ModusToolbox™ Machine Learning Pack release notes

ModusToolbox™ version 3.8.0 or later

Machine Learning Pack version 3.3.0

About this document

This release is the Machine Learning pack to be installed in addition to the ModusToolbox™ tools package. This document describes the features and known limitations for tools included with this pack.

ModusToolbox™ software is a set of tools that enable you to integrate our devices into your existing development methodology. ModusToolbox™ software consists of various libraries and middleware on GitHub, as well as an IDE and tools package installed on your computer. For more details about what is included with ModusToolbox™ software, refer to the ModusToolbox™ tools package user guide.

Installation

You must have already installed the ModusToolbox™ tools package before installing this pack. If you do not already have the ModusToolbox™ tools package installed, refer to the ModusToolbox™ installation guide for instructions before proceeding.

See additional requirements in the ModusToolbox™ Machine Learning user guide.

What's Included

This pack provides the latest Machine Learning features for use with ModusToolbox™ software, including:

- Machine Learning Configurator version 3.1
- Machine Learning Core Tools version 3.3 (macOS ml-coretools are dedicated for Mac M1/M2/M3 (arm64 architecture) machines only)

What's New/changed

Version 3.3.0 machine learning pack includes the following updates and features:

- Upgraded Vela from 4.5.0 to 5.0.0
- Updated ifx_tfl_pack to 0.17.1
- Exposed VELA options: Python(Cpp) Backend selection and Arena cache size strategy
- Added auto-selection for the interpreter engine for TFLM inference
- Added "sparsity setting ignore" for pre-converted *.tflite* input models
- Fixed the proper batch dimension handling in cycle estimators
- Fixed the double-quantization of already-quantized input data
- Fixed the MAE results file for multi-output models and updated CSV output handling
- Fixed the dequantization pointer arithmetic for the output data type-size handling

- Dropped Mac x86 support
- Updated QEMU images to MTB-ML 3.4

Known Issues

Problem	Workaround
We have discovered an imprecise estimation of model arena size for some LSTM models. Known error: MTB ML streaming init failure: 138280968	Expand the arena size in the model header file: *mtb_ml_gen/<model_name>tflm_model<quantization_type>.h* (Macro name: <model_name>_ARENA_SIZE)
The deployment and validation fail on MacOS Ventura.	The workaround is to install dependencies as following: “brew install cairo, snappy, libjpeg, gnutls, libslirp, colima”
The 1D CSV data in folder format is not supported.	Use the ML CSV or NPZ data formats instead.
The LSTM models in the interpreter-less mode are not supported. The internal state reset is not properly handled and reported arena size is unreasonably large.	Do not use interpreter-less mode for the LSTM models.
The unrolled/unfused LSTM models are not supported in streaming mode.	Use either rolled/fused or non-streaming LSTM models.
The pruned models with enabled sparsity packing feature consume up-to-twice more memory for the arena size.	Ensure that more memory is available or disable the sparsity.
The TFLM warnings may appear in the log: “[WARNING] Found untraced functions such as …”	You can ignore the warnings.
The un-installer for the Machine Learning pack does not un-install on Mac devices.	All Machine Learning pack files need to be manually removed. The Machine Learning pack is located in two locations: /Applications/ModusToolbox-Machine-Learning-Pack, and /Applications/ModusToolbox/packs/ModusToolbox-Machine-Learning-Pack.
When using a TFLM int8x8 model with local data (set REGRESSION_DATA_SOURCE to USE_LOCAL_DATA in main.c) in the Machine Learning Neural Network Profiler code example, the output of the model is compared to quantized reference int8x8 model results.	Use the Machine Learning Neural Network Profiler streaming feature (set REGRESSION_DATA_SOURCE to USE_STREAM_DATA in main.c) to stream data from the ML configurator. This will give the correct accuracy.

NNLite NPU Limitations

TFLite/NNLite Generic Constraints

- Operators supported:
    - CONV1D (as a case of CONV2D)
    - CONV_2D
    - FULLY_CONNECTED
    - AVERAGE_POOL_1D (as a case of AVERAGE_POOL_2D)
    - AVERAGE_POOL_2D
    - MAX_POOL_1D
    - MAX_POOL_2D
    - UNIDIRECTIONAL_LSTM
    - DEPTHWISE_CONV
    - LayerNorm
    - Unfused activations: RELU, RELU6, LEAKY_RELU, RELU_0_TO_1, RELU_N1_TO_1, LOGISTIC, SIGMOID, SOFTMAX, TANH
    - Pointwise operations: ADD, SUB, MUL.

