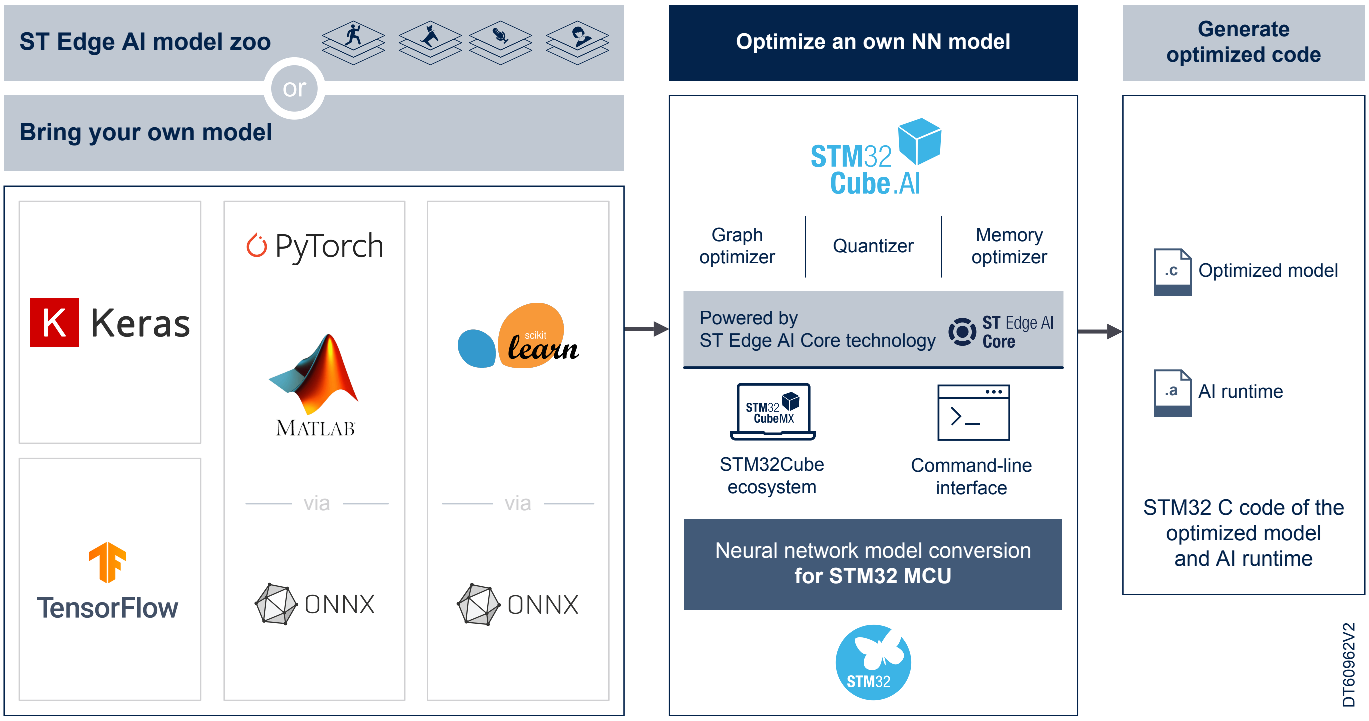
X-CUBE-AI

Data brief

Artificial intelligence (AI) software expansion for STM32CubeMX



Product status link
X-CUBE-AI

Features

- Generation of an STM32-optimized library from pretrained neural network and classical machine learning (ML) models
- Support for STMicroelectronics Neural-ART Accelerator neural processing unit (NPU) for AI/ML model acceleration in hardware
- Native support for various deep learning frameworks such as Keras and TensorFlow™ Lite, and support for all frameworks that can export to the ONNX standard format such as PyTorch™, MATLAB®, and more
- Support for various built-in scikit-learn models such as isolation forest, support vector machine (SVM), and K-means via ONNX
- Support for 32-bit float and 8-bit quantized neural network formats (TensorFlow™ Lite and ONNX Tensor-oriented QDQ)
- Support for deeply quantized neural networks (down to 1-bit) from QKeras and Larq
- Relocatable option enabling standalone model update during the product life cycle by creating a model binary code separated from the application code
- Possibility to use larger networks by storing weights in external flash memory and activation buffers in external RAM
- Easy portability across different STM32 microcontroller series through STM32Cube integration
- With a TensorFlow™ Lite neural network, code generation using either the STM32Cube.AI runtime or TensorFlow™ Lite for Microcontrollers runtime
- Free-of-charge, user-friendly license terms

Description

X-CUBE-AI is an STM32Cube Expansion Package designed to evaluate, optimize, and compile edge AI models for STM32 microcontrollers and the Neural-ART Accelerator. When optimizing NN models for the Neural-ART Accelerator NPU, the tool generates the microcode that maps AI operations to the NPU when possible, falling back to CPU when necessary. This scheduling is performed at the operator level to maximize AI hardware acceleration. X-CUBE-AI is part of the STM32Cube.AI ecosystem and extends STM32CubeMX capabilities by automatically converting pretrained artificial intelligence algorithms into C code. It also integrates a generated optimized library into the user's project.

The easiest way to use X-CUBE-AI is to download it inside the STM32CubeMX tool (version 5.4.0 or newer) as described in the user manual *Getting started with X-CUBE-AI Expansion Package for artificial intelligence (AI)* (UM2526).

The X-CUBE-AI Expansion Package also offers several means to validate artificial intelligence algorithms both on a desktop PC and an STM32. With X-CUBE-AI, it is also possible to measure performance on STM32 devices without any user-specific handmade C code.

ST Edge AI Suite

