AN13924

在i.MX RT1060和RT1170上使用高效神经网络进行多人员检测

第0版—2023年5月8日 **应用笔记**

文档信息

信息	内容
关键词	边缘计算平台、RT1060、RT1170、eIQ、机器学习、uVITA、COCO、Pascal-VOC、人员检测
摘要	恩智浦的跨界MCU是一个理想的边缘计算平台，能提供卓越的算力。

1 介绍

恩智浦的跨界MCU是一个理想的边缘计算平台，能提供卓越的算力。为了进一步展示i.MX RT系列MCU在机器学习方面的能力，本应用笔记介绍了一个在i.MX RT1060和i.MX RT1170上使用高效神经网络进行多人员检测的示例。

1. 此示例提供了一种采用高效网络架构ShuffleNet-V2[1]的轻量级人员检测模型，与Arm平台上的大多数现有网络相比，此架构的速度更快，同时还能降低内存访问的成本。
2. 所给的模型通过eIQ Glow工具转换为目标文件后，对于i.MX RT1060和i.MX RT1170的Arm Cortex-M7内核来说，可以得到更高的性能并减少内存的占用。通过进一步的实验分析，展示了在不同量化选项下，目标平台的量化精度、内存占用以及延迟等。
3. 提出了一种基于微控制器的视觉智能算法（uVITA）应用流程，旨在打造跨不同微控制器平台的多人员检测解决方案。由此，摄像头可以实时捕获帧，同时显示器会同步显示帧，无论不同平台上的视觉算法速度快或慢。

此应用软件包提供的内容总结如下：

- 提供了一个轻量级的人员检测模型，此模型采用了一种高效且内存访问成本友好的神经网络。
- 给出了详细步骤和实验分析，演示如何在微控制器上使用eIQ Glow工具将对象检测模型转换为目标文件。
- 提出了一个基于微控制器的视觉智能算法应用流程，在i.MX RT1060EVK和i.MX RT1170EVK上构建了多人员检测的工程。

表1. 术语表

术语	说明
ML	机器学习
CNN	卷积神经网络
MAC	内存访问成本
RAM	随机存取存储器
NMS	非最大值抑制

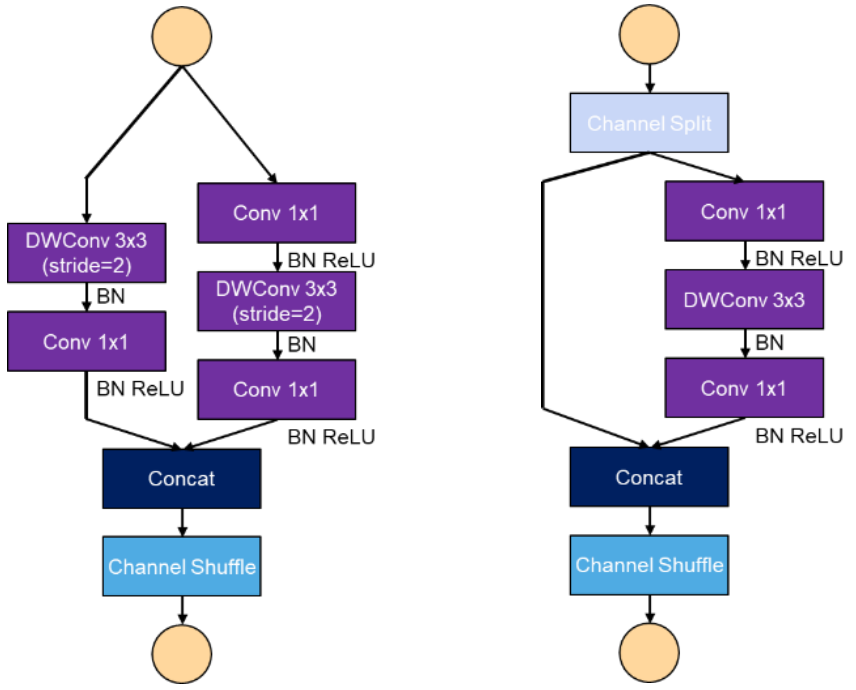
2 多人员检测神经网络

多人员检测在诸如机器人和安全等应用领域中发挥着重要的作用。研究表明，深度卷积神经网络（CNN）在这些目标检测任务中通常具有更高的准确性。因此，业界提出了许多基于CNN的方法来提高目标检测的性能，包括Yolo[2]、ResNet[3]、SSD[4]等。除了检测精度之外，计算复杂度也是一个重要的因素尤其是对于边缘设备的应用。因此，提出了许多轻量级CNN，如Xception、MobileNet[5]和ShuffleNet[6]，以优化速度-精度之间的平衡。其中，ShuffleNet-V2呈现出一种更好的轻量级和高精度特性[1]。此外，通过Arm平台上的验证，ShuffleNet-V2的内存访问成本（MAC）更低。因此，在本应用中，采用ShuffleNet-V2架构来训练多人员检测。

