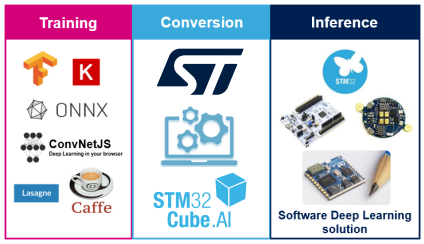
X-CUBE-AI

Data brief

Artificial Intelligence (AI) software expansion for STM32Cube



Features

- Generation of an STM32-optimized library from pre-trained Neural Network models
- Native support of various Deep Learning frameworks such as Keras, TensorFlow™ Lite, Caffe, ConvNetJs and Lasagne, and suppport of all frameworks that can export to the ONNX standard format such as PyTorch™, Microsoft® Cognitive Toolkit, MATLAB® and more
- Supports 8-bit quantization of Keras networks and TensorFlow™ Lite quantized networks
- Allows the use of 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
- Free, user-friendly license terms

Description

X-CUBE-AI is an STM32Cube Expansion Package part of the STM32Cube.AI ecosystem and extending STM32CubeMX capabilities with automatic conversion of pre-trained Neural Network and integration of generated optimized library into the user's project. The easiest way to use it is to download it inside the STM32CubeMX tool (version 5.0.1 or newer) as described in user manual *Getting started with X-CUBE-AI Expansion Package for Artificial Intelligence (AI)* (UM2526).

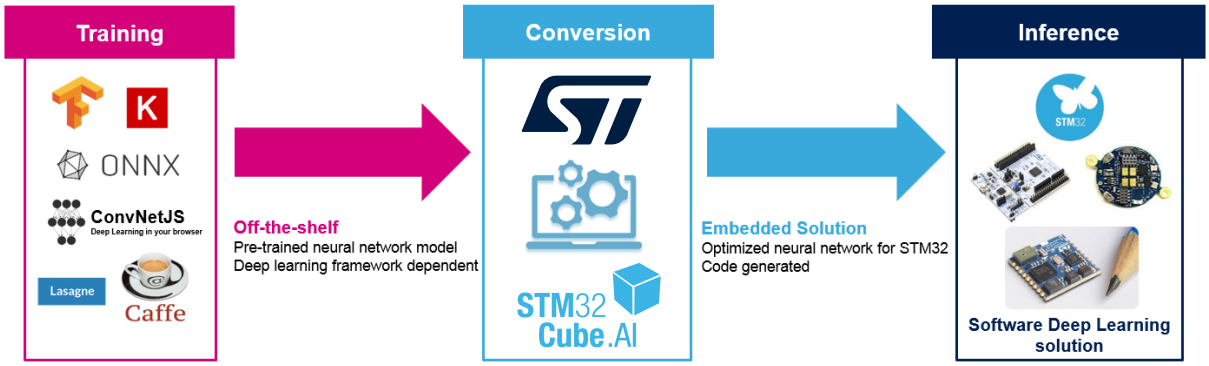
The X-CUBE-AI Expansion Package offers also several means to validate Neural Network models both on desktop PC and STM32, as well as measure performance on STM32 devices without user handmade ad hoc C code.

Product status link
X-CUBE-AI

1 Detailed description

Figure 1 sketches the integration of X-CUBE-AI in STM32 AI environment.

Figure 1. X-CUBE-AI overview



1.1 Ordering 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 significantly improve designer's productivity 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 the conception to the realization, among which:
  - 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
  - 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 & MPU Packages, comprehensive embedded-software platforms specific to each microcontroller and microprocessor series (such as STM32CubeF7 for the STM32F7 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 the HW
  - A consistent set of middleware components such as RTOS, USB, FAT file system, graphics and TCP/IP
  - 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 & 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 generator optimized in computation and memory (RAM and Flash) that converts pre-trained Neural Networks from most used Deep Learning frameworks (such as Keras, TensorFlow™ Lite, Caffe, Lasagne, ConvNetJs and any model exported in the ONNX format) into a library that is automatically integrated in the final user project. The project is automatically setup, ready for compilation and execution on the STM32 microcontroller.

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 size) for a user's Neural Network.

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

- System performance project running on the STM32 MCU allowing the accurate measurement of the Neural Network inference CPU 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 embedded environment
- Application template project allowing the building of an application including multi-network support

8-bit quantized networks reduce the required Flash memory size and improve the inference time without significant loss on 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	Owner	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.	MIT
ONNX	Copyright © 2019 ONNX Project Contributors	MIT
Caffe	Copyright (c) 2014-2017 The Regents of the University of California (Regents) Copyright (c) 2014-2017, the respective contributors	BSD-2-Clause
Lasagne	Copyright (c) 2014-2015 Lasagne contributors	MIT
ConvNetJs	Copyright (c) 2014 Andrej Karpathy	MIT
matplotlib	Copyright (c) 2012-2013 Matplotlib Development Team; All Rights Reserved	Python Software Foundation, Version 2(1)
numpy	Copyright © 2005-2018, NumPy Developers. All rights reserved.	BSD-3-Clause
scikit-learn	Copyright (c) 2007–2018 The scikit-learn developers. All rights reserved.	
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	MIT
tensorflow(2)	Copyright 2018 The TensorFlow Authors. All rights reserved.	Apache License 2.0
Theano	Copyright (c) 2008–2017, Theano Development Team All rights reserved.	BSD-3-Clause

Software component	Owner	License
	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.	
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	Version	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?.

IMPORTANT NOTICE – PLEASE 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 acknowledgement.

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, please 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.