AN14241

如何在MCX N94x上将客户ML模型与NPU无缝集成

第1版—2024年3月1日 **应用笔记**

文档信息

信息	内容
关键词	MCX N94x、NPU、机器学习（ML）
摘要	本文介绍了如何在FRDM-MCXN947板上将客户ML模型与NPU无缝集成。

1 介绍

本文介绍了如何在FRDM-MCXN947板上将客户ML模型与NPU无缝集成。MCX N94x和MCX N54x系列基于运行频率高达150MHz的双高性能Arm Cortex-M33内核，具有2MB的片上闪存和可选的完整ECC RAM，以及集成的专有神经处理单元（NPU）。与单独使用CPU内核相比，集成NPU的机器学习（ML）吞吐量提高了40倍，显著缩短了工作时间并降低了总体功耗。

表1. 要求

软件要求	硬件要求
MCUXpresso IDE eIQ Toolkit MCUXpresso SDK for FRDM-MCXN947	FRDM-MCXN947开发板 硬件调试器：USB-C

2 NPU概述

eIQ Neutron神经处理单元（NPU）是一种高度可扩展的加速器内核架构，可提供机器学习（ML）的加速。此架构提供了功耗和性能优化的NPU，并可与恩智浦大多数微控制器和应用处理器产品组合无缝集成。

eIQ Neutron NPU支持多种神经网络类型，例如CNN、RNN、TCN、Transformer网络等。eIQ机器学习软件开发环境完全支持使用eIQ Neutron NPU进行ML应用开发。MCX N94中所用的NPU是Neutron N1-16。其框图结构如图1所示。

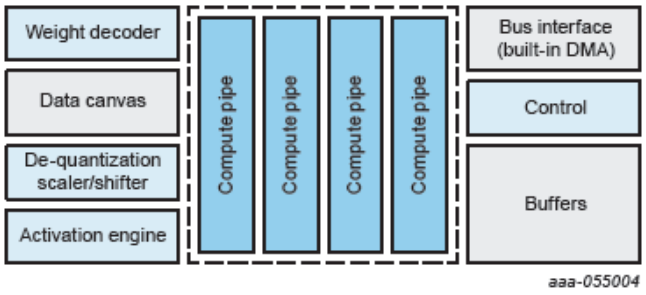

图1. Neutron N1-16结构框图

eIQ Neutron N1-16 NPU内置于MCX N94中，配备了4个计算管道，每个计算管道包含4个INT8 MAC（乘累加）块，共16个MAC块。因此，MCX N94能够以每秒4.8G（150MHz * 4 * 4 * 2）的速度执行INT8运算。

MCUXpresso软件开发套件（MCUXpresso SDK）为开发人员提供了一个综合性软件包，其中包含预集成的TensorFlow Lite for Microcontrollers (TFLM)。

Neutron库也集成到了TFLM中。表2列出了NPU所支持的运算符。

表2. NPU所支持的运算符

运算符	运算符输入类型	MCXN947/MCXN548 NPU
ADD	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有

表2. NPU所支持的运算符（续）

运算符	运算符输入类型	MCXN947/MCXN548 NPU
AVERAGE_POOL_2D	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
CONV_2D	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
DEPTHWISE_CONV_2D	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
FULLY_CONNECTED	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
UNIDIRECTIONAL_SEQUENCE_LSTM	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	无
LOGISTIC (Sigmoid)	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
MAX_POOL_2D	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	有
MUL	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	无
SOFTMAX	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	无
SVDF	Float	无
	Uint8(PTQ)	无
	Int8(PCQ)	无