2.1 采用ShuffleNet-V2的神经网络

为了导出一个人员检测的轻量级机器学习模型，我们训练了一个高效神经网络，其采用了ShuffleNet-V2架构以实现速度-精度之间的平衡。ShuffleNet是一种最先进的网络架构，广泛应用于手机等低端设备[1]。

图1所示为使用ShuffleNet-V2训练的模型中的构建模块。其中，**Block-1**和**Block-2**构成了神经网络的主要结构，用于维护多个通道，避免了密集的卷积和过多的分组[1]。这种方式有助于减少MAC。具体来说，**Block-1**有助于缩小特征图的大小，并仅保留有用的信息。同时，引入了**通道洗牌**操作，以实现不同通道组之间的信息交换并提高准确性[1]。**Block-2**引入了一种名为“通道拆分”的简单运算符，将特征分为两个分支。一个分支保持不变，作为身份分支，而另一个分支则尝试探索更多信息。



(a) Block -1　　(b) Block -2

图1. ShuffleNet-V2人员检测模型的主要结构

然后，将提取的特征发送到带有5×5并行卷积的Inception结构中，如图2中的**Block-3**所示。它旨在整合不同感知领域的特征，从而使单个检测头能够适应不同尺度的对象检测。最后，使用带有3个分支的无锚检测头，如图2中的**Block-4**所示，其中带有*softmax*激活层的第一个分支负责检测置信度。第二分支的输出提供被检测对象的坐标。同时，带有*softmax*激活层的最后一个分支负责检测类别。在此应用中，只有一个对象，即人体，因此最后一个分支实际上不起作用。

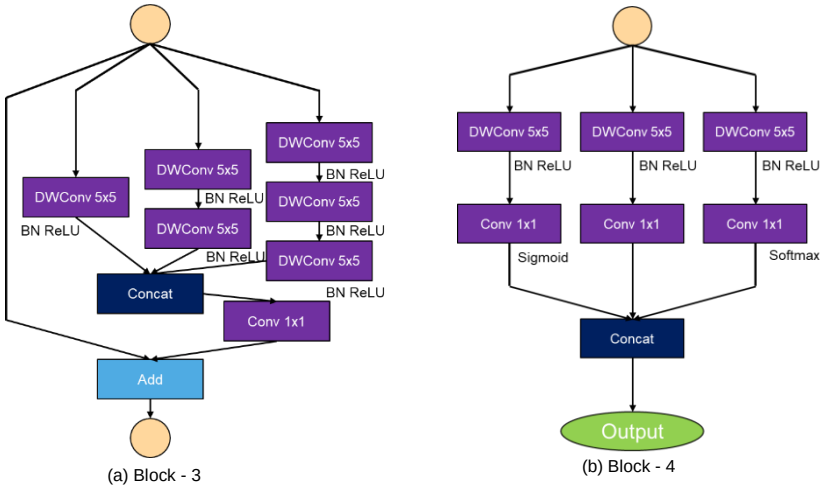


图2. ShuffleNet-V2人员检测模型的构建模块

将构建模块反复堆叠，以构建完整的多人员检测器。表2总结了整个网络的结构。请注意，在提出的人员检测器中，输入的高度和宽度分别设置为192和320，保持了约为9:16的高宽比。这是因为RT1170EVK或RT1060EVK上的摄像头的高宽比也都是9:16左右。因此，摄像头和人员检测器的输入之间的匹配将不会出现失真。

表2. 带有层信息的人员检测模型的整体架构

索引	层	输出大小	核心大小	步幅	重复	输出通道
0	Image	192×320	—	—	—	3
1	Conv1	96×160	3×3	2	1	24
	MaxPool	48×80	3×3	2		
2	Block-1	24×40	3×3和1×1	2	1	48
	Block-2	24×40	3×3和1×1	1	3	48
3	Block-1	12×20	3×3和1×1	2	1	96
	Block-2	12×20	3×3和1×1	1	7	96
4	Block-1	6×10	3×3和1×1	2	1	192
	Block-1	6×10	3×3和1×1	2	1	192
5	Concat	12×20	—	—	—	336
	Conv2	12×20	1×1	1	1	96
6	Block-3	12×20	5×5和1×1	1	1	96
7	Block-4	12×20	5×5和1×1	1	1	6

在所给的人员检测器网络中，特征图的输出大小为12×20，对于网络的输入分辨率（192×320），保持着16的下采样率。此外，所给网络的最终输出有6个通道。其中，第一个通道和最后一个通道分别提供对象的置信度和类别。置信度和类别信息位于相应的网格中，如图3所示。其他4个通道分别对应于中心位置的X坐标和Y坐标，以及对象的宽度和高度。然后，通过6个输出通道的信息提取出相关对象对应的候选框，如图3所示。最后，利用非最大值抑制（NMS）策略对候选框进行筛选，以得到检测结果。