X-CUBE-AI is part of STMicroelectronics ST Edge AI Suite, which is an integrated collection of software tools designed to facilitate the development and deployment of embedded AI applications. This comprehensive suite supports both optimization and deployment of machine learning algorithms and neural network models, from data collection to the final deployment on hardware, streamlining the workflow for professionals across various disciplines.

The ST Edge AI Suite supports various STMicroelectronics products: STM32 microcontrollers and microprocessors, Neural-ART Accelerator, Stellar microcontrollers, and smart sensors.

The ST Edge AI Suite represents a strategic move to democratize edge AI technology, making it a pivotal resource for developers looking to harness the power of AI in embedded systems efficiently and effectively.

1 General information

1.1 Download information

X-CUBE-AI is available for free download from the *www.st.com* website.

1.2 What is STM32Cube?

STM32Cube is an STMicroelectronics original initiative to improve designer productivity significantly by reducing development effort, time, and cost. STM32Cube covers the whole STM32 portfolio.

STM32Cube includes:

- A set of user-friendly software development tools to cover project development from conception to realization, among which are:
  - STM32CubeMX, a graphical software configuration tool that allows the automatic generation of C initialization code using graphical wizards
  - STM32CubeIDE, an all-in-one development tool with peripheral configuration, code generation, code compilation, and debug features
  - STM32CubeCLT, an all-in-one command-line development toolset with code compilation, board programming, and debug features
  - STM32CubeProgrammer (STM32CubeProg), a programming tool available in graphical and command-line versions
  - STM32CubeMonitor (STM32CubeMonitor, STM32CubeMonPwr, STM32CubeMonRF, STM32CubeMonUCPD), powerful monitoring tools to fine-tune the behavior and performance of STM32 applications in real time
- STM32Cube MCU and MPU Packages, comprehensive embedded-software platforms specific to each microcontroller and microprocessor series (such as STM32CubeN6 for the STM32N6 series), which include:
  - STM32Cube hardware abstraction layer (HAL), ensuring maximized portability across the STM32 portfolio
  - STM32Cube low-layer APIs, ensuring the best performance and footprints with a high degree of user control over hardware
  - A consistent set of middleware components such as ThreadX, FileX, LevelX, NetX Duo, USBX, USB PD, video encoder API, and OpenBL
  - All embedded software utilities with full sets of peripheral and applicative examples
- STM32Cube Expansion Packages, which contain embedded software components that complement the functionalities of the STM32Cube MCU and MPU Packages with:
  - Middleware extensions and applicative layers
  - Examples running on some specific STMicroelectronics development boards