注：

- *PTQ — 按张量量化（非对称8位量化）。*
- *PCQ — 按通道量化（对称8位量化）。*

有关更多信息，请参阅SDK的`middleware/eiq/doc`中的每一个TensorFlow Lite用户指南。

3 软件环境的搭建

软件环境的搭建是基于Windows的。

1. 下载MCUXpresso IDE，并选择要下载的Windows版本，如图2所示。

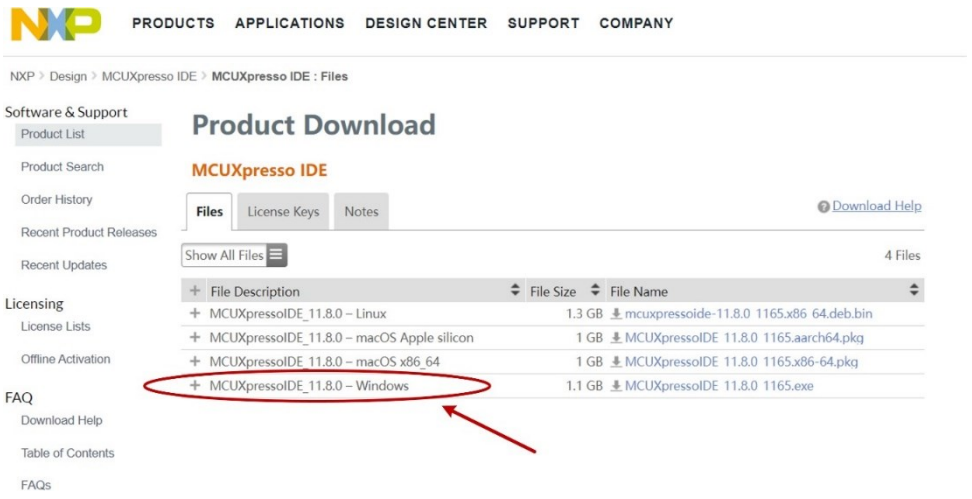

图2. MCUXpresso IDE的下载

2. 下载完成后，双击下载的安装包，选择安装位置，并执行安装直至安装完成。

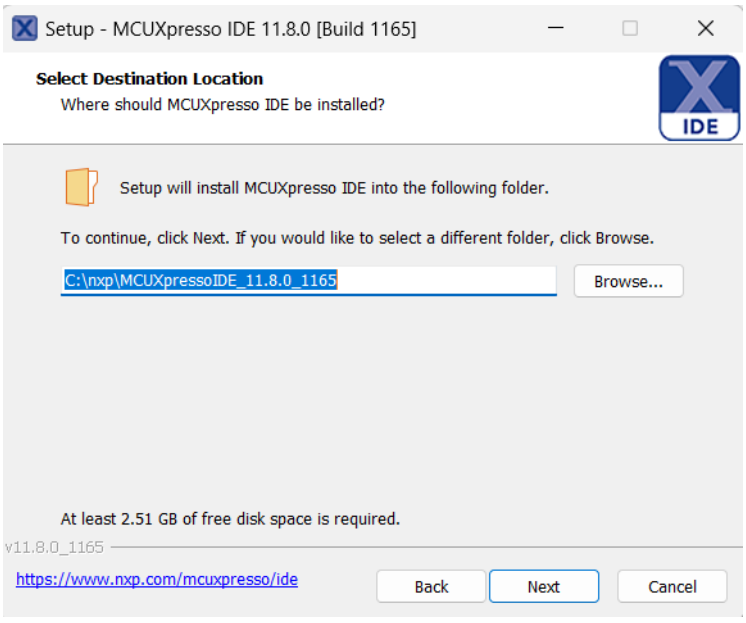

图3. MCUXpresso IDE的安装

***注：**使用USB-C数据线为FRDM-MCXN947板供电和调试，如图4所示。USB-C数据线的另一端必须插入开发计算机中。*

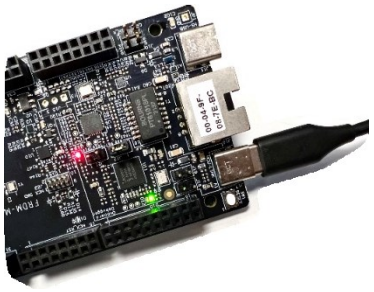

图4. USB-C数据线和FRDM-MCXN947开发板

3. 要安装所需的SDK，请在MCUxpresso SDK Builder中选择FRDM-MCXN947，如图5所示。

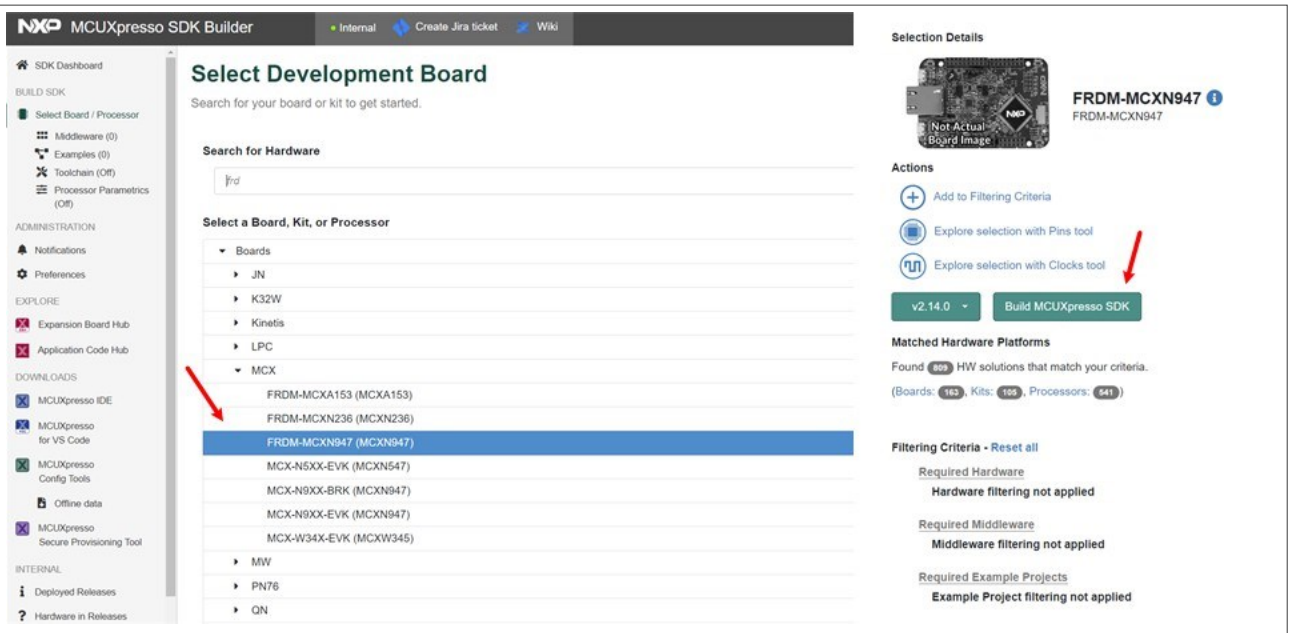

图5. 选择开发板