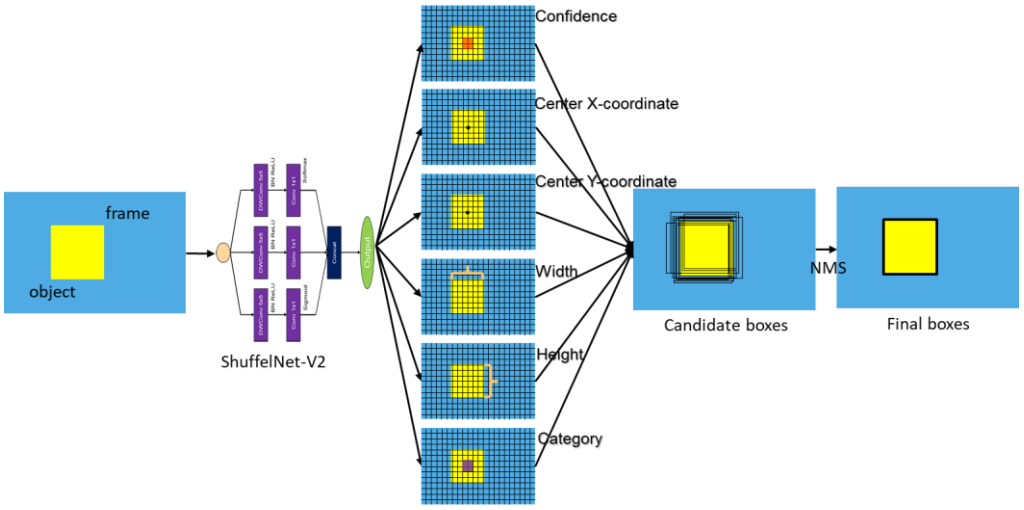


图3. 采用ShuffleNet-V2的人员检测机器学习算法的流程

2.2 神经网络的预处理与后处理

所给的神经网络使用COCO和PASCAL-VOC这两个用于多目标检测的流行数据集进行训练。这些数据集中有很多种对象。但我们只需要“人员”这一个类别，在该应用中对人员检测器进行训练。因此，提供了一个预处理功能，来准备与人员相关的标签，而所有的其他对象都会被视为背景。

为了评估训练模型的性能，*Scripts*文件夹中提供了一个静态图像测试和一个动态视频测试。在测试模型或将模型部署到一个真实的边缘设备之前，用户应注意3个要点。第一个要点是在将图像数据发送到模型之前对其进行预处理。在所提出的人员检测器中，图像的预处理如下所示：

$$Input = Im / 255 \tag{1}$$

换句话说，在将图像发送到该模型之前，必须将其归一化到0和1之间。另一个要点是模型输出的后处理。由于该人员检测器以无锚的方式提取对象的候选框，因此后处理比传统的Yolo方法略微简单些[2]。如图3所示，每个输出网格中的最终混合置信度计算公式如下：

$$mc_{i,j} = \omega \times confidence_{i,j} + (1 - \omega) \times category_{i,j} \tag{2}$$

在等式2中，$confidence_{i,j}$表示图3中第一个通道的第*i*行和第*j*列的值。$category_{i,j}$表示图3中最后一个通道的第*i*行和第*j*列的值。$\omega < 1$表示*confidence*通道的权重。通过每个网格中的最终混合置信度的特定的阈值$mc_{i,j}$，可以筛选相关对象的候选框。然后，计算出相应的中心坐标如下：

$$cx_{i,j} = i + \left(\frac{2}{\left(1+\exp\left(-2 \times x\_offset_{i,j}\right)\right)} - 1 \right) \Big/ output\_w \tag{3}$$

$$cy_{i,j} = j + \left(\frac{2}{\left(1+\exp\left(-2 \times y\_offset_{i,j}\right)\right)} - 1 \right) \Big/ output\_h \tag{4}$$

同时，在(i, j)处激活的相关对象的高度和宽度如下所示：

$$h_{i,j} = \text{sigmoid}\left(sh_{i,j}\right) \tag{5}$$

$$w_{i,j} = \text{sigmoid}\left(sw_{i,j}\right) \tag{6}$$

2.3 算法的性能

通过给定的人员检测模型以及预处理/后处理，可以根据一些静态图像示例得出算法结果，如图4所示。它说明了模型对每个框中人员的置信度和位置坐标的预测结果。这些结果证明了该模型在各种环境条件下都能进行稳定且可靠的人员检测。此外，该应用中还有一个视频测试脚本，可以让用户通过它亲自验证该人员检测器的性能。

图4. 人员检测器的算法性能

3 使用Glow NN进行eIQ推理

为了在i.MX RT跨界MCU上部署一个神经网络，恩智浦eIQ ML软件开发环境提供了友好高效的工具，如Glow、TensorFlow Lite Micro或DeepViewRT。在此应用中，支持使用Glow进行提前编译，将原始神经网络转换为目标文件，并进一步将模型部署到MCU上。

