STM32Cube.AI: The full power of neural networks available at the edge without compromise

Markus Mayr
Introduction to Edge AI
A new way to add environmental 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

**Input Data** + **Procedural algorithm** ➔ **Desired output from the system**

**Machine Learning**  
Rules learnt from real-world data

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

**Input Data** + **Desired output from the system** ➔ **General ML model trained for the specific problem**
Distributed Artificial Intelligence approach

Leverage billions of devices at the Edge!

**CLOUD**
- Cloud services
- On-premise cloud servers

**EDGE**
- Gateways
- Smart devices
- Ultra-low-power devices and sensors
- Real time
- Local processing

<table>
<thead>
<tr>
<th>Grid</th>
<th>Thousands</th>
<th>Millions</th>
<th>Billions</th>
<th>100 Billions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud services</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>On-premise cloud servers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gateways</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smart devices</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ultra-low-power devices and sensors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Analytics
- Storage
- Compute
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

5
STM32 AI ecosystem

<table>
<thead>
<tr>
<th>Applications</th>
<th>Frameworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicative Examples (Function Packs)</td>
<td>NANOEDGE AI STUDIO</td>
</tr>
<tr>
<td>Pre and post Processing libraries</td>
<td>STM32 Cube.AI</td>
</tr>
<tr>
<td>AI Model converter</td>
<td></td>
</tr>
</tbody>
</table>

Edge Hardware

- STM32 series
- Discovery kit
- STM32 Nucleo board
- Camera add-on
STM32Cube.AI
The STM32CubeMX expansion pack for ML

Power Consumption Calculator
MCU Selector
Pinout Configuration
Clock Tree Initialization
Peripherals Configuration
Middleware Parameters
Code Generation

STM32Cube.AI
STM32Cube.AI
STM32Cube.AI
AI development workflow

1. Data preparation
   - Data acquisition
   - Data processing

2. Model development
   - Model selection and training
   - Model testing

3. Model Implementation
   - Model library creation
   - Model inference
AI development workflow – STM32Cube.AI

1. Data preparation
   - Data acquisition
   - Data processing
   - Data logging tools

2. Model development
   - Model selection and training
   - Model testing
   - Keras
   - TensorFlow Lite
   - ONNX

3. Model Implementation
   - Model library creation
   - Model inference
   - Edge AI toolkit
   - STM32
   - STM32Cube.AI
A tool to seamlessly integrate AI in your projects

Machine Learning
- 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
The 3 pillars of STM32Cube.AI

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

<table>
<thead>
<tr>
<th>HW Target</th>
<th>NUCLEO-STM32H743ZI2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Low complexity handwritten digit reading</td>
</tr>
<tr>
<td>Freq</td>
<td>480 MHz</td>
</tr>
<tr>
<td>Accuracy</td>
<td>&gt;97% for all quantized models</td>
</tr>
<tr>
<td>Tested database</td>
<td>MNIST dataset</td>
</tr>
</tbody>
</table>
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
STM32Cube.AI
Get the best performance on STM32

Image Classif v0.5

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Flash (KB)</th>
<th>RAM (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>61</td>
<td>49</td>
</tr>
<tr>
<td>37</td>
<td>77</td>
<td>53</td>
</tr>
</tbody>
</table>

Visual Wake Word v0.5

<table>
<thead>
<tr>
<th>Latency (ms)</th>
<th>Flash (KB)</th>
<th>RAM (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>42</td>
<td>37</td>
</tr>
<tr>
<td>31</td>
<td>214</td>
<td>98</td>
</tr>
</tbody>
</table>

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
STM32Cube.AI is compatible with all STM32 series