4. 选择eIQ中间件，点击**DOWNLOAD SDK**，如图6所示。

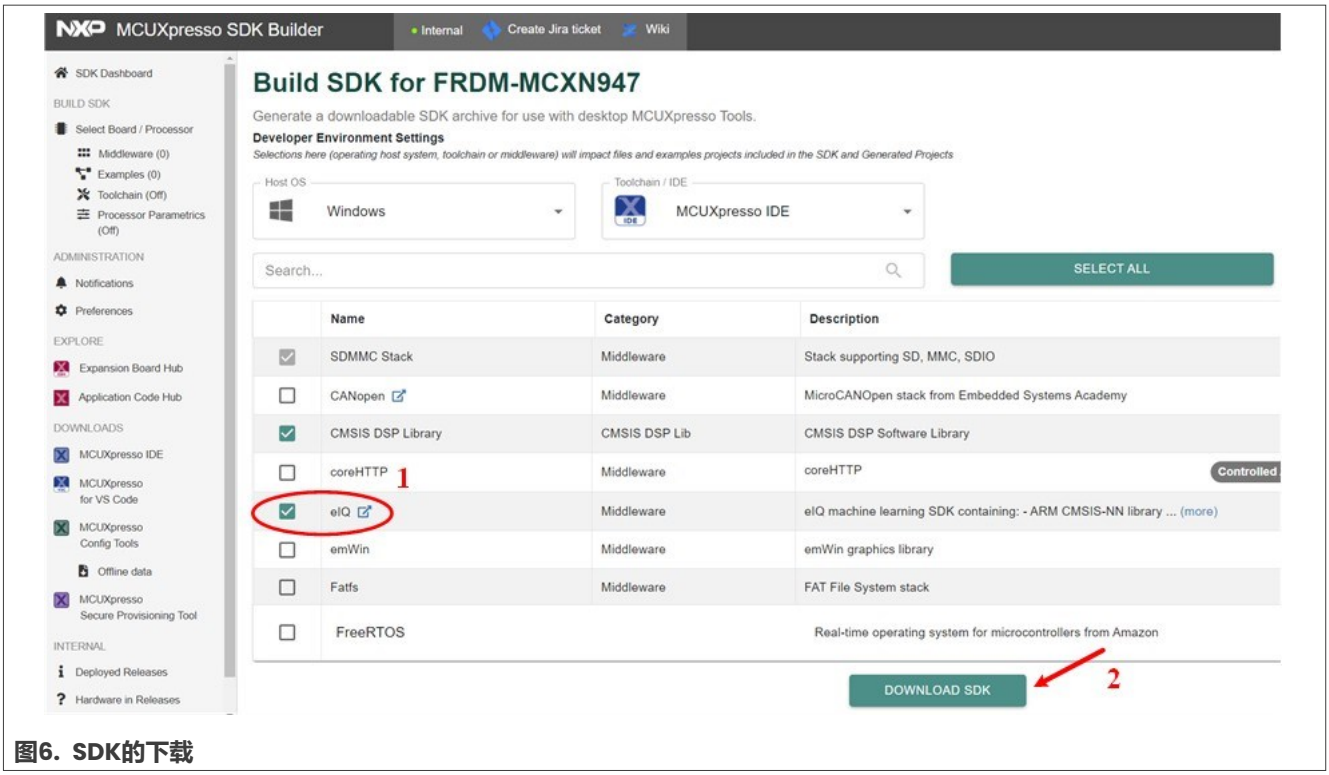

图6. SDK的下载

5. SDK构建完成后，下载SDK压缩文件，如图7所示。

图7. 下载SDK压缩文件

6. 打开MCUXpresso IDE，将SDK文件拖放到MCUXpresso IDE的SDK安装窗口中进行安装，如图8所示。

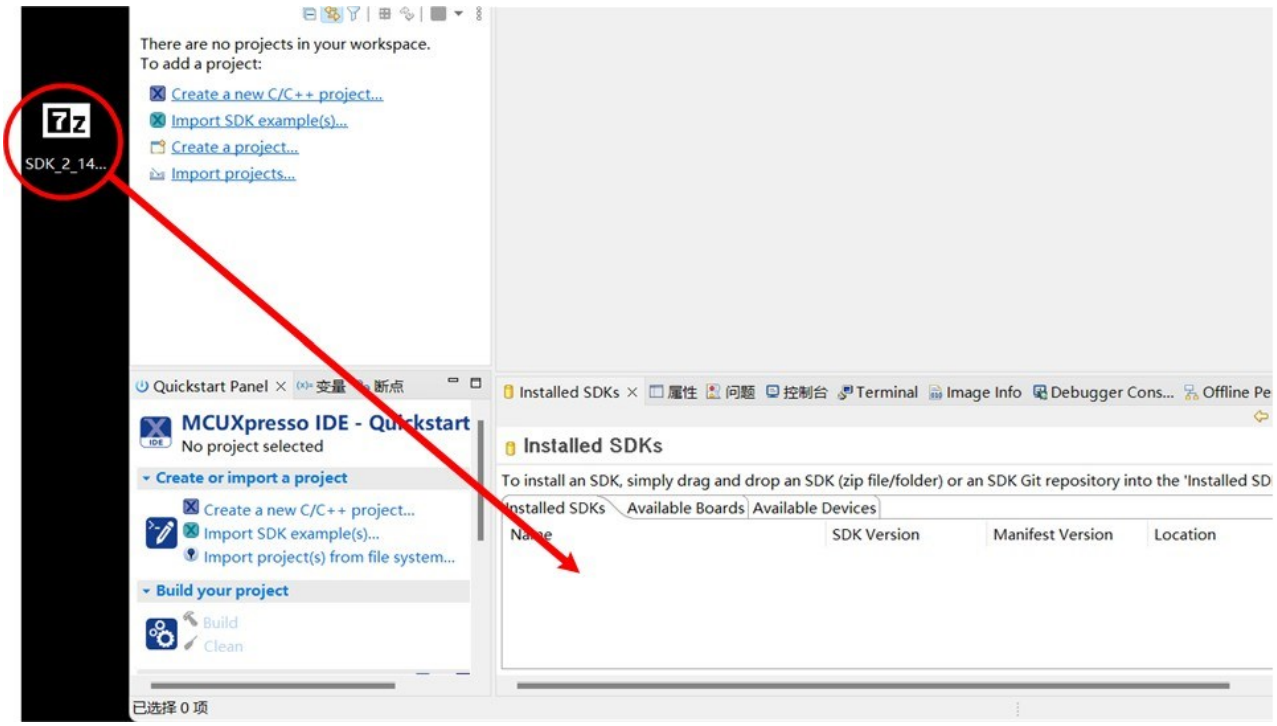

图8. SDK的安装

4 转换客户模型

要使用NPU加速模型，必须首先使用eIQ工具包中所含的Neutron Converter工具将模型进行转换。

在EIQ-TOOLKIT中下载eIQ Toolkit安装包，双击下载的安装包，选择安装位置，然后执行安装直至安装完成，如图9所示。

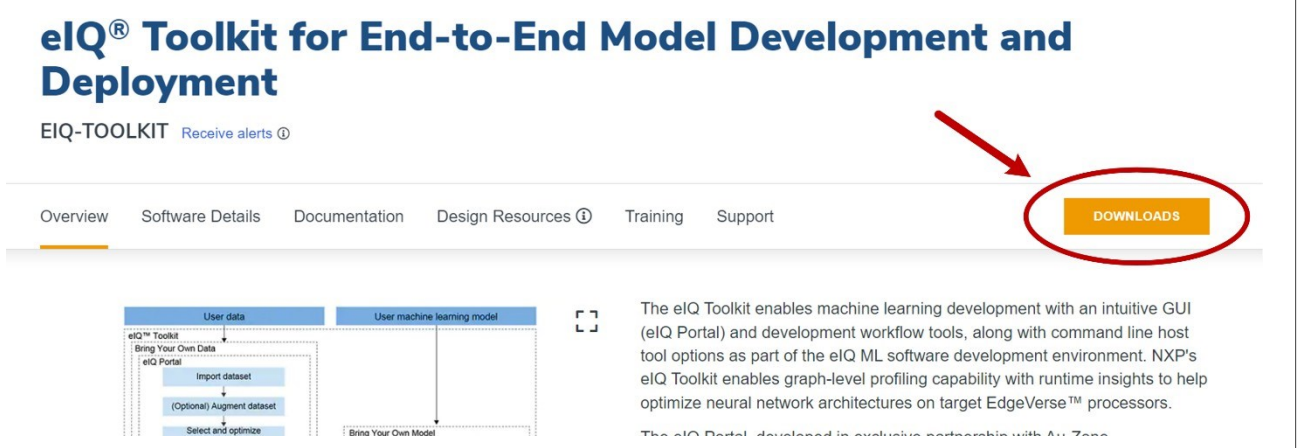

图9. eIQ Toolkit安装包

转换客户模型有两种方法：

- 使用命令行工具：neutron-converter
- 使用eIQ门户应用程序

4.1 命令行工具：neutron-converter

1. 打开eIQ工具包文件夹eIQ_Toolkit_v2.1.0\bin\neutron-converter。此文件夹有两个版本的转换器。根据SDK版本来选择匹配的转换器。对应关系如表3所示。