3.1 使用Glow NN进行量化和编译

Glow能够在边缘设备上进行神经网络模型的推理。要使用Glow编译该模型，通常需要两个阶段将模型转换为目标文件。在第一个阶段，Glow优化器使用给定的校准数据和模型进行量化分析。为了帮助用户重现被量化的目标文件，该应用程序提供了`onnx`格式的浮点模型和分辨率为192×320的132个图像的校准数据。这些图像是从WIDER FACE数据集生成的。用户可以首先使用以下命令生成*ym*l文件：

```
image-classifier.exe -input-image-dir=Data/Calibration -image-mode=0to1 -image-
layout=NCHW -image-channel-order=BGR -model=Model/Onnx/dperson_shufflenetv2.onnx
 -model-input-name=input.1 -dump-profile=Model/Glow/dperson_shufflenetv2.yml
```

有关Glow操作的更多帮助信息，请参阅《*eIQ Glow Ahead of Time*用户指南》（EIQGLOWAOTUG）。一旦导出了*ym*l文件，就可以进入第二阶段，利用专用的后端硬件功能来进行优化。在此应用中，目标平台是Arm Cortex-M7内核。因此，可以通过将float32模型编译成int8 bundle来生成二进制目标文件（bundle），从而以更少的内存消耗和更快的推理速度获得Cortex-M7的支持。为此，使用以下命令生成8位bundle。

```
model-compiler.exe -model=Model/Onnx/dperson_shufflenetv2.onnx -model-
input=input.1,float,[1,3,192,320] -emit-bundle=Model/Glow/int8_bundle -
backend=CPU -target=arm -mcpu=cortex-m7 -float-abi=hard -load-profile=Model/
Glow/dperson_shufflenetv2.yml -quantization-schema=symmetric_with_power2_scale -
quantization-precision-bias=Int8
```

另一个bundle的编译选项是利用Arm CMSIS-NN库，通过以下命令来提升性能。

```
model-compiler.exe -model=Model/Onnx/dperson_shufflenetv2.onnx -model-
input=input.1,float,[1,3,192,320] -emit-bundle=Model/Glow/int8_cmsis_bundle -
backend=CPU -target=arm -mcpu=cortex-m7 -float-abi=hard -load-profile=Model/
Glow/dperson_shufflenetv2.yml -quantization-schema=symmetric_with_power2_scale -
quantization-precision-bias=Int8 -use-cmsis
```

然后，从Glow编译器的输出中导出glow bundle。在-emit-bundle指定的目录中生成4个文件。在本应用程序中，这4个文件分别为：

- **dperson_shufflenetv2.h**——bundle头文件（API）。
- **dperson_shufflenetv2.o**——bundle目标文件（code）。
- **dperson_shufflenetv2.weights.bin**——二进制格式的模型权重。
- **dperson_shufflenetv2.weights**——文本格式为C文本数组的模型权重。

dperson_shufflenetv2.h文件包含内存占用情况和推理函数API。**dperson_shufflenetv2.o**文件是一个目标文件，其中包含以库形式编译的模型代码。通常，目标文件的大小大于其自身所需的闪存大小。

3.2 内存占用和延迟分析

在此应用中，所给的人员检测器显示了所需的内存及模型延迟方面的轻量级特征，如表3所示。众所周知，Glow不会动态分配内存。因此，由Glow生成的量化模型所需的内存大小在bundle头文件中提供。表3对这些信息进行了汇总。

可以发现，当不使用CMSIS-NN时，人员检测器的固定权重占用了Glow生成的235904字节；而使用CMSIS-NN时，则占用246848字节。在推理的过程中，可以从Flash或RAM读取权重，而权重会占用指定大小的Flash区域。另一种Flash的占用情况是由以库格式生成的目标代码产生的，在不使用CMSIS-NN的情况下需要76192个字节，而使用CMSIS-NN时需要25840个字节。

输入和输出数据缓冲区所需的内存为743040字节，此部分必须从RAM中分配。输入/输出缓冲区与所给模型的输入分辨率和输出特征图的大小有关。例如，人员检测模型的输入分辨率为192×320×3，输出形状为12×20×6。缓冲区总大小为：

$$192 \times 320 \times 3 \times 4 + 12 \times 20 \times 6 \times 4 = 743040 \text{ bytes} \quad (7)$$

激活缓冲区被视为中间计算所需的临时存储器区，必须从RAM中分配。对于所给的模型，在不使用CMSIS-NN的情况下，激活缓冲区的大小为552960字节，在使用CMSIS-NN时为645120字节。

表3. 人员检测模型的内存占用和延迟

Glow的编译选项	权重 (Flash)	输入/输出 (SDRAM)	激活 (SDRAM)	库 (Flash)	延迟
8位 无CMSIS-NN	235,904	743,040	552,960	76,912	778 ms (RT1060)
					495 ms (RT1170)
8位 有CMSIS-NN	246,848	743,040	645,120	25,840	353 ms (RT1060)
					237 ms (RT1170)

