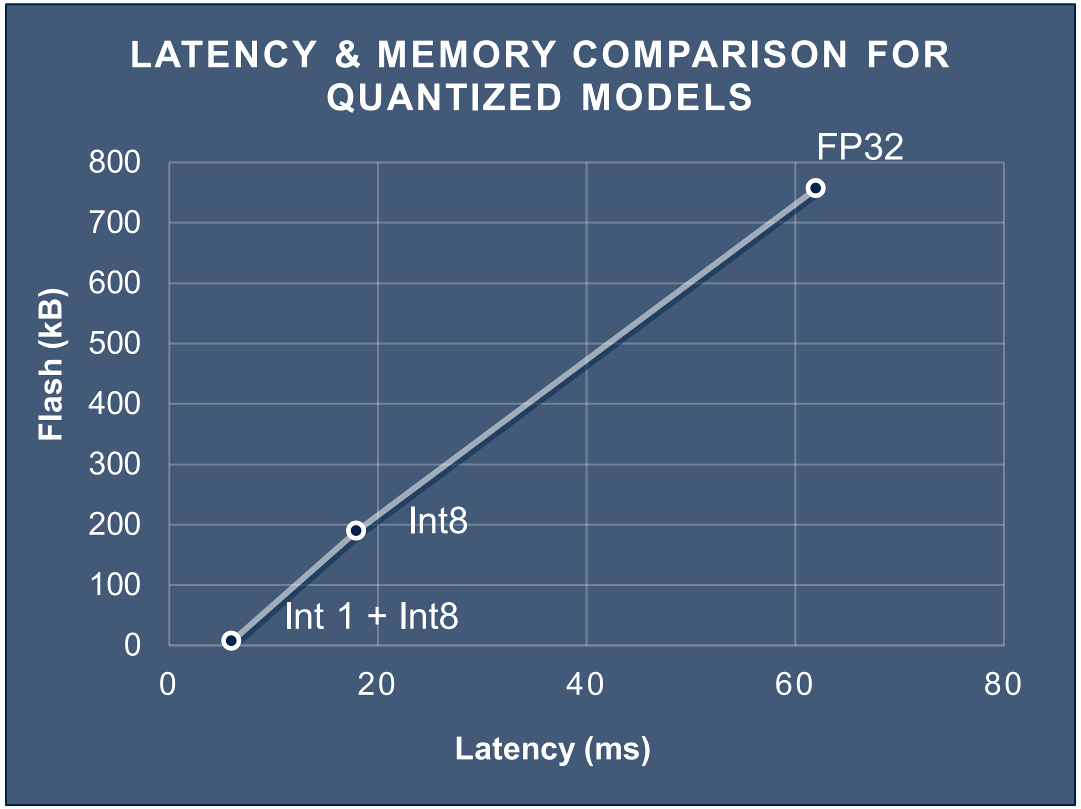
Fully automated process in the STM32Cube.AI workflow

- Your original graph is optimized at the very early stage for optimal integration into STM32 MCU/MPU
- Loss-less conversion

life.augmented

Quantized model support

Simply use quantized networks to reduce memory footprint and inference time



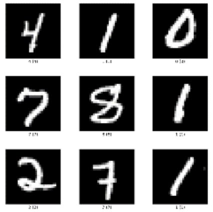
STM32Cube.AI support quantized Neural Network models with **all parameter formats**:

- FP32
- Int8
- Mixed binary Int1 to Int8 (Qkeras*, Larq.dev*)

**Please contact edge.ai@st.com to request the relevant version of STM32Cube.AI*

HW Target: NUCLEO-STM32H743ZI2
Model: Low complexity handwritten digit reading
Freq: 480 MHz
Accuracy: >97% for all quantized models

