第57卷第2期，2023年4月

ANALOG DEVICES 超越一切可能™ | ADI AnalogDialogue

卷积神经网络的硬件转换：什么是机器学习？—第三部分

Ole Dreessen，现场应用工程师

摘要

本系列文章由三部分组成，主要探讨卷积神经网络(CNN)的特性和应用。CNN主要用于模式识别和对象分类。作为系列文章的第三部分，本文重点解释如何使用硬件转换卷积神经网络(CNN)，并特别介绍使用带CNN硬件加速器的人工智能(AI)微控制器在物联网(IoT)边缘实现人工智能应用所带来的好处。系列文章的前两篇文章为《卷积神经网络简介：什么是机器学习？——第一部分》和《训练卷积神经网络：什么是机器学习？——第二部分》。

简介

AI应用通常需要消耗大量能源，并以服务器农场或昂贵的现场可编程门阵列(FPGA)为载体。AI应用的挑战在于提高计算能力的同时保持较低的功耗和成本。当前，强大的智能边缘计算正在使AI应用发生巨大转变。与传统的基于固件的AI计算相比，以基于硬件的卷积神经网络加速器为载体的智能边缘AI计算具备惊人的速度和强大的算力，开创了计算性能的新时代。这是因为智能边缘计算能够让传感器节点在本地自行决策而不受5G和Wi-Fi网络数据传输速率的限制，为实现之前难以落地的新兴技术和应用场景提供了助力。例如，在偏远地区，传感器级别的烟雾/火灾探测或环境数据分析已成为现实。这些应用支持电池供电，能够工作很多年的时间。本文通过探讨如何采用带专用CNN加速器的AI微控制器实现CNN的硬件转换来说明如何实现这些功能。

采用超低功耗卷积神经网络加速器的人工智能微控制器

MAX78000是一款有超低功耗CNN加速器的AI微控制器片上系统，能在资源受限的边缘设备或物联网应用中实现超低功耗的神经网络运算。其应用场景包括目标检测和分类、音频处理、声音分类、噪声消除、面部识别、基于心率等健康体征分析的时间序列数据处理、多传感器分析以及预测性维护。

图1为MAX78000的框图，其内核为带浮点运算单元的Arm® Cortex®-M4F内核，工作频率高达100 MHz。为了给应用提供足够的存储资源，MAX78000还配备了512 kB的闪存和128 kB的SRAM。该器件提供多个外部接口，例如I²C、SPI、UART，以及用于音频的I²S。此外，器件还集成了60 MHz的RISC-V内核，可以作为一个智能的直接存储器访问(DMA)引擎从/向各个外围模块和存储（包括闪存和SRAM）复制/粘贴数据。由于RISC-V内核可以对AI加速器所需的传感器数据进行预处理，因而Arm内核在此期间可以处于深度睡眠模式。推理结果也可以通过中断触发Arm内核在主应用程序中执行操作，通过无线传输传感器数据或向用户发送通知。

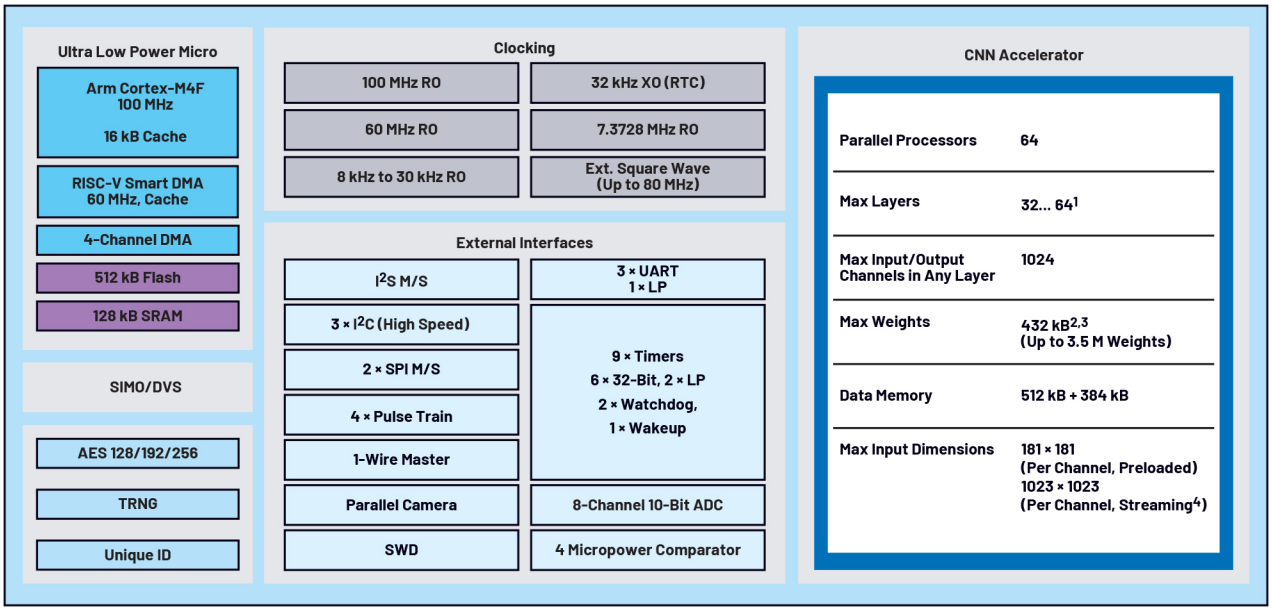

图 1. MAX78000 的结构框图

具备用于执行卷积神经网络推理的专用硬件加速器单元是MAX7800x系列微控制器的一个显著特征，这使其有别于标准的微控制器架构。该CNN硬件加速器可以支持完整的CNN模型架构以及所有必需的参数（权重和偏置），配备了64个并行处理器和一个集成存储器。集成存储器中的442 kB用于存储参数，896 kB用于存储输入数据。不仅存储在SRAM中的模型和参数可以通过固件进行调整，网络也可以实时地通过固件进行调整。器件支持的模型权重为1位、2位、4位或8位，存储器支持容纳多达350万个参数。加速器的存储功能使得微控制器无需在连续的数学运算中每次都要通过总线获取相关参数——这样的方式通常伴有高延迟和高功耗，代价高昂。CNN加速器可以支持32层或64层的网络，具体层数取决于池化函数。每层的可编程图像输入/输出大小最多为1024 × 1024像素。