分配到不同的RAM区域的激活缓冲区，可能会对延迟产生不同的性能影响。例如，在片上RAM（OCRAM）中分配激活缓冲区可以将RT1170的延迟从230 ms减少到161 ms，如图5所示。这是可以理解的，因为OCRAM中的模型推理计算效率高于SDRAM。更重要的是，将激活缓冲区分配到OCRAM中可以避免CPU与DMA、PXP等其他资源之间的内存访问冲突。这一点将在下一节中详细讨论。

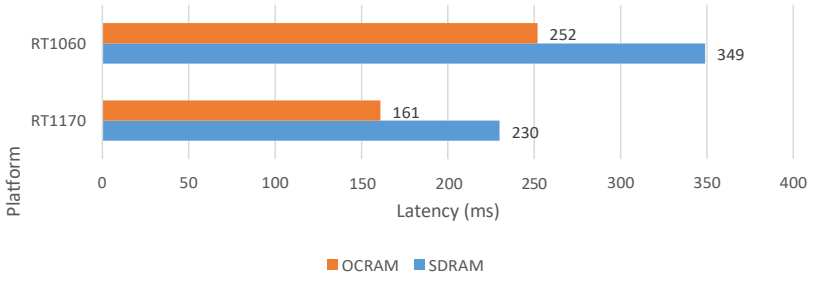


图5. 将激活缓冲区分配在不同的RAM区域时的模型推理延迟

3.3 量化精度的验证

在把一个量化模型部署到边缘设备之前，应首先验证其量化精度。图6所示为原始浮点模型和使用Glow量化的两个不同版本分别给出的预测结果。可以发现，无论是否使用CMSIS-NN，Glow的量化结果都与浮点模型给出的结果相对一致，特别是对于背景简单和人员不重叠的样本。由于8位格式的模型存在一定的信息损失，因此会出现如图6(c)中的红色和绿色框所示的失配情况。尽管如此，与原始浮点模型相比，量化模型的整体性能仍然更加可靠，如图6(a)、(b)和(d)所示。