1.3 How does this package complement STM32Cube?

The X-CUBE-AI Expansion Package extends STM32CubeMX by providing an automatic neural network library and classical machine learning library generator optimized in computation and memory (RAM and flash) that converts pretrained artificial intelligence algorithms from most used AI frameworks (such as Keras, TensorFlow™ Lite, scikit-learn, and any model exported in the ONNX format) into a library that is automatically integrated in the final user's project. The project is automatically set up, ready for compilation and execution on the STM32 microcontroller or Neural-ART Accelerator.

X-CUBE-AI also extends STM32CubeMX by adding, for the project creation, specific MCU and board filtering to select the right devices that fit specific criteria requirements (such as RAM or flash memory size) for a user's AI model.

The X-CUBE-AI tool can generate three kinds of projects:

- System performance project running on the STM32 MCU or NPU allowing the accurate measurement of the neural network inference CPU or NPU load and memory usage
- Validation project that validates incrementally the results returned by the neural network, stimulated by either random or user test data, on both desktop PC and STM32 Arm® Cortex®-M-based MCU or STM32 MCU with Neural-ART Accelerator embedded environment
- Application and mixed-precision quantized template project allowing the building of an application including multi-network support

8-bit quantized networks and binarized neural networks reduce the required flash memory size and improve the inference time without significant loss in the network accuracy.

The tool also offers a complete flexibility of the generated code, allowing optimal usage of internal and external memory.

The X-CUBE-AI tool includes a command-line interface for performing all the analysis, generation, validation, and quantization steps.

Note: *Arm is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere.*

2 License

X-CUBE-AI is delivered under the *Mix Ultimate Liberty+OSS+3rd-party V1* software license agreement (SLA0048).

The software components provided in this package come with different license schemes as shown in Table 1.

Table 1. Software component license agreements

Software component	Copyright	License
h5py	Copyright (c) 2008 Andrew Collette and contributors http://h5py.alven.org (see note). All rights reserved. *Note: refer to http://docs.h5py.org/en/stable/licenses.html.*	BSD-3-Clause
Keras	All contributions by François Chollet: Copyright (c) 2015 - 2018, François Chollet. All rights reserved. All contributions by Google: Copyright (c) 2015 - 2018, Google, Inc. All rights reserved. All contributions by Microsoft: Copyright (c) 2017 - 2018, Microsoft, Inc. All rights reserved. All other contributions: Copyright (c) 2015 - 2018, the respective contributors. All rights reserved.	The MIT License
Larq	Copyright © 2020 The Larq Authors	Apache License 2.0
ONNX	Copyright © 2019 ONNX Project Contributors	The MIT License
matplotlib	Copyright (c) 2012-2013 Matplotlib Development Team; All Rights Reserved	Python Software Foundation, Version 2(1)
numpy		BSD-3-Clause
QKeras	Copyright © 2019 The QKeras Authors	Apache License 2.0
scikit-learn	Copyright (c) 2007–2018 The scikit-learn developers. All rights reserved.	BSD-3-Clause
scikit-image	Copyright (C) 2011, the scikit-image team All rights reserved.	
scipy	Copyright © 2003-2013 SciPy Developers. All rights reserved.	
six	Copyright (c) 2010-2018 Benjamin Peterson	The MIT License
tensorflow(2)	Copyright 2018 The TensorFlow Authors. All rights reserved.	Apache License 2.0