- Input(s) must be of type: int8 or int16
- Weight(s) must be of type: int8
- Input(s) and Output tensors must have a defined shape
- Input(s) and Output tensors must not be greater than 4D
- Input(s), Output and Weight tensors must have quantization parameters
- Output tensors must have quantization scales that fit within float32 precision
- Input Tensor batch size must be 1
- for packed case in weights quantization parameters bits_per_item=8

TFLite CONV1D, CONV_2D Constraints

- Only dilation_h = dilation_w = 1 (no dilation)
- TFLite FULLY_CONNECTED (Dense)
- Only the 2D tensor is supported for the output dims definition
- Fused activations supported: ReLU

TFLite AVERAGE_POOL_2D Constraints

- Padding is not supported (TFLM limitation)

TFLite MAX_POOL_2D Constraints

- Padding is not supported (TFLM limitation)

TFLite DEPTHWISE CONV_2D Constraints

- Padding is not supported (TFLM limitation)

TFLite UNIDIRECTIONAL_LSTM Constraints

- Only the int8x8 quantization is supported (TFLite limitation)

TFLM Pointwise Operations (ADD, SUB, MUL) Constraints

- NNLite can only support broadcast 1 --> for leading major dimensions. It replicates the values of the smaller tensor for the larger tensor.

TFLite RELU Constraints

- run-time switch to HW (NNLite) ReLU: zero_point = -128 only
- run-time switch to SW (TFLM) ReLU: arbitrary zero_point

Note: *To check the NNLite support level for a particular model, run "Desktop Validation" in the ML Configurator. It will produce a software/hardware (SW/HW) direction for every operation inside the model graph.*

```
[INFO] Layer DepthwiseConv with weights of shape: 1x2x3x3 - NNLite v2 - accelerated by NPU
Layer Convolution with filter of shape: 6x1x1x3 - NNLite v2 - accelerated by NPU
Layer DepthwiseConv with weights of shape: 1x2x3x12 - NNLite v2 - accelerated by NPU
Layer FullyConnected with weights of shape: 12x1524 - NNLite v2 - accelerated by NPU
Layer FullyConnected with weights of shape: 5x12 - NNLite v2 - accelerated by NPU
Layer Activation(Softmax) with input flat shape 5  - NNLite v2 - accelerated by NPU
```

Documentation

This pack also includes the following documents located in the installation and on the Infineon website:

- ModusToolbox™ Machine Learning user guide
- ModusToolbox™ Machine Learning Configurator user guide

Trademarks
All referenced product or service names and trademarks are the property of their respective owners.

Edition 2026-07-06

Published by

Infineon Technologies AG

81726 Munich, Germany

© 2026 Infineon Technologies AG. All Rights Reserved.

Do you have a question about this document?

Go to
https://www.infineon.com/support/

Document reference

002-36487 Rev. *G

IMPORTANT NOTICE

The information given in this document shall in no event be regarded as a guarantee of conditions or characteristics (“Beschaffenheitsgarantie”) .

With respect to any examples, hints or any typical values stated herein and/or any information regarding the application of the product, Infineon Technologies hereby disclaims any and all warranties and liabilities of any kind, including without limitation warranties of non-infringement of intellectual property rights of any third party.

In addition, any information given in this document is subject to customer’s compliance with its obligations stated in this document and any applicable legal requirements, norms and standards concerning customer’s products and any use of the product of Infineon Technologies in customer’s applications.

The data contained in this document is exclusively intended for technically trained staff. It is the responsibility of customer’s technical departments to evaluate the suitability of the product for the intended application and the completeness of the product information given in this document with respect to such application.

For further information on the product, technology, delivery terms and conditions and prices please contact your nearest Infineon Technologies office (**www.infineon.com**).

WARNINGS

Due to technical requirements products may contain dangerous substances. For information on the types in question please contact your nearest Infineon Technologies office.

Except as otherwise explicitly approved by Infineon Technologies in a written document signed by authorized representatives of Infineon Technologies, Infineon Technologies’ products may not be used in any applications where a failure of the product or any consequences of the use thereof can reasonably be expected to result in personal injury.