表3. SDK和转换器

SDK	转换器
SDK-2.13.1	v1.0.0
SDK-2.14.0	v1.2.0

2. 使用命令行输入转换器的文件夹，将客户模型复制到该文件夹中，然后执行命令以转换模型，如图10所示。

图10. 执行转换命令

4.2 eIQ门户应用程序

1. 打开eIQ门户应用程序，并单击顶部的**PLUG-INS**按钮，如图11所示。

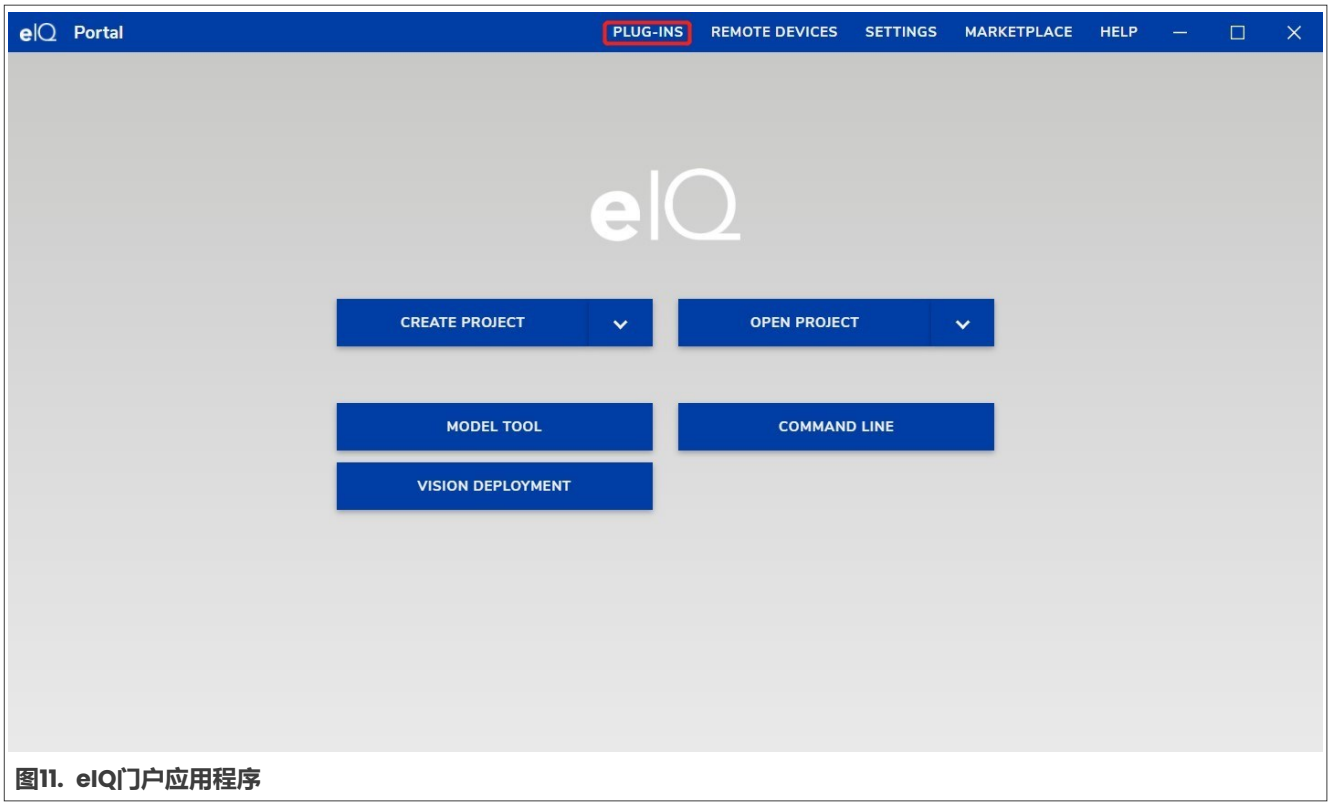

图11. eIQ门户应用程序

2. 在转换服务器初始化成功后，关闭此窗口并在MODEL　TOOL中打开客户模型。神经网络图将显示在窗口中，如图12所示。

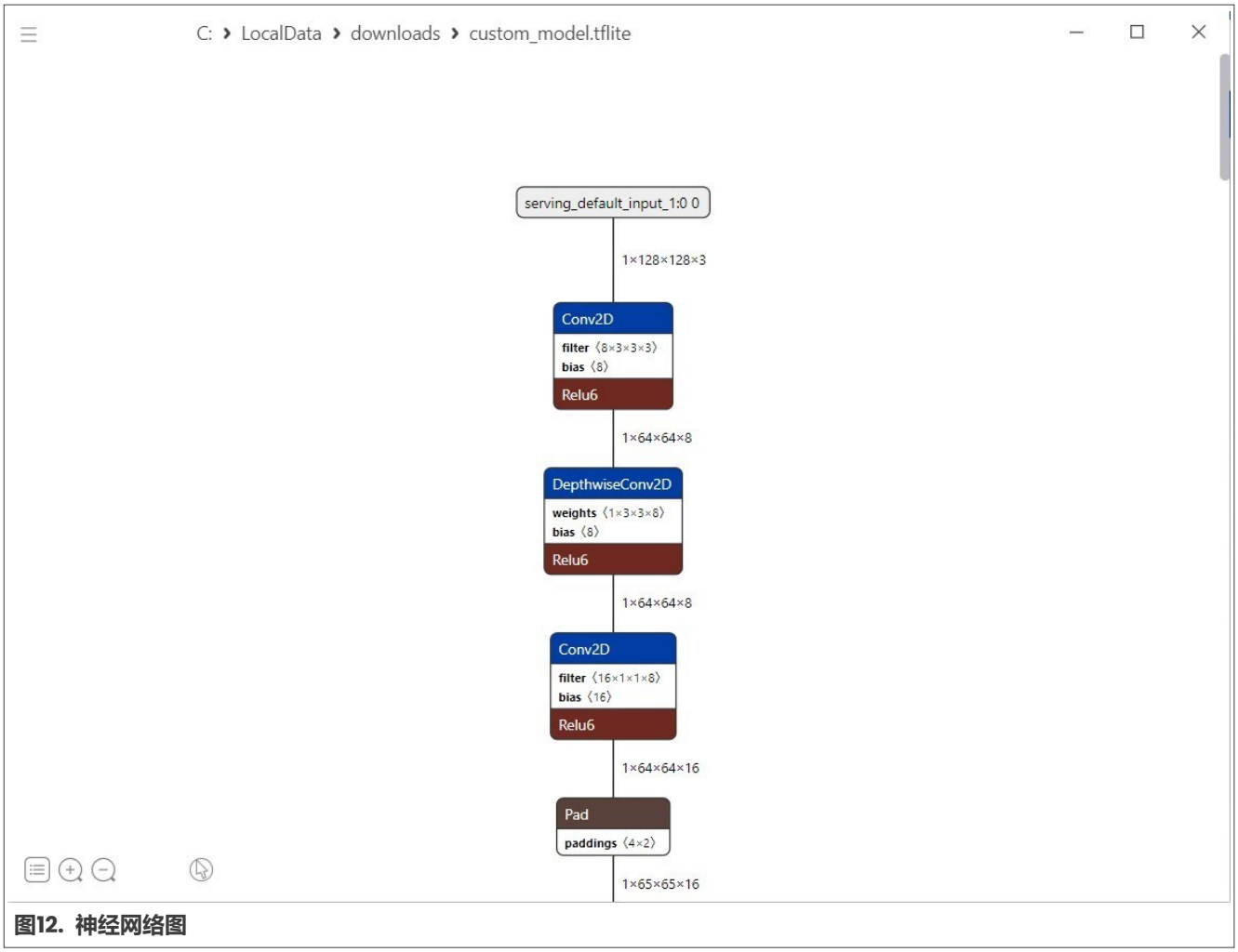

图12. 神经网络图