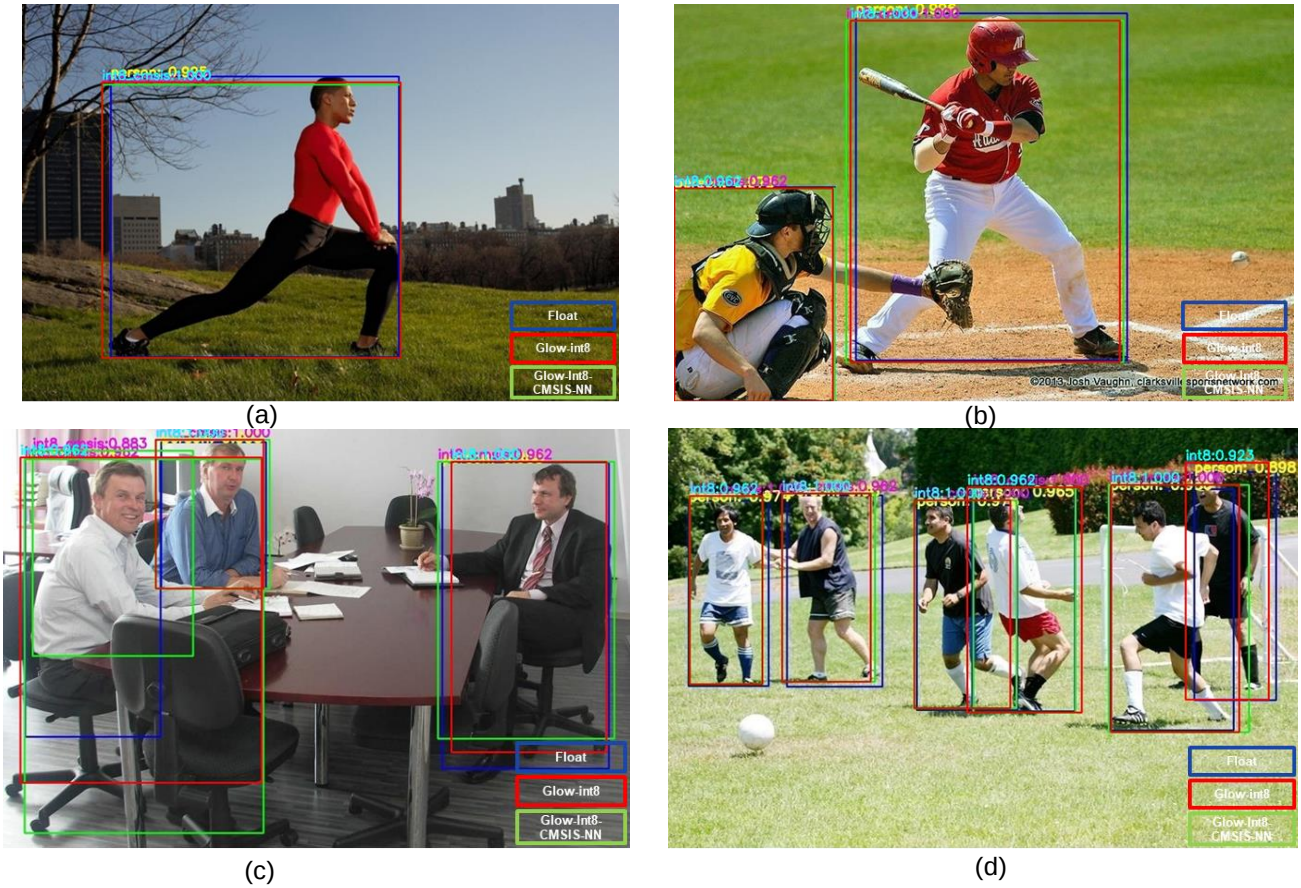


图6. 人员检测模型的量化性能

4 人员检测器的应用

本节提供了附加的指导和解释，介绍了集成在真实边缘设备上的机器学习人员检测器。该应用软件包还有另一个“入门”文档，可帮助用户轻松复现示例应用。

4.1 系统设计

恩智浦的跨界MCU通过丰富的硬件资源提供高性能的智能功能。例如，RT1060EVK和RT1170EVK的处理器分别搭载了主频高达600 MHz和1 GHz的Arm Cortex-M7内核，如表4所示。此外，所给的平台为视觉应用提供了充足的内存。RT1060和RT1170中内置了通用2D硬件加速（PXP）功能，以帮助快速实现通用的图像处理功能并节省CPU带宽。所支持的2D处理功能包括图像旋转、图像缩放和色彩空间转换等。如表4所示，RT1060EVK和RT1170EVK上所使用的摄像头分别为MT9M114和OV5640。在本应用中，RT1060EVK上的摄像头分辨率设置为480*272，与其显示器的分辨率保持一致。同样，RT1170EVK上的摄像头和显示器的分辨率均设置为1280*720。

表4. 恩智浦跨界MCU中面向机器学习视觉应用的硬件资源

	RT1060EVK	RT1170EVK
处理器	MIMXRT1062DVL6A 600 MHz Arm Cortex-M7内核	MIMXRT1176DVMAA 1 GHz Arm Cortex-M7内核

表4. 恩智浦跨界MCU中面向机器学习视觉应用的硬件资源（续）

	RT1060EVK	RT1170EVK
		400 MHz Arm Cortex-M4内核
存储器	• 1 MB片上RAM • 256 MB SDRAM存储器 • 512 MB Hyper Flash • 64 MB QSPI Flash	• 2 MB片上RAM • 512 MB SDRAM存储器 • 512 MB Octal Flash • 128 MB QSPI Flash
摄像头	MT9M114或OV7725	OV5640
显示器	TFT：RK043FN02H-CT 分辨率：480*272	TFT：RK055HDMIPI4M 分辨率：1280*720
通用2D (PXP)	• 图像旋转（90°、180°、270°） • 图像缩放 • 色彩空间转换 • ...	• 图像旋转（90°、180°、270°） • 图像缩放 • 色彩空间转换 • ...

为了使基于机器学习的人员检测器可以轻松部署在不同的开发板上，我们提出了一个基于微控制器的跨平台视觉智能算法（uVITA）系统，来管理摄像头、显示器以及算法的任务。此外，uVITA系统还能够在机器学习视觉应用方面提供更好的用户体验。例如，摄像头应实时捕获帧。同时，无论算法的速度是快（在RT1170上）还是慢（在RT1060上），显示器都应同步显示捕获的帧。我们提出的系统架构如图7所示，其中摄像头的任务负责捕获图像帧，并将其发送到算法任务和显示任务，并提供相应的图像格式和大小。同时，算法的任务是利用馈送的数据对机器学习的模型进行推理。然后，从模型中提取结果，并使用提出的后处理功能对预测结果进行筛选。最后，显示任务负责在显示器上显示图像帧和算法结果。由于图7所示的3个任务是在FreeRTOS的管理下并行运行的，因此如果摄像头和显示器的优先级高于算法任务，则会实时处理它们的进程。