Software component	Copyright	License
Theano	Copyright (c) 2008–2017, Theano Development Team All rights reserved. Contains code from NumPy, Copyright (c) 2005-2016, NumPy Developers. All rights reserved. Contains CnMeM under the same license with this copyright: Copyright (c) 2015, NVIDIA CORPORATION. All rights reserved. Contains frozendict code from slezica's python-frozendict (https://github.com/slezica/python-frozendict/blob/master/frozendict/__init__.py), Copyright (c) 2012 Santiago Lezica. All rights reserved.	BSD-3-Clause
typing	Copyright (c) 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014 Python Software Foundation; All Rights Reserved	Python Software Foundation, Version 2
Jinja2	Copyright (c) 2009 by the Jinja Team	BSD-3-Clause
networkx	Copyright (C) 2004-2012, NetworkX Developers Aric Hagberg <hagberg@lanl.gov> Dan Schult <dschult@colgate.edu> Pieter Swart <swart@lanl.gov> All rights reserved.	BSD-3-Clause

1. *Matplotlib only uses BSD-compatible code, and its license is based on the PSF license.*
2. *TensorFlow is a trademark of Google Inc.*

Revision history

Table 2. **Document revision history**

Date	Revision	Changes
17-Dec-2018	1	Initial release.
3-Jan-2019	2	Updated *Description*.
19-Jul-2019	3	Added the support of TensorFlow™ Lite, quantization of Keras networks, and command-line interface.
11-Oct-2019	4	Updated *Features* and *How does this package complement STM32Cube?*: • Added the support of TensorFlow™ Lite quantized networks • Added the use of external memories to support larger networks
18-Dec-2019	5	Added ONNX support: • Updated *Features* and *License* • Updated figures in *Detailed description* and cover page
10-Jun-2020	6	Updated *Features* and *How does this package complement STM32Cube?* for Deep Learning frameworks. Updated *What is STM32Cube?*
5-Mar-2021	7	Updated the entire document for deprecated toolboxes (Caffe, Lasagne, ConvNetJs): figures, *Features*, *Description*, *How does this package complement STM32Cube?* and *License*. Added code generation using the STM32Cube.AI runtime or TensorFlow™ Lite for Microcontrollers runtime for TensorFlow™ Lite Neural Networks in *Features*.
15-Sep-2021	8	Added the support for open-source models from scikit-learn and the generation of classical Machine Learning models in *Features* and *How does this package complement STM32Cube?* Updated the cover image and *Figure 1*.
20-Jul-2022	9	Added the support of mixed-precision quantization and binarized neural network (BNN) for STM32 from QKeras and Larq in *Features*, *License*, and *How does this package complement STM32Cube?* Updated the cover image and removed *Figure 1. X-CUBE-AI overview*.
14-Feb-2023	10	Updated the support for 8-bit quantized neural network format in *Features*. Updated the cover image and *What is STM32Cube?*
05-Dec-2024	11	Added support for STMicroelectronics Neural-ART Accelerator NPU and 32-bit float network format, and described X-CUBE-AI as part of the ST Edge AI Suite: • Updated the title and the cover image • Updated Features and Description • Updated How does this package complement STM32Cube?

IMPORTANT NOTICE – READ CAREFULLY

STMicroelectronics NV and its subsidiaries ("ST") reserve the right to make changes, corrections, enhancements, modifications, and improvements to ST products and/or to this document at any time without notice. Purchasers should obtain the latest relevant information on ST products before placing orders. ST products are sold pursuant to ST's terms and conditions of sale in place at the time of order acknowledgment.

Purchasers are solely responsible for the choice, selection, and use of ST products and ST assumes no liability for application assistance or the design of purchasers' products.

No license, express or implied, to any intellectual property right is granted by ST herein.

Resale of ST products with provisions different from the information set forth herein shall void any warranty granted by ST for such product.

ST and the ST logo are trademarks of ST. For additional information about ST trademarks, refer to www.st.com/trademarks. All other product or service names are the property of their respective owners.

Information in this document supersedes and replaces information previously supplied in any prior versions of this document.