3. 点击左上角的菜单，在转换菜单中选择TensorFlow Lite For Neutron，如图13所示。

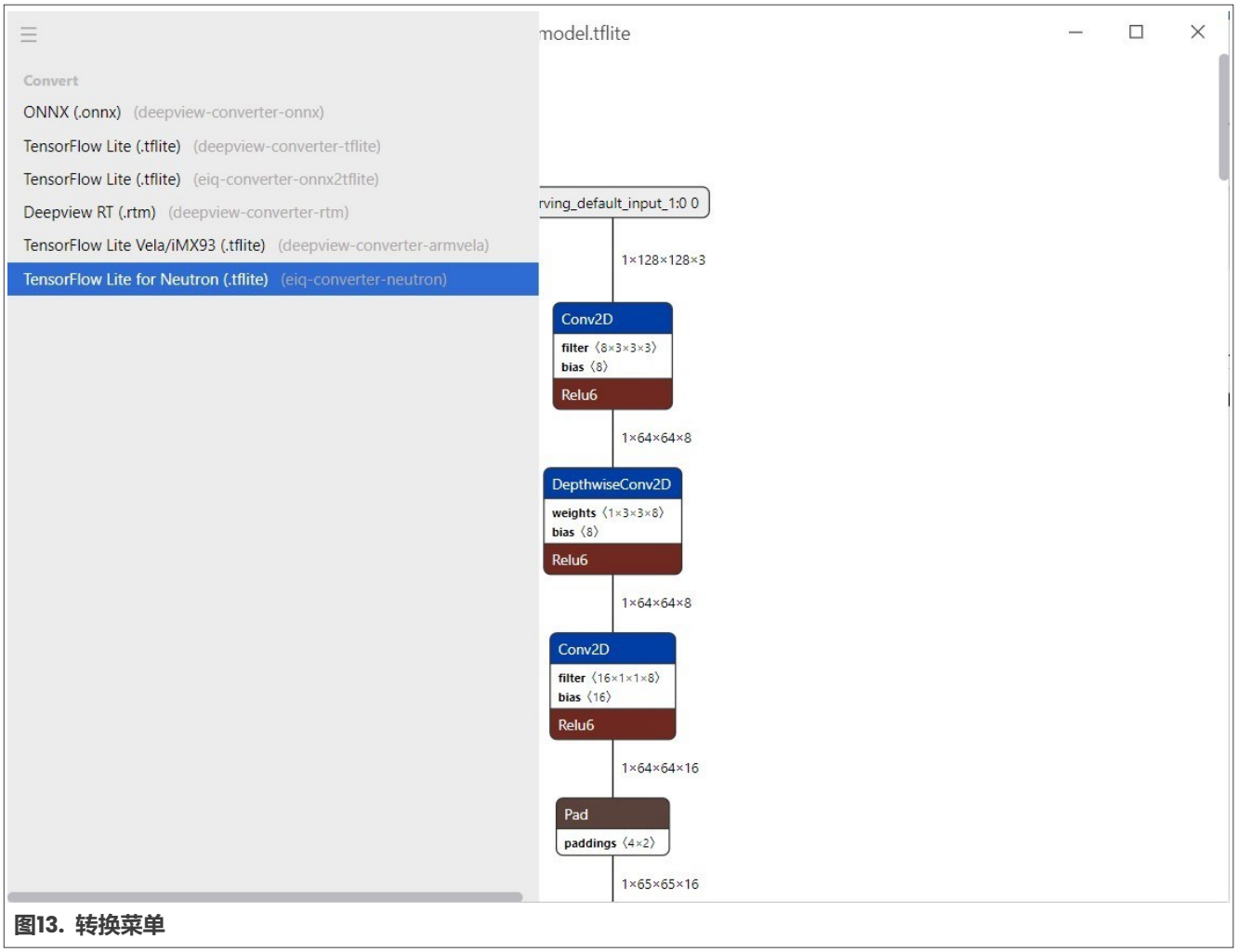

图13. 转换菜单

4. 单击**Convert**按钮，并为转换后的模型选择保存路径。当转换成功后，转换后的图形会显示在窗口中，如图14所示。

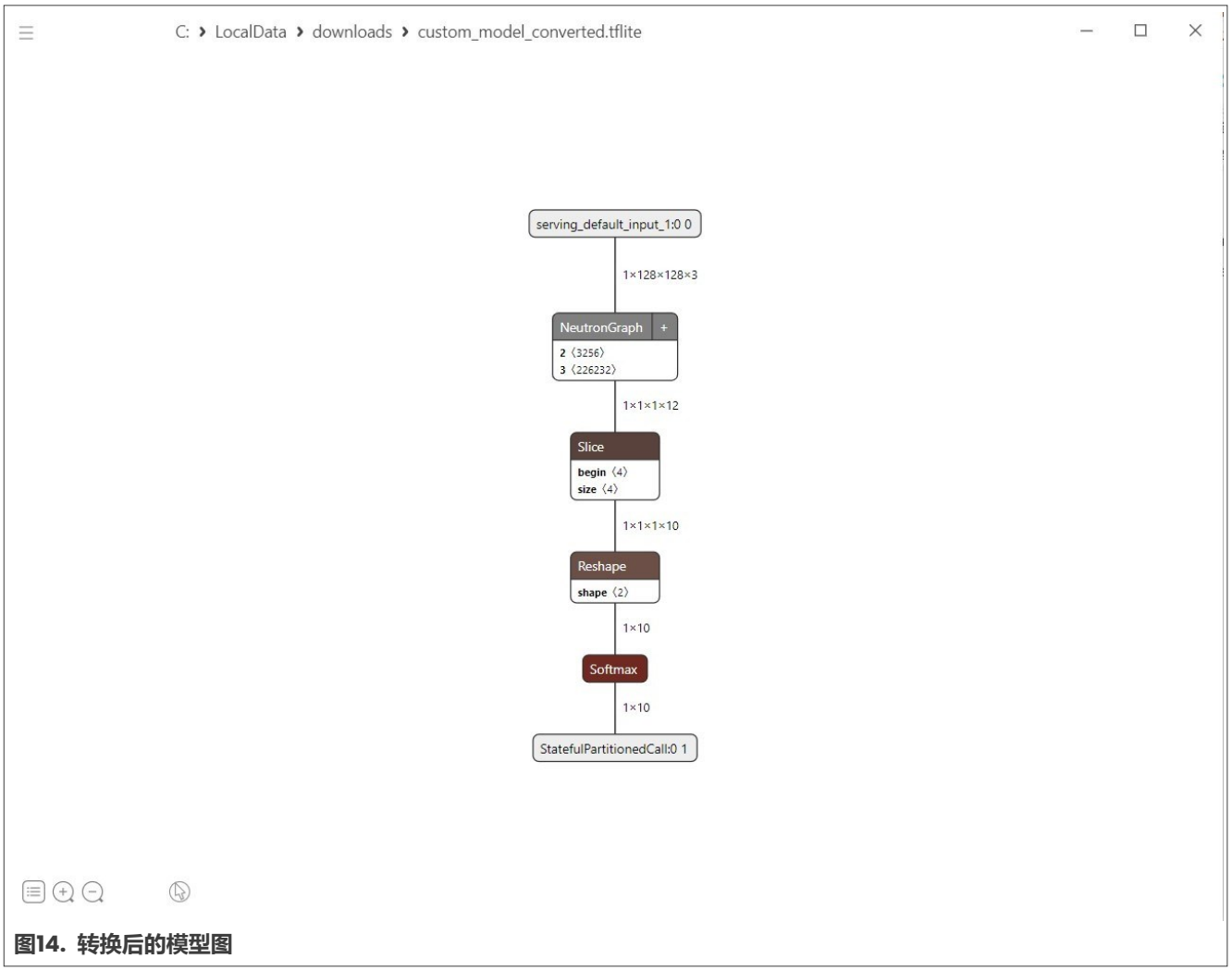

图14. 转换后的模型图

5. 如图14所示，Neutron Converter工具会改变模型的图形，将可在NPU上运行的运算符转换为自定义的NeutronGraph节点。转换后的模型命名为`custom_model_converted.tflite`。
6. 转换后，运算符的数量会减少，量化因子的表达也更有效。模型转换前后的对比如表4所示。