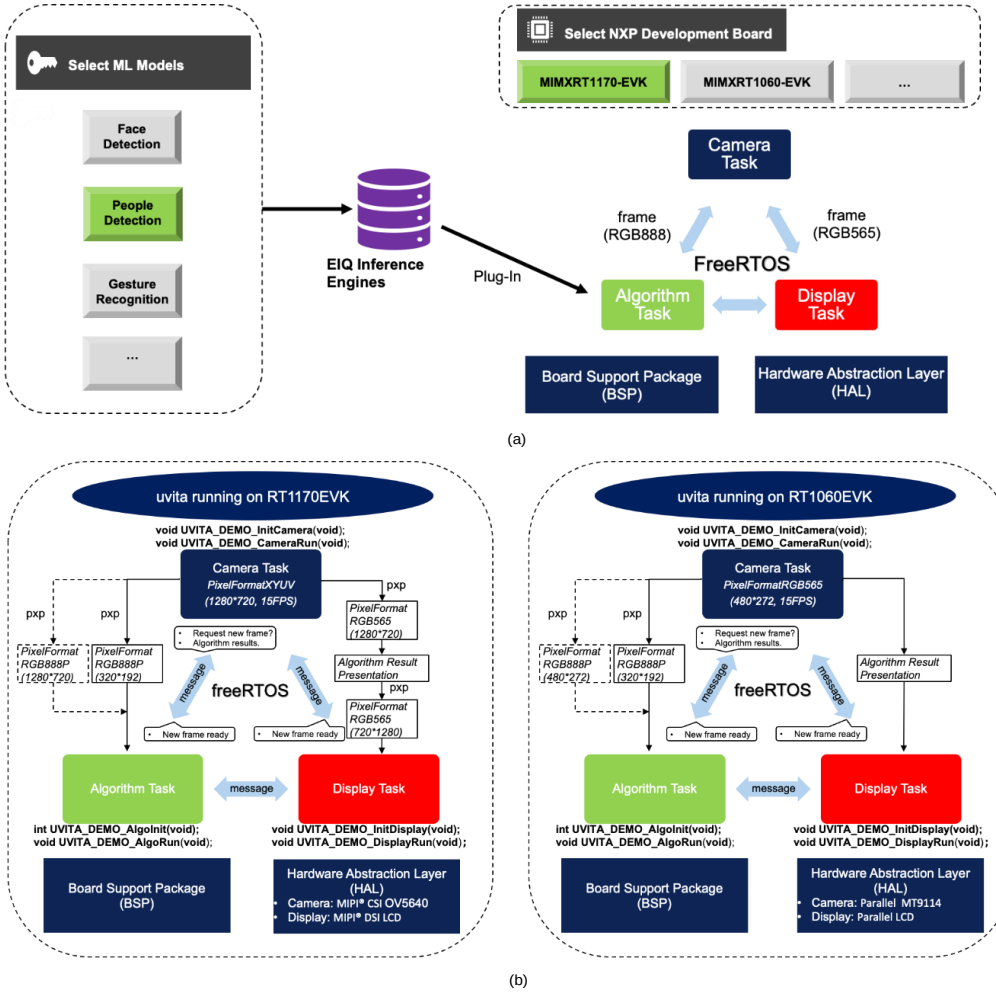


图7. 基于微控制器的视觉智能算法（uVITA）系统

图像比例和图像格式的转换是通过恩智浦跨界MCU支持的PXP加速功能来实现的。根据人员检测器的输入要求，通过PXP功能将摄像头接收到的图像帧直接转换为分辨率为320*192的RGB888格式。此外，摄像头捕获的图像帧通过PXP功能转换为RGB565格式显示在显示器上，帧率为15PFS。因此，尽可能地节省了CPU资源，使其能够以更大的带宽对人员检测器的神经网络进行推理。由于显示面板处于垂直模式，因此需要旋转90°才能在RT1170EVK的显示器上显示帧。

4.2 整体性能

在此应用中，首先要讨论在MCUXpresso IDE工程中人员检测器对内存需求的影响。如表5所示，对于RT1060上的应用工程，所有缓冲区在一开始均已设置为零。然后，在SDK工程的基础上，添加了摄像头、显示器和FreeRTOS的支持，目的是关注应用工程对内存需求的影响。用户可以确定特定的机器学习模型是否适合特定的电路板。除了表5中列出的内存需求外，还需要额外1020K字节的存储区来承载摄像头捕获的帧数据并将其显示在显示器上。具体来说，摄像头的分辨率设置为272*480，格式为RGB565，因此其数据缓冲区占用2*272*480*2字节。此外，显示器的分辨率设置为272*480，格式为RGB565，因此其数据缓冲区占用了2*272*480*2字节。上述所有的缓冲区均在SDRAM中，共占用1020 kB。

表5. MCUXpresso SDK为基于RT1060EVK的机器学习人员检测器编译的工程大小

说明	Flash（字节）	RAM（字节）	更改（字节）	详细信息
裸机	109,144	26,372	基线	基线SDK工程，提供摄像头、显示器和FreeRTOS支持。
在算法任务中为静态输入图像增加存储区	109,144	211,292	+184,320 RAM	算法任务的帧缓冲区为RGB888格式，分辨率为192*400，占用192*320*3=184320字节。
添加 *.o库*	134,984	211,292	+25,840 Flash	Glow编译的人员检测库 *.o*文件。 ***注：** 这比电脑硬盘上 .o文件的大小要小。*
添加输入/输出和激活缓冲区	134,984	1,599,604	+1,388,312 RAM	为可变权重（模型输入/输出数据，743040字节）和激活（模型中间结果，645120字节）静态地分配存储区。
在Flash中添加权重	381,832	1,599,604	+246,848 Flash	如果从Flash读取权重，不会影响RAM的占用，但需要246848字节的非易失性存储器。
在RAM中添加权重	381,832	1,846,452	+246,848 RAM	如果从RAM读取权重，则该工程需要246848个额外的RAM字节。这是可选的，但可能会减少推理时间。

