Optimize neural networks on STM32 with STM32Cube.AI
Introduction to Edge AI
Signals turning into data

Embedded applications will collect more data in the future

Growing demand for data-driven insights

Increasing use of sensors

Proliferation of IoT devices
AI is offering the best approach to process this growing amount of data.

Traditional approaches show their limitations:
- when dealing with large datasets
- when the phenomena are too complex

Algorithms and predefined models to analyze data and make predictions or decisions.

Machine learning algorithms to automatically learn patterns and relationships from the data.

AI-based data processing offers a more flexible and powerful approach to analyzing and making decisions from large data collection.
The rise of Edge AI

- **Ultra-low latency**
  - Real-time applications

- **Reduced data transmission**
  - Generate meaningful information

- **Enhanced privacy and security**
  - No data sharing in the cloud

- **Power efficiency**
  - Low-data / Low-power

- **Improved accuracy**
  - Analyze data from a wide range of sensors and sources

Edge AI will benefit many application domains:

- **Industrial maintenance**
  - Condition monitoring
  - Predictive maintenance

- **Control systems**
  - From home heating systems to industrial machines

- **Internet of Things (IoT)**
  - Smart cities, smart buildings, connected homes, and industrial automation
AI development workflow – ST software offering

1. Data preparation
   - Data acquisition
   - Data processing

2. Data science
   - Model selection and training
   - Model validation

3. Model implementation
   - Model library creation
   - Model inference

---

NANOEDGE AI STUDIO
Automated Edge AI software

STM32 Cube.AI
Edge AI toolkit

STM32
All STM32 MCUs
ST ecosystem ease your AI to reach production level

### Edge AI toolkit
for model optimization on STM32

<table>
<thead>
<tr>
<th>Key benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Get optimized C-code from your trained model</td>
</tr>
<tr>
<td>✓ Desktop and online versions</td>
</tr>
<tr>
<td>✓ Benchmark service on remote hardware (online version)</td>
</tr>
<tr>
<td>✓ On-device performance validation</td>
</tr>
</tbody>
</table>

### Automated ML software
for end-to-end Edge AI solution design on STM32

<table>
<thead>
<tr>
<th>Key benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ The easiest way to integrate AI into your system</td>
</tr>
<tr>
<td>✓ Save resources and development cost</td>
</tr>
<tr>
<td>✓ Reach the highest performance with the automated model finder embedded in the tool</td>
</tr>
</tbody>
</table>

### Application domain

- All

- Time series (except voice and speech)

### Business model

- Free of charge

- Free for prototyping on STM32 dev boards

- Production requires right of use
A proven technology adopted by multiple clients

**INDUSTRIAL | DEMO**
Fan anomaly detection based on vibrations
Learn to detect abnormal behavior at the edge on a vibrating machine.

**INDUSTRIAL | CUSTOMER**
AI solution for industrial predictive maintenance with NKE Watteco
Predictive maintenance solution for industrial equipment.

**TRANSPORTATION | CUSTOMER**
AI solution for monitoring automatic doors with Crouzet
Predictive maintenance on motors for automatic door motors.

**INDUSTRIAL | DEMO**
Anomaly detection in an electric motor
Current sensing to detect abnormal behaviors in motors.

**INDUSTRIAL | CUSTOMER**
AI solution for industrial predictive maintenance with Oxytronic
Predictive maintenance solution for industrial equipment.

**SMART OFFICE | CUSTOMER**
People flow counting Sensor with Schneider Electric
An innovative approach to measure people flows using an in-house thermal sensor.

**SMART CITY | DEMO**
Acoustic scene classification
Identify different environments (indoor, outdoor, in-car) using a simple microphone.

**WEARABLES | DEMO**
Human Activity Recognition
Easily identify 5 different activities with a 3D accelerometer.

**INDUSTRIAL | DEMO**
People presence detection (visual wake word)
Human detection on high-performance MCU.

**INDUSTRIAL | DEMO**
Aftermarket wireless digit reader
Equip meters with aftermarket wireless & low-power readers.
STM32Cube.AI
AI optimization tool for STM32 portfolio
STM32Cube.AI overview

STM32Cube.AI
The original desktop front end AI optimizer for STM32

STM32Cube.AI Developer Cloud
The brand-new online AI services front end for STM32

X-CUBE-AI
for STM32Cube.MX

X-CUBE-AI
Command Line Interface

ST model zoo
Web GUI + REST API
Board farm

Core engine technology
One tool – two versions to deploy AI on STM32

Load your trained neural network model

or pick one from STM32 model zoo (AI models library)