Tested database: MNIST dataset

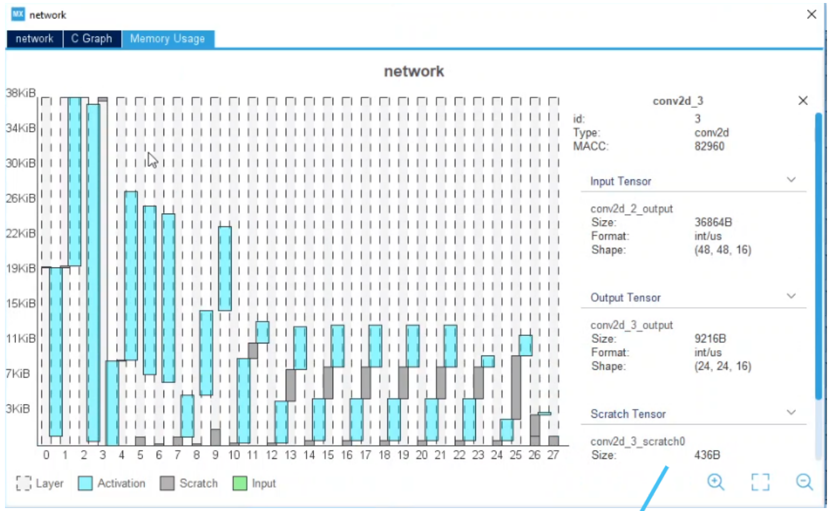


MNIST dataset

life.augmented

Memory optimizer

Optimize performance easily with the memory allocation tool



conv2d_3

id: 3
Type: conv2d
MACC: 82960

Input Tensor

conv2d_2_output
Size: 36864B
Format: int/us
Shape: (48, 48, 16)

Output Tensor

conv2d_3_output
Size: 9216B
Format: int/us
Shape: (24, 24, 16)

Scratch Tensor

conv2d_3_scratch0
Size: 436B

Model RAM consumption per layer

- Easily identify most critical layers

Model memory allocation

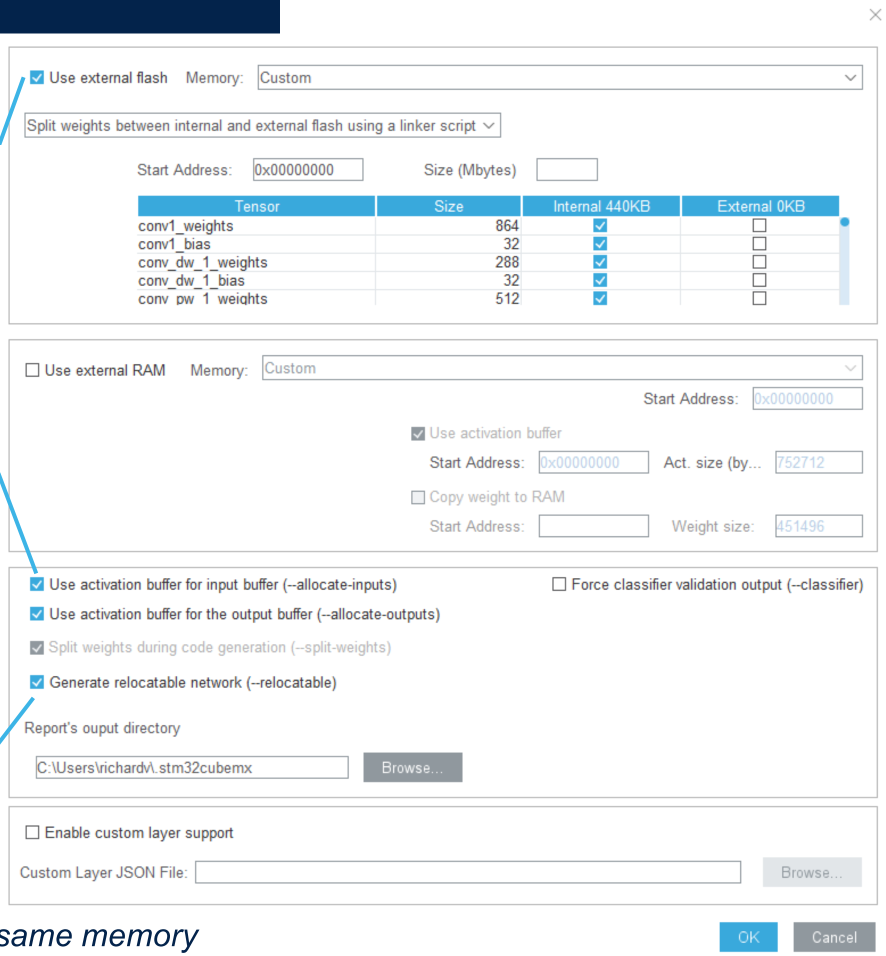
- Set your external memory
- Map in non-contiguous internal flash section
- Partition internal vs external flash memories