### MPU

**STM32H7**
- Up to 3224 CoreMark
- Up to 550 MHz Cortex-M7
- 240 MHz Cortex-M4

### High Perf MCUs

<table>
<thead>
<tr>
<th></th>
<th>CoreMark</th>
<th>Clock Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STM32F0</strong></td>
<td>106</td>
<td>48 MHz Cortex-M0</td>
</tr>
<tr>
<td><strong>STM32G0</strong></td>
<td>142</td>
<td>64 MHz Cortex-M0+</td>
</tr>
<tr>
<td><strong>STM32F1</strong></td>
<td>177</td>
<td>72 MHz Cortex-M3</td>
</tr>
<tr>
<td><strong>STM32F2</strong></td>
<td>Up to 398</td>
<td>120 MHz Cortex-M3</td>
</tr>
<tr>
<td><strong>STM32F4</strong></td>
<td>Up to 608</td>
<td>180 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32F7</strong></td>
<td>1082</td>
<td>216 MHz Cortex-M7</td>
</tr>
</tbody>
</table>

### Mainstream MCUs

<table>
<thead>
<tr>
<th></th>
<th>CoreMark</th>
<th>Clock Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STM32L0</strong></td>
<td>75</td>
<td>32 MHz Cortex-M0+</td>
</tr>
<tr>
<td><strong>STM32L1</strong></td>
<td>93</td>
<td>32 MHz Cortex-M3</td>
</tr>
<tr>
<td><strong>STM32L4</strong></td>
<td>273</td>
<td>80 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32L4+</strong></td>
<td>409</td>
<td>120 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32L5</strong></td>
<td>443</td>
<td>110 MHz Cortex-M33</td>
</tr>
</tbody>
</table>

### Ultra-low Power MCUs

<table>
<thead>
<tr>
<th></th>
<th>CoreMark</th>
<th>Clock Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STM32L0</strong></td>
<td>75</td>
<td>32 MHz Cortex-M0+</td>
</tr>
<tr>
<td><strong>STM32L1</strong></td>
<td>93</td>
<td>32 MHz Cortex-M3</td>
</tr>
<tr>
<td><strong>STM32L4</strong></td>
<td>273</td>
<td>80 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32L4+</strong></td>
<td>409</td>
<td>120 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32L5</strong></td>
<td>443</td>
<td>110 MHz Cortex-M33</td>
</tr>
</tbody>
</table>

### Wireless MCUs

<table>
<thead>
<tr>
<th></th>
<th>CoreMark</th>
<th>Clock Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STM32WL</strong></td>
<td>162</td>
<td>48 MHz Cortex-M4</td>
</tr>
<tr>
<td><strong>STM32WB</strong></td>
<td>216</td>
<td>64 MHz Cortex-M4</td>
</tr>
</tbody>
</table>

### Latest product generation

- **STM32MP1**
  - 4158 CoreMark
  - Up to 800 MHz Cortex-A7
  - 209 MHz Cortex-M4

- **STM32U5**
  - 651 CoreMark
  - 160 MHz Cortex-M33

### Mixed-signal MCUs

- **STM32F0**
  - 106 CoreMark
  - 48 MHz Cortex-M0

- **STM32G0**
  - 142 CoreMark
  - 64 MHz Cortex-M0+

- **STM32F1**
  - 177 CoreMark
  - 72 MHz Cortex-M3

- **STM32F2**
  - Up to 398 CoreMark
  - 120 MHz Cortex-M3

- **STM32F4**
  - Up to 608 CoreMark
  - 180 MHz Cortex-M4

- **STM32F7**
  - 1082 CoreMark
  - 216 MHz Cortex-M7

- **STM32L0**
  - 75 CoreMark
  - 32 MHz Cortex-M0+
Integrate your ML models more easily with our application-oriented code examples

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

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

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

**Face recognition**
- Face detection and recognition
- Fully functional without cloud connection
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)

Development zone

- Getting started
  Be guided step-by-step to learn STM32 ecosystem
- Development zone
  Get started on application development and project sharing

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
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

www.st.com/STM32ai

wiki.st.com/stm32

github.com/STMicroelectronics

Videos

STM32Cube.AI blog articles
Our technology starts with You

Find out more at www.st.com/stm32ai