Optimize and validate your NN model

Generate optimized code for STM32

STM32Cube.AI for desktop

STM32Cube.AI Developer Cloud

STM32Cube ecosystem

Command Line Interface

Online platform

REST API

Benchmarking tool

Optimized model code for STM32
STM32Cube.AI
Get the best AI performance on STM32

Image Classif v1.0 MLPerf Tiny

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Flash (Kbytes)</th>
<th>RAM (Kbytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>121</td>
<td>155</td>
<td>55</td>
</tr>
</tbody>
</table>

Visual Wake Word v1.0 MLPerf Tiny

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Flash (Kbytes)</th>
<th>RAM (Kbytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>72</td>
<td>58</td>
</tr>
<tr>
<td>283</td>
<td>378</td>
<td>102</td>
</tr>
</tbody>
</table>

**UP TO 60% faster inference time***

**UP TO 20% space freed-up in flash and RAM***

**HW Target:** STM32H7A3
**Flash:** 2 Mbytes
**RAM:** 1.4 Kbytes
**Freq:** 280 MHz

**SW Version:** X-Cube.AI v 7.3.0
TFLm v2.10.0

* versus TensorFlow Lite for microcontroller
STM32Cube.AI is free of charge, available both in graphical interface and in command line.
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
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*)

LATENCY & MEMORY COMPARISON FOR QUANTIZED MODELS

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

Tested database: MNIST dataset

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

MNIST dataset
Memory optimizer

Optimize performance easily with the memory allocation tool

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

Model RAM consumption per layer
- Easily identify most critical layers

* Requires input and activation buffers in same memory
Performance benchmarking made simple
STM32Cube.AI Developer Cloud

The unique possibility to evaluate the performance of models remotely, on real STM32 boards

Get the real inference time from optimized models running on STM32

Benchmark models on a large variety of STM32 boards
• Find the most appropriate board for your application

Get access to the most recent devices
• A board farm is constantly updated with the latest available boards
Start with edge AI optimized models
STM32 model zoo

A collection of application-oriented models optimized for STM32

- Human activity
  - Motion Sensing
- Image classification
  - Computer vision
- Audio event detection
  - Audio classification
- Object detection
  - Computer vision

Hosted on Github

Model training scripts
- Scripts to generate and validate

Application code example
- Designed to host optimized NN models
- Automatically generated from the trained models
- Easy to deploy for end-to-end evaluation
We provide everything to kick off your project

**Design documentation**
- 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
What’s new in STM32Cube.AI v8.0.0?

Bringing a higher degree of versatility with STM32Cube.AI

# ONNX quantized models' support

Introducing the support of ONNX Tensor-oriented file format (QDQ):
- ONNX models quantized with ONNX runtime post-training quantization.

# Up-to-date and improved code generation

- Support for TensorFlow 2.11 models
- Support Keras.io 2.11
- Support ONNX Runtime 1.13.1
- New kernel performance improvements.
# Making Edge AI accessible to all STM32 portfolio

## Take advantage of STM32Cube.AI on all STM32 series

<table>
<thead>
<tr>
<th>STM32 Series</th>
<th>CoreMark</th>
<th>Clock Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High Perf MCUs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM32F2</td>
<td>Up to 398 CoreMark</td>
<td>120 MHz Cortex-M3</td>
</tr>
<tr>
<td>STM32F3</td>
<td>245 CoreMark</td>
<td>72 MHz Cortex-M4</td>
</tr>
<tr>
<td>STM32F4</td>
<td>Up to 608 CoreMark</td>
<td>180 MHz Cortex-M4</td>
</tr>
<tr>
<td>STM32F4L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM32F7</td>
<td>1082 CoreMark</td>
<td>216 MHz Cortex-M7</td>
</tr>
<tr>
<td>STM32F7L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM32F1</td>
<td>177 CoreMark</td>
<td>72 MHz Cortex-M3</td>
</tr>
<tr>
<td>STM32F1L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM32F0</td>
<td>114 CoreMark</td>
<td>48 MHz Cortex-M0+</td>
</tr>
<tr>
<td>STM32F0L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM32F0G</td>
<td>106 CoreMark</td>
<td>48 MHz Cortex-M0</td>
</tr>
<tr>
<td>STM32F0G0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Don't go alone

We have created a network of companies to support you

Trust our **authorized partners** to ensure the success of your project. Learn more at [st.com/stm32ai](http://st.com/stm32ai)

Wish to discuss a co-development partnership for ML/AI projects? Contact us at [edge.ai@st.com](mailto:edge.ai@st.com)
Releasing your creativity

/STM32

@ST_World

community.st.com

stm32ai.st.com

wiki.st.com/stm32

github.com/STMicroelectronics

Videos

STM32Cube.AI blog articles
Our technology starts with You

Find out more at stm32ai.st.com