在RT1170EVK上，人员检测器的内存需求与RT1060EVK类似，但存在一些区别。主要区别在于用于承载摄像头帧和显示器帧数据的额外缓冲区的大小，这些帧的分辨率高于RT1060EVK的摄像头帧和显示器帧。对于RT1170EVK，摄像头和显示器的分辨率均为1280*720，采用YUYV格式，因此其数据缓冲区占用2*720*1280*4字节。同时，显示器的分辨率设置为720*1280，格式为RGB565，因此其数据缓冲区占用2*720*1280*2字节。此外，在将帧发送到显示器之前还有一个额外的缓冲区，用于保存单帧以显示算法结果。这个缓冲区需要720*1280*2字节。因此，上述的所有缓冲都均在SDRAM中处理，共占用12600 kB。

另一个问题是在实际的边缘应用中人员检测器的延迟影响。由于多任务系统占用了CPU资源和内存访问的带宽，因此可能无法完全为模型推理任务提供服务。例如，当使用CMSIS-NN优化的Glow编译模型的激活缓冲区分配在SDRAM中时，在RT1170EVK上编译模型的理想延迟为230 ms。然而，当摄像头任务和显示器任务同时运行时，算法任务中编译模型的延迟会增加到280 ms。其主要原因在于CPU与PXP和DMA等其他硬件加速器之间的内存访问存在带宽限制。因此，更理想的内存配置是将编译模型的激活缓冲区分配在OCRAM中，同时将摄像头和显示器的数据缓冲区放在SDRAM中。这样就可以避免内存访问的冲突。如表5所示，当将激活缓冲区分配在OCRAM中时，可以减少延迟的影响。

表6. 在RT1170EVK上的人员检测器的延迟影响

模型	权重（246848）	激活（645120）	延迟（理想值）	延迟（实际应用）
Shufflenetv2 EIQ-Glow 8位 带CMSIS-NN	Flash	SDRAM	230 ms	281 ms
	Flash	OCRAM	161 ms	165 ms

5 结论

在本应用笔记中，介绍了基于恩智浦i.MX RT1060和RT1170跨界MCU的多人员检测器。所给的人员检测器首先使用一个基于ShuffleNet-V2架构的高效神经网络来实现，并优化了速度-精度之间的平衡。然后将eIQ-Glow的量化和编译过程引入到经过训练的人员检测器中，从而获得MCU上相应的可执行代码。同时，分析了转换后模型的内存占用、延迟和量化精度。最后，分别在RT1060EVK和RT1170EVK上演示了如何使用所提出的uVITA系统构建人员检测器。因此，摄像头可以实时捕获帧。同时，无论算法速度的快慢，显示器都会同步显示捕获的帧。此应用程序可以作为一个原型，用户在此基础上使用恩智浦的跨界MCU构建自己的机器学习视觉程序。有了自主开发的机器学习模型，用户可以基于eIQ ML软件开发环境构建与此应用类似的智能产品。

6 参考资料

1. Ma N、Zhang X、Zheng H T等。“Shufflenet v2：高效CNN架构设计实用指南”（Shufflenet v2: Practical guidelines for efficient cnn architecture design），《欧洲计算机视觉大会论文集》。2018：116-131。
2. Redmon J、Divvala S、Girshick R等。“只需查看一次：统一的实时对象检测”（You only look once: Unified, real-time object detection），《IEEE计算机视觉与模式识别大会论文集》。2016：779-788。
3. He, K.、Zhang, X.、Ren, S.、Sun, J：“深度残差网络中的身份映射”（Identity mappings in deep residual networks），“欧洲计算机视觉大会”。第630–645页。施普林格出版社（2016）。
4. Liu W、Anguelov D、Erhan D等。Ssd：“单点多框检测器”（Single shot multibox detector），“欧洲计算机视觉大会”。施普林格出版社，卡姆，2016：21-37。
5. Howard，A.G.、Zhu，M.、Chen，B.、Kalenichenko，D.、Wang，W.、Weyand，T.、Andreetto，M.、Adam，H.：“Mobilenets：面向移动视觉应用的高效卷积神经网络”（Mobilenets: Efficient convolutional neural networks for mobile vision applications）。arXiv预印本arXiv：1704.04861（2017）。
6. Zhang, X.、Zhou, X.、Lin, M.、Sun, J.：“Shufflenet：一种面向移动设备的高效卷积神经网络”（Shufflenet: An extremely efficient convolutional neural network for mobile devices）。arXiv预印本arXiv：11707.01083（2017）。

7 关于本文中源代码的说明

本文中所示的示例代码具有以下版权和BSD-3-Clause许可：

8 修订历史

表7汇总了对本文档的修订。

表7. 修订历史

版本号	日期	实质性变更
0	2023年5月8日	初版发布

9 Legal information

9.1 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.

9.2 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 http://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 data sheet 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.

9.3 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: http://www.nxp.com.cn

Date of release: 8 May 2023
Document identifier: AN13924