表4. 转换前后的模型

	原始模型	转换后
文件大小	299KB	227KB
张量内存库（Tensor Arena）	157KB	144KB

5 将模型集成到示例工程中

1. 导入`tflm_ciar10`工程。要导入`tflm_ciar10`工程，请按照图15和图16中的步骤操作。

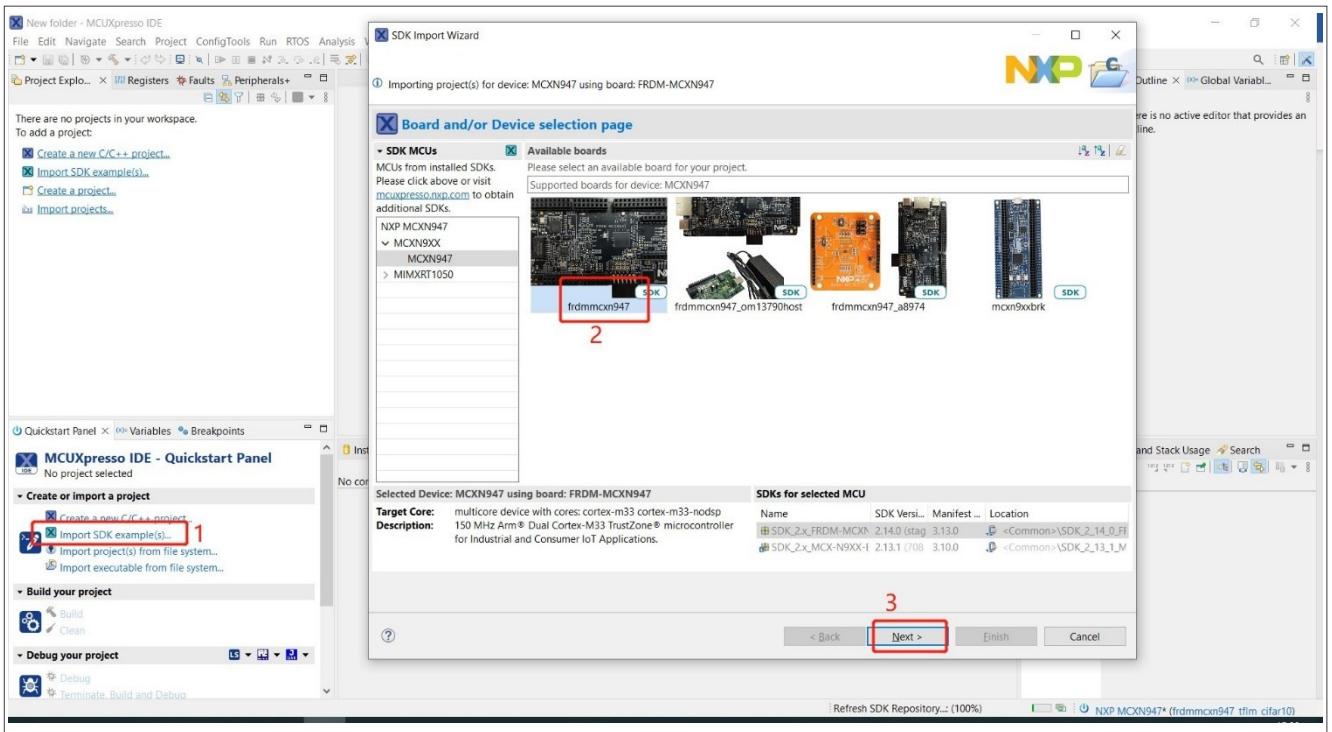

图15. 导入tflm_ciar10工程

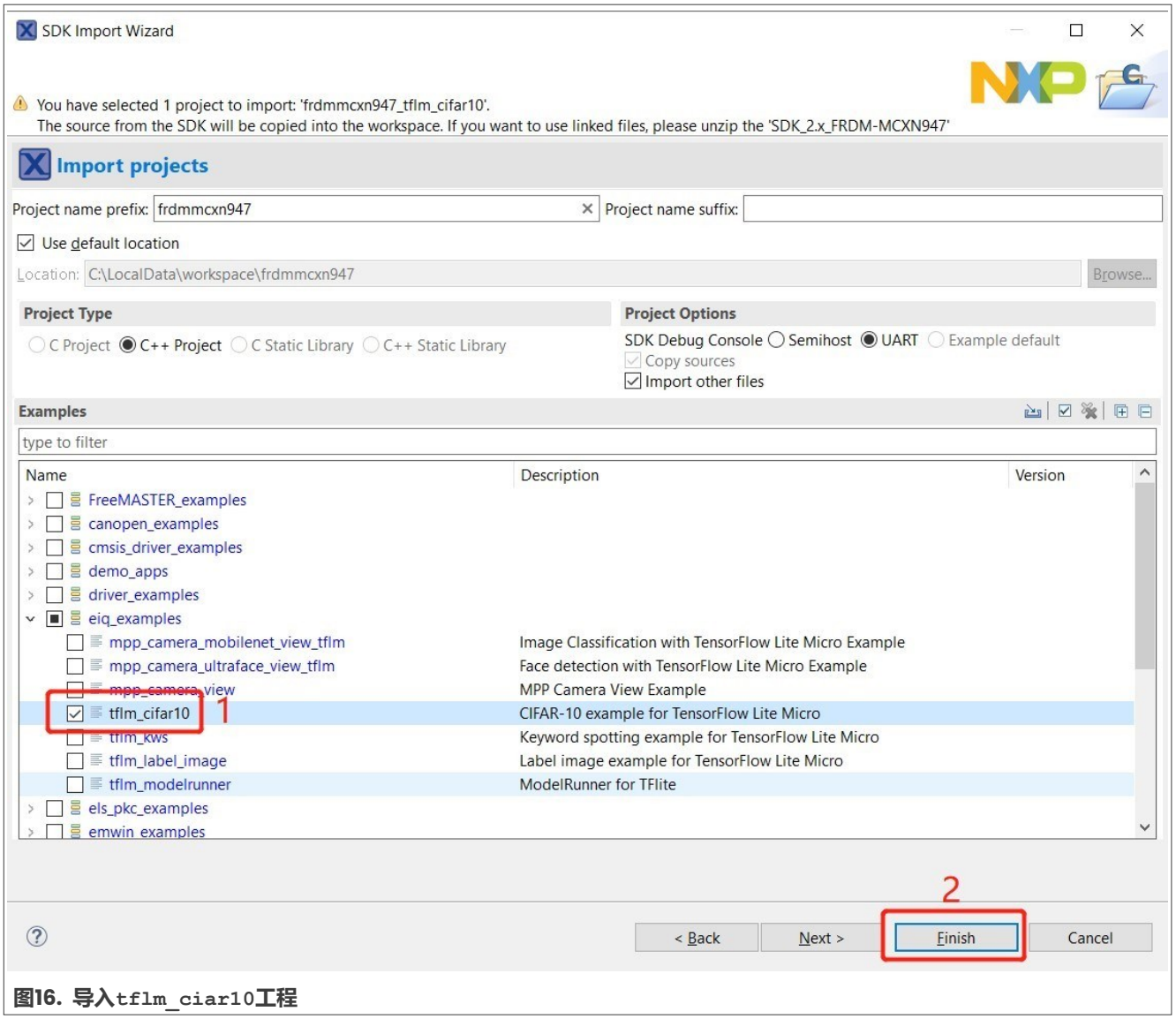

图16. 导入tflm_ciar10工程

2. 将转换后的模型文件复制到工程文件夹中，创建名为model_data.s的程序集文件，并将模型文件包含到数据部分custom_model_data中。导出第32行至第35行的指令，如图17所示。