Re-use model input buffer to store activation data*

- Minimize RAM requirements

Relocatable network

- A separate binary is generated for the library and the network to enable standalone model upgrade



☑ Use external flash Memory: Custom

Split weights between internal and external flash using a linker script

Start Address: 0x00000000 Size (Mbytes):

Tensor	Size	Internal 440KB	External 0KB
conv1_weights	864	☑	☐
conv1_bias	32	☑	☐
conv_dw_1_weights	288	☑	☐
conv_dw_1_bias	32	☑	☐
conv_pw_1_weights	512	☑	☐

☐ Use external RAM Memory: Custom

Start Address: 0x00000000

☑ Use activation buffer

Start Address: 0x00000000 Act. size (by... 752712

☐ Copy weight to RAM

Start Address: Weight size: 451496

☑ Use activation buffer for input buffer (--allocate-inputs)

☐ Force classifier validation output (--classifier)

☑ Use activation buffer for the output buffer (--allocate-outputs)

☑ Split weights during code generation (--split-weights)

☑ Generate relocatable network (--relocatable)

Report's ouput directory: C:\Users\richardv\.stm32cubemx Browse...

☐ Enable custom layer support

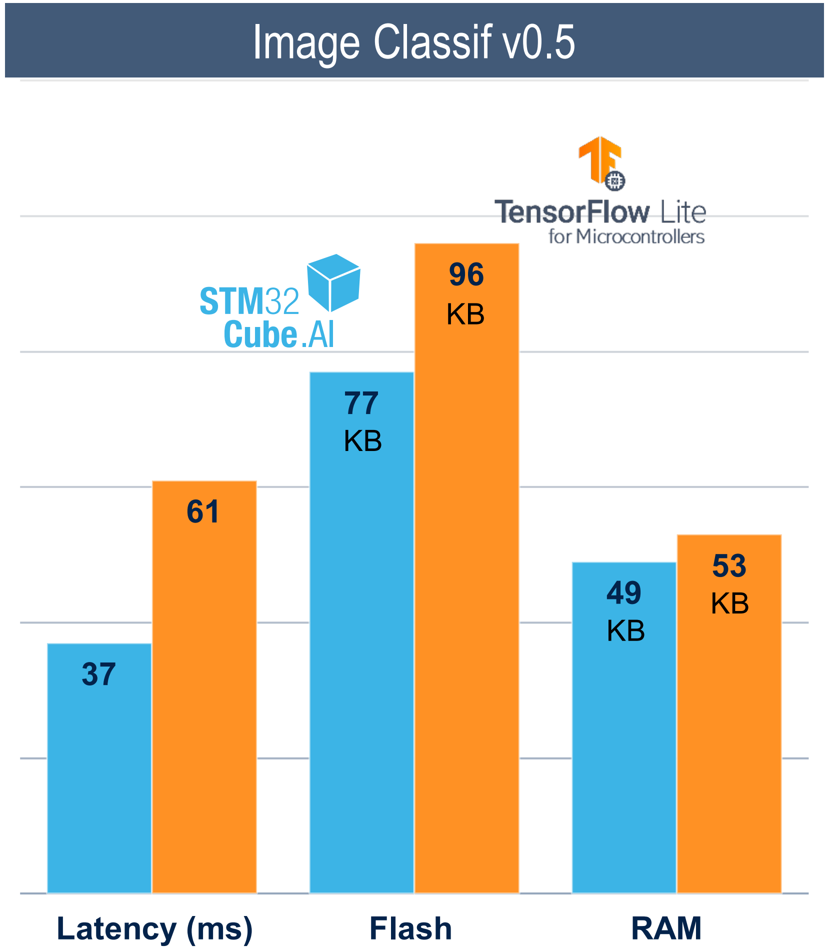
Custom Layer JSON File: Browse...