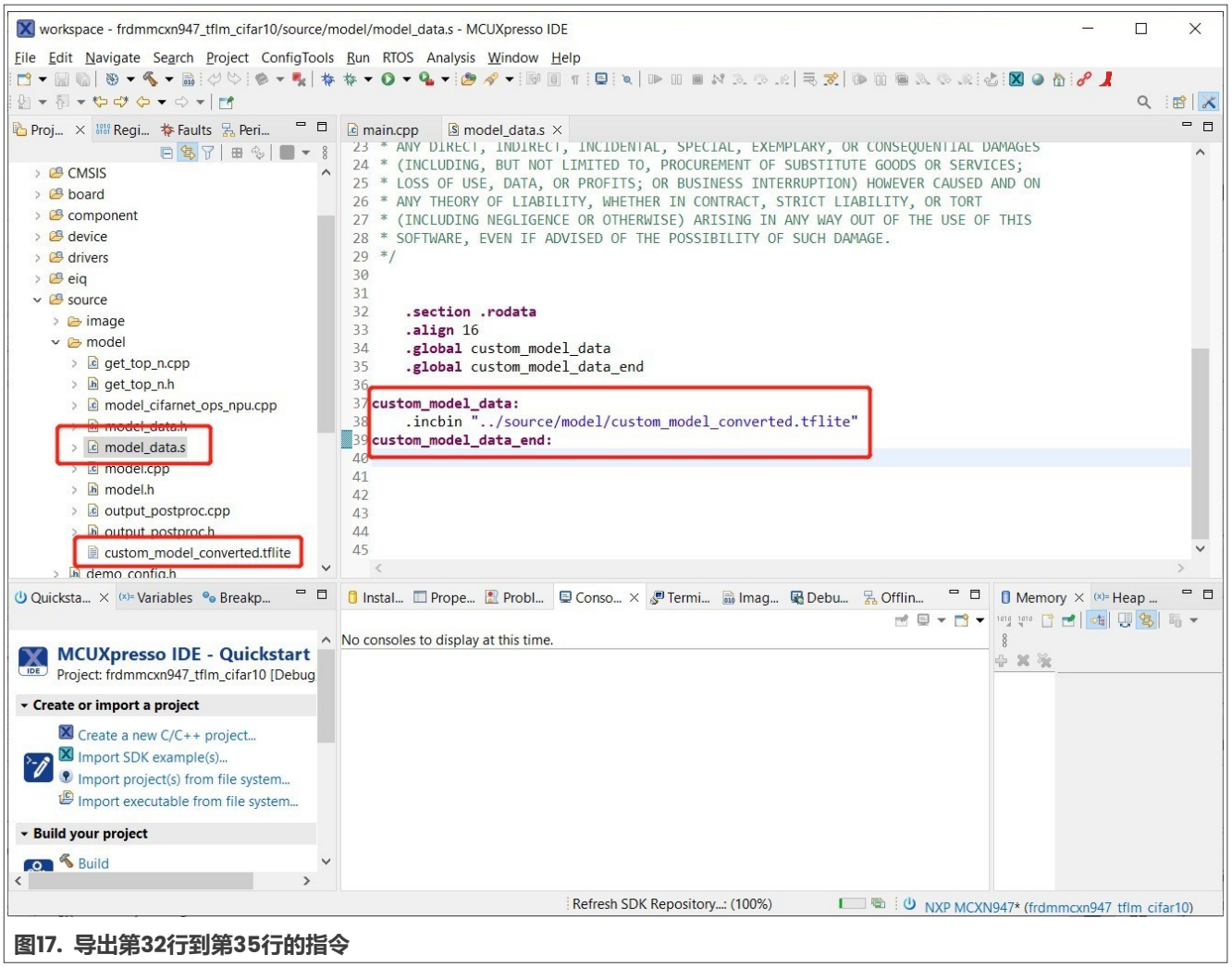

图17. 导出第32行到第35行的指令

3. 修改`model.cpp`，改为使用客户模型数据，如图18所示。

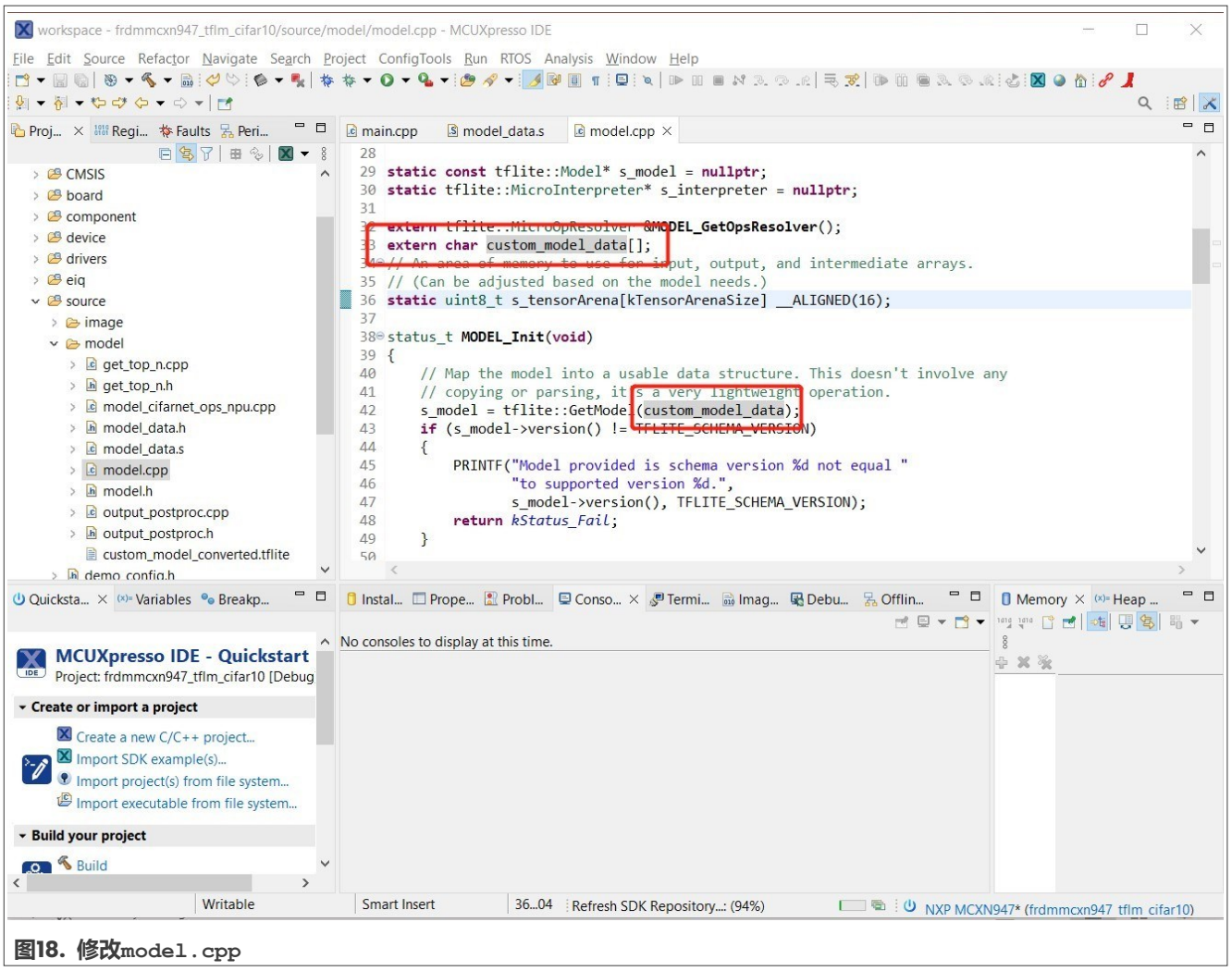

图18. 修改model.cpp

4. 修改model_cifarnet_ops_npu.cpp，确保客户模型所用的所有运算符都添加到s_microOpResolver中，如图19所示。

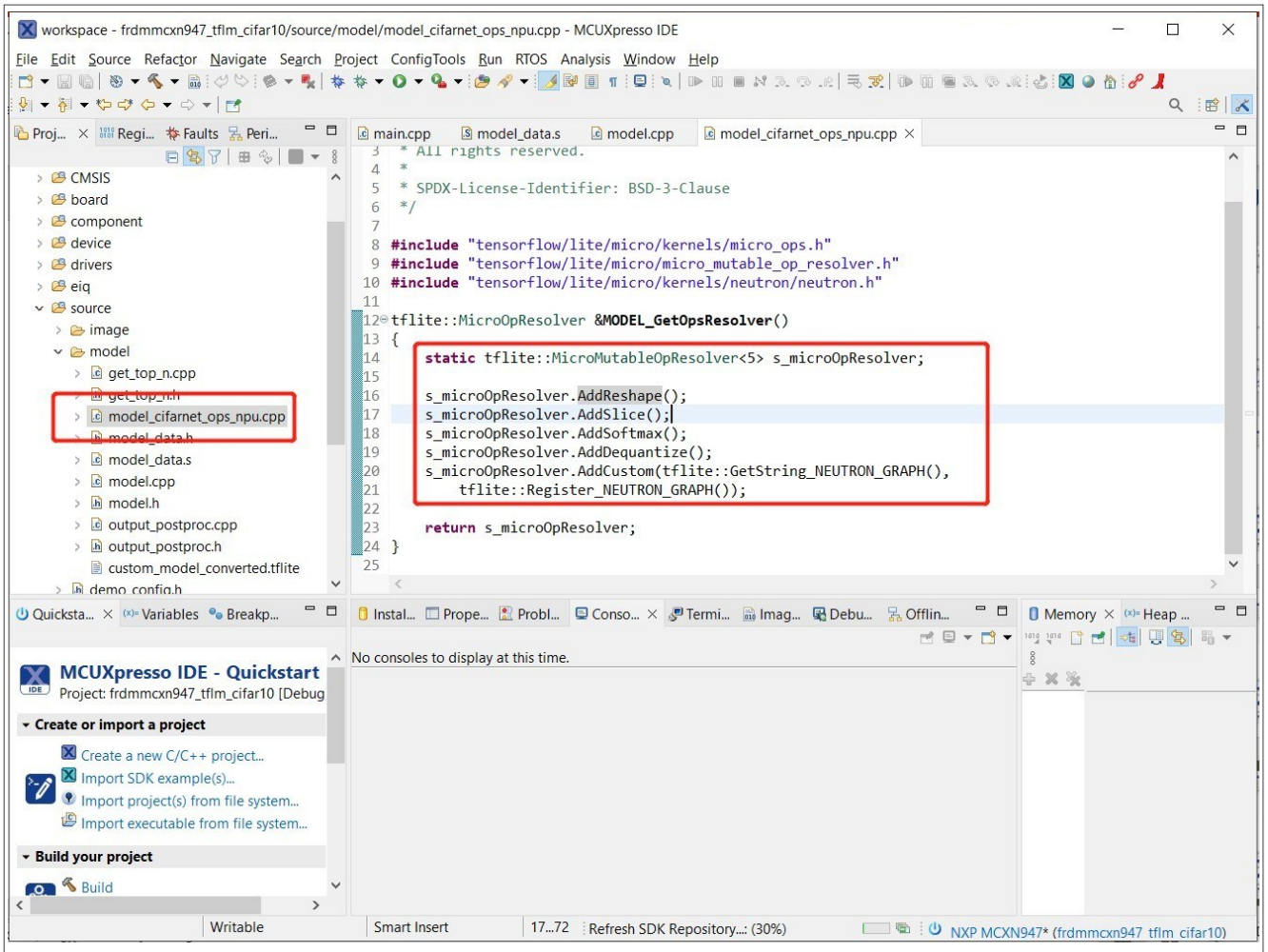

图19. 修改model_cifarnet_ops_npu.cpp

5. 准备一些测试数据，并使用如下Python脚本程序来加载选定的JPG文件（包含在CIFAR10测试数据集中），将其解压并与数组定义一起导出为一个C语言数组。然后，将这个数组保存到与Python脚本文件处于同一文件夹的bird.h文件中。确保图像的路径相对于执行脚本的位置是正确的，如图20所示。

```
import cv2
import numpy as np
img = cv2.imread('test/bird/batch_5_num_113.jpg')
img = cv2.resize(img, (128, 128))
img = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
with open('bird.h', 'w') as fout:
    print('#define STATIC_IMAGE_NAME "bird"', file=fout)
    print('static const uint8_t bird [] = {', file=fout)
    img.tofile(fout, ', ', '0x%02X')
    print('};\n', file=fout)
```