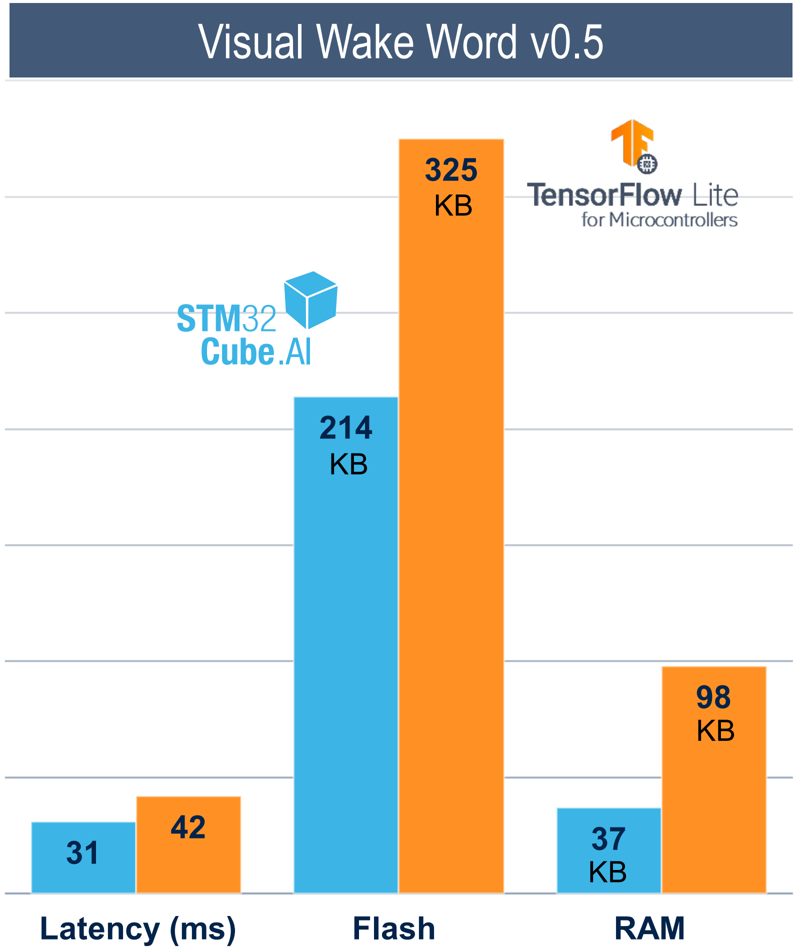
OK Cancel

** Requires input and activation buffers in same memory*

life.augmented

STM32Cube.AI
Get the best performance on STM32





HW Target: STM32H723
Flash: 1Mbyte
RAM: 564 Kbytes
Freq: 550 MHz

SW Version:
X-Cube.AI v 7.0.0
TFLm v2.5.0

** the lower the better*

life.augmented

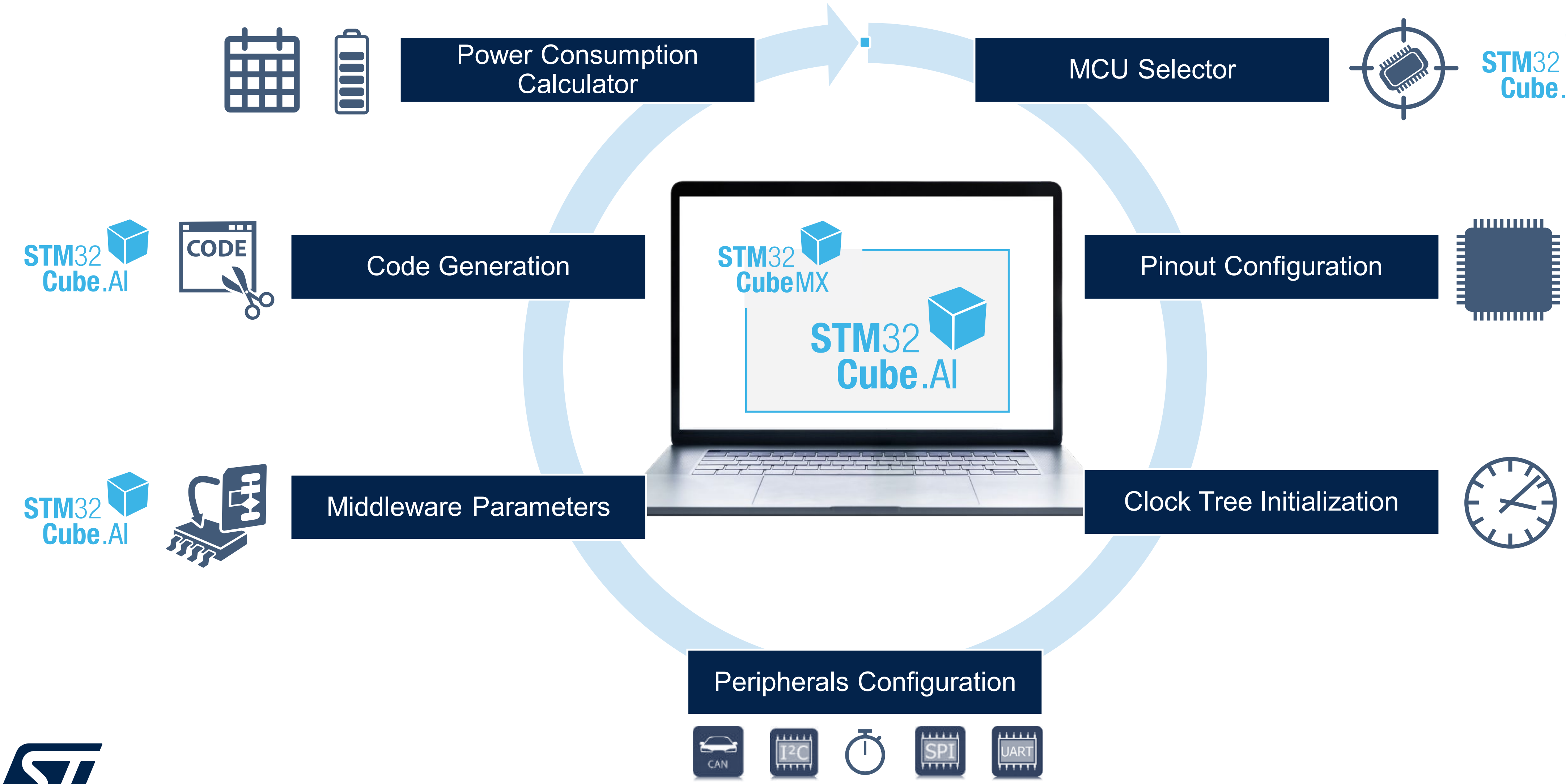
Making Edge AI accessible to all STM32 portfolio