图20. 执行脚本

6. 将头文件（bird.h）复制到工程的图像文件夹中，并修改image_loade.c。切换到使用Bird类图像数据，如图21所示。

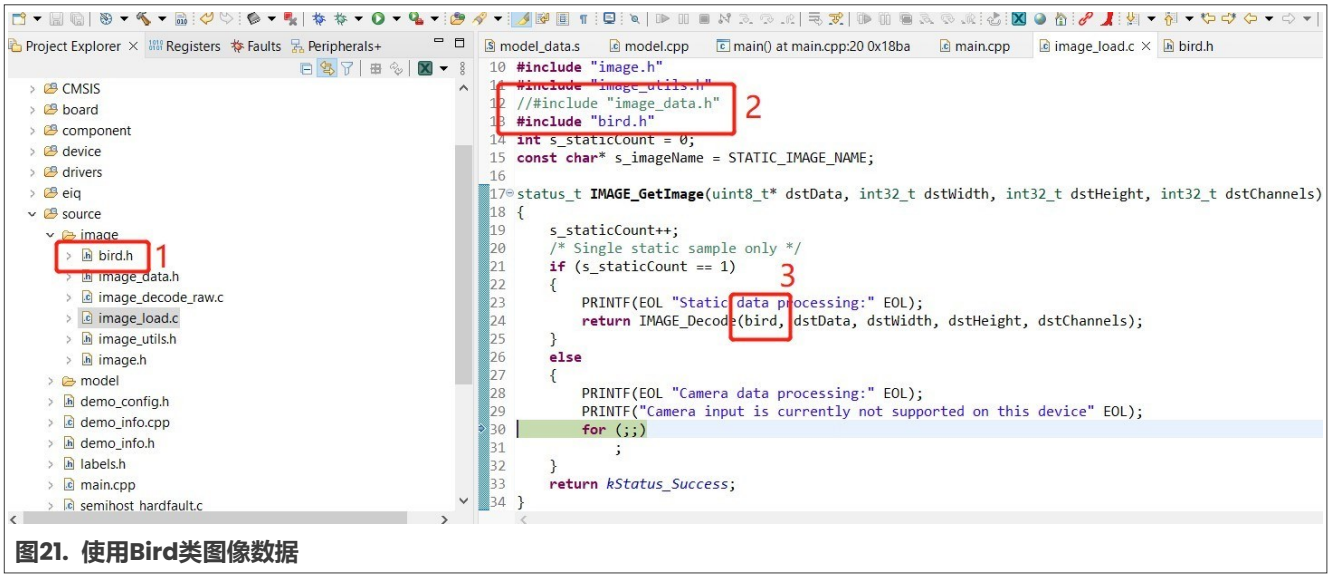

图21. 使用Bird类图像数据

7. 编译工程并将其下载到FRDM-MCXN947开发板。

6 结果

在开发板上运行程序的结果如图22所示，推理时间约为9毫秒，结果得分（置信度）为98%。

图22. 在开发板上运行程序的结果

7 修订历史

表5. 修订历史

文档ID	发布日期	说明
AN14241 v.1	2024年3月1日	初版发布

Legal information

Definitions

Draft — A draft status on a document indicates that the content is still under internal review and subject to formal approval, which may result in modifications or additions. NXP Semiconductors does not give any representations or warranties as to the accuracy or completeness of information included in a draft version of a document and shall have no liability for the consequences of use of such information.

Disclaimers

Limited warranty and liability — Information in this document is believed to be accurate and reliable. However, NXP Semiconductors does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information and shall have no liability for the consequences of use of such information. NXP Semiconductors takes no responsibility for the content in this document if provided by an information source outside of NXP Semiconductors.

In no event shall NXP Semiconductors be liable for any indirect, incidental, punitive, special or consequential damages (including - without limitation - lost profits, lost savings, business interruption, costs related to the removal or replacement of any products or rework charges) whether or not such damages are based on tort (including negligence), warranty, breach of contract or any other legal theory.

Notwithstanding any damages that customer might incur for any reason whatsoever, NXP Semiconductors' aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms and conditions of commercial sale of NXP Semiconductors.

Right to make changes — NXP Semiconductors reserves the right to make changes to information published in this document, including without limitation specifications and product descriptions, at any time and without notice. This document supersedes and replaces all information supplied prior to the publication hereof.

Suitability for use — NXP Semiconductors products are not designed, authorized or warranted to be suitable for use in life support, life-critical or safety-critical systems or equipment, nor in applications where failure or malfunction of an NXP Semiconductors product can reasonably be expected to result in personal injury, death or severe property or environmental damage. NXP Semiconductors and its suppliers accept no liability for inclusion and/or use of NXP Semiconductors products in such equipment or applications and therefore such inclusion and/or use is at the customer's own risk.

Applications — Applications that are described herein for any of these products are for illustrative purposes only. NXP Semiconductors makes no representation or warranty that such applications will be suitable for the specified use without further testing or modification.

Customers are responsible for the design and operation of their applications and products using NXP Semiconductors products, and NXP Semiconductors accepts no liability for any assistance with applications or customer product design. It is customer's sole responsibility to determine whether the NXP Semiconductors product is suitable and fit for the customer's applications and products planned, as well as for the planned application and use of customer's third party customer(s). Customers should provide appropriate design and operating safeguards to minimize the risks associated with their applications and products.

NXP Semiconductors does not accept any liability related to any default, damage, costs or problem which is based on any weakness or default in the customer's applications or products, or the application or use by customer's third party customer(s). Customer is responsible for doing all necessary testing for the customer's applications and products using NXP Semiconductors products in order to avoid a default of the applications and the products or of the application or use by customer's third party customer(s). NXP does not accept any liability in this respect.

Terms and conditions of commercial sale — NXP Semiconductors products are sold subject to the general terms and conditions of commercial sale, as published at https://www.nxp.com.cn/profile/terms, unless otherwise agreed in a valid written individual agreement. In case an individual agreement is concluded only the terms and conditions of the respective agreement shall apply. NXP Semiconductors hereby expressly objects to applying the customer's general terms and conditions with regard to the purchase of NXP Semiconductors products by customer.

Export control — This document as well as the item(s) described herein may be subject to export control regulations. Export might require a prior authorization from competent authorities.

Suitability for use in non-automotive qualified products — Unless this document expressly states that this specific NXP Semiconductors product is automotive qualified, the product is not suitable for automotive use. It is neither qualified nor tested in accordance with automotive testing or application requirements. NXP Semiconductors accepts no liability for inclusion and/or use of non-automotive qualified products in automotive equipment or applications.

In the event that customer uses the product for design-in and use in automotive applications to automotive specifications and standards, customer (a) shall use the product without NXP Semiconductors' warranty of the product for such automotive applications, use and specifications, and (b) whenever customer uses the product for automotive applications beyond NXP Semiconductors' specifications such use shall be solely at customer's own risk, and (c) customer fully indemnifies NXP Semiconductors for any liability, damages or failed product claims resulting from customer design and use of the product for automotive applications beyond NXP Semiconductors' standard warranty and NXP Semiconductors' product specifications.

Translations — A non-English (translated) version of a document, including the legal information in that document, is for reference only. The English version shall prevail in case of any discrepancy between the translated and English versions.

Security — Customer understands that all NXP products may be subject to unidentified vulnerabilities or may support established security standards or specifications with known limitations. Customer is responsible for the design and operation of its applications and products throughout their lifecycles to reduce the effect of these vulnerabilities on customer's applications and products. Customer's responsibility also extends to other open and/or proprietary technologies supported by NXP products for use in customer's applications. NXP accepts no liability for any vulnerability. Customer should regularly check security updates from NXP and follow up appropriately.

Customer shall select products with security features that best meet rules, regulations, and standards of the intended application and make the ultimate design decisions regarding its products and is solely responsible for compliance with all legal, regulatory, and security related requirements concerning its products, regardless of any information or support that may be provided by NXP.

NXP has a Product Security Incident Response Team (PSIRT) (reachable at PSIRT@nxp.com) that manages the investigation, reporting, and solution release to security vulnerabilities of NXP products.

NXP B.V. — NXP B.V. is not an operating company and it does not distribute or sell products.

Trademarks

Notice: All referenced brands, product names, service names, and trademarks are the property of their respective owners.

NXP — wordmark and logo are trademarks of NXP B.V.

目录

For more information, please visit: https://www.nxp.com.cn

Date of release: 1 March 2024
Document identifier: AN14241