STM32Cube.AI is compatible with all STM32 series

MPU							STM32**MP1** 4158 CoreMark Up to 800 MHz Cortex –A7 209 MHz Cortex –M4
High Perf MCUs	STM32**F3** 245 CoreMark 72 MHz Cortex-M4	STM32**G4** 569 CoreMark 170 MHz Cortex-M4	STM32**F2** Up to 398 CoreMark 120 MHz Cortex-M3	STM32**F4** Up to 608 CoreMark 180 MHz Cortex-M4	STM32**F7** 1082 CoreMark 216 MHz Cortex-M7	STM32**H7** Up to 3224 CoreMark Up to 550 MHz Cortex -M7 240 MHz Cortex -M4	
	Optimized for mixed-signal Applications						
Mainstream MCUs	STM32**F0** 106 CoreMark 48 MHz Cortex-M0	STM32**G0** 142 CoreMark 64 MHz Cortex-M0+	STM32**F1** 177 CoreMark 72 MHz Cortex-M3				
Ultra-low Power MCUs	STM32**L0** 75 CoreMark 32 MHz Cortex-M0+	STM32**L1** 93 CoreMark 32 MHz Cortex-M3	STM32**L4** 273 CoreMark 80 MHz Cortex-M4	STM32**L4+** 409 CoreMark 120 MHz Cortex-M4	STM32**L5** 443 CoreMark 110 MHz Cortex-M33	STM32**U5** 651 CoreMark 160 MHz Cortex-M33	
Wireless MCUs			STM32**WL** 162 CoreMark 48 MHz Cortex-M4 48 MHz Cortex-M0+	STM32**WB** 216 CoreMark 64 MHz Cortex-M4 32 MHz Cortex-M0+			

■ Latest product generation

STM32Cube.AI
The STM32CubeMX expansion pack for ML



Power Consumption Calculator

MCU Selector

Code Generation

Pinout Configuration

Middleware Parameters

Clock Tree Initialization

Peripherals Configuration

Integrate your ML models more easily with our application-oriented code examples

Time series-based monitoring

FP-AI-MONITOR1

- Predictive maintenance and much more sensor-monitoring apps
- Runs Libraries from NanoEdge™ AI Studio

Audio and Sensing

FP-AI-SENSING1

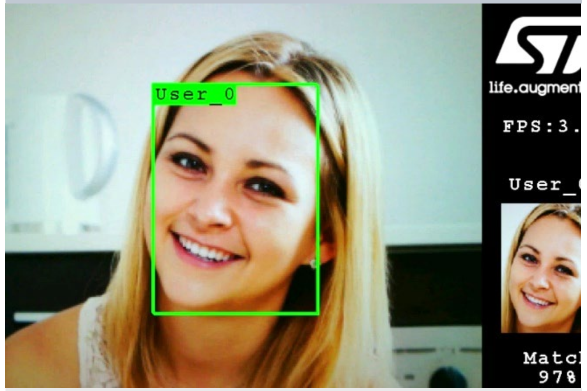
- Human Activity Recognition
- Acoustic Scene Classification
- Data logging, labeling and result on BLE applications

Computer Vision

FP-AI-VISION1

- Food recognition (CNN)
- Person presence detection (CNN)
- People counting (Object detection NN)
- Image processing Library

Face recognition



FP-AI-FACEREC1

- Face detection and recognition
- Fully functional without cloud connection

We provide everything to kick off your project

Design documentation

Getting started

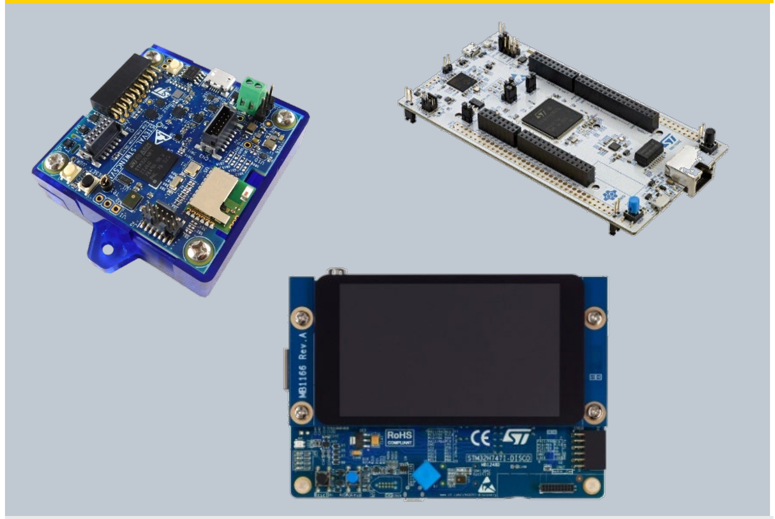
Be guided step-by-step to learn STM32 ecosystem

Development zone

Get started on application development and project sharing

- **Wiki by ST** is a great forum to learn and start developing AI on STM32!
- Videos of application examples
- Massive Open Online Course (MOOC)

Hardware and software tools



- Evaluation platforms for STM32 MCU/MPU
- Extra sensor boards
- Full software suite

Support & Updates

- **ST Community**: STM32 ML & AI group
- Distributor certified FAE
- Support center
- Newsletter

STM32 Cube.AI

What’s new in STM32Cube.AI v7.1.0?

v7.1.0

Bringing STM32Cube.AI to all STM32 and improving overall performances

Now supporting entry level STM32

- Introducing the support for STM32 arm Cortex-M0 and arm Cortex-M0+
- STM32Cube.AI can now generate optimized code for STM32C0, STM32F0, STM32L0 and STM32G0 series

Improved user experience and performance tuning

- Added advanced support for splitting the activation buffer over several memory segments (multi-heap support) allowing full manipulation of the different on-board memories of the STM32H7 for example.

Up-to-date and improved code generation

- Support TensorFlow Lite micro v2.7 runtime and ONNX 1.9
- Support of more Keras, TensorFlow Lite and ONNX layers (refer to documentation for exhaustive list)
- Extend support of scikit-learn algorithms with new ONNX-ML operators

life.augmented

Don't go alone

We have created a network of companies to support you

Partner Program

Trust our **authorized partners** to ensure the success of your project. Learn more at st.com/stm32ai

Wish to discuss a co-development partnership for ML/AI projects? Contact us at edge.ai@st.com

Releasing your creativity

- /STM32
- @ST_World
- community.st.com
- www.st.com/STM32ai
- wiki.st.com/stm32
- github.com/STMicroelectronics
- Videos
- STM32Cube.AI blog articles

STM32

life.augmented

Our technology starts with You

Find out more at www.st.com/stm32ai

life.augmented