CNN硬件转换：功耗和推理速度比较

CNN推理是一项包含大型矩阵线性方程运算的复杂计算任务。Arm Cortex-M4F微控制器的强大能力可以使得CNN推理在嵌入式系统的固件上运行。但这种方式也有一些缺点：在微控制器上运行基于固件的CNN推理时，计算命令和相关参数都需要先从存储器中检索再被写回中间结果，这会造成大量功耗和时延。

表1对三种不同解决方案的CNN推理速度和功耗进行了比较。所用的模型基于手写数字识别训练集MNIST开发，可对视觉输入数据中的数字和字母进行分类以获得准确的输出结果。为确定功耗和速度的差异，本文对三种解决方案所需的推理时间进行了测量。

表1. 手写数字识别的CNN推理时间和推理功耗，基于MNIST数据集

方案	推理速度 (ms)	每次推理功耗 (µWs)
(1) MAX32630，MNIST网络部署在固件中	574	22887
(2) MAX78000，MNIST网络部署在硬件中	1.42	20.7
(3) MAX78000，MNIST网络部署在硬件中且优化了低功耗性能	0.36	1.1

方案一使用集成Arm Cortex-M4F处理器的MAX32630进行推理，其工作频率为96 MHz。方案二使用MAX78000的CNN硬件加速器进行推理，其推理速度（即数据输入与结果输出之间的时间）比方案一加快了400倍，每次推理所需的能量也仅为方案一的1/1100。方案三对MNIST网络进行了低功耗优化，从而最大限度地降低了每次推理的功耗。虽然方案三推理结果的准确性从99.6%下降到了95.6%，但其速度快了很多，每次推理只需0.36 ms，推理功耗降也低至

仅1.1 µW。两节AA碱性电池（总共6 Wh能量）可以支持应用进行500万次的推理（忽略系统其它部分的功耗）。

这些数据说明了硬件加速器的强大计算能力可以大大助益无法利用或连接到连续电源的应用场景。MAX78000就是这样一款产品，它支持边缘AI处理，无需大量功耗和网络连接，也无需冗长的推理时间。

MAX78000 AI微控制器的使用示例

MAX78000支持多种应用，下面本文围绕部分用例展开讨论。其中一个用例是设计一个电池供电的摄像头，需要能检测到视野中是否有猫出现，并能够通过数字输出打开猫门允许猫进入房屋。

图2为该设计的示例框图。在本设计中，RISC-V内核会定期开启图像传感器并将图像数据加载到MAX78000的CNN加速器中。如果系统判断猫出现的概率高于预设的阈值，则打开猫门然后回到待机模式。

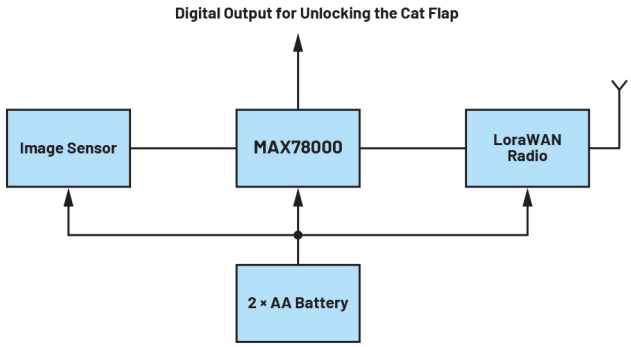

图 2. 智能宠物门框图

开发环境和评估套件

边缘人工智能应用的开发过程可分为以下几个阶段：

第一阶段：AI—网络的定义、训练和量化

第二阶段：Arm固件 — 将第一阶段生成的网络和参数导入C/C++应用程序，创建并测试固件

开发过程的第一阶段涉及建模、训练和评估AI模型等环节。此阶段开发人员可以利用开源工具，例如PyTorch和TensorFlow。MAX78000的GitHub网页也提供全面的资源帮助用户在考虑其硬件规格的同时使用PyTorch构建和训练AI网络。网页也提供一些简单的AI网络和应用，例如面部识别(Face ID)，供用户参考。

图3显示了采用PyTorch进行AI开发的典型过程。首先是对网络进行建模。必须注意的是，MAX7800x微控制器并非都配置了支持所有PyTorch数据操作的相关硬件。因此，必须首先将ADI公司提供的ai8x.py文件包含在项目中，该文件包含MAX78000所需的PyTorch模块和运算符。基于此可以进入下一步骤构建网络，使用训练数据对网络进行训练、评估和量化。这一步骤会生成一个检查点文件，其中包含用于最终综合过程的输入数据。最后一步是将网络及其参数转换为适合CNN硬件加速器的形式。值得注意的是，虽然任何PC（笔记本、服务器等）都可用于训练网络，但如果没有CUDA显卡，训练网络可能会花费很长的时间——即使对于小型网络来说也有可能需要几天甚至几周的时间。

开发过程的第二阶段是通过将数据写入CNN加速器并读取结果的机制来创建应用固件。第一阶段创建的文件通过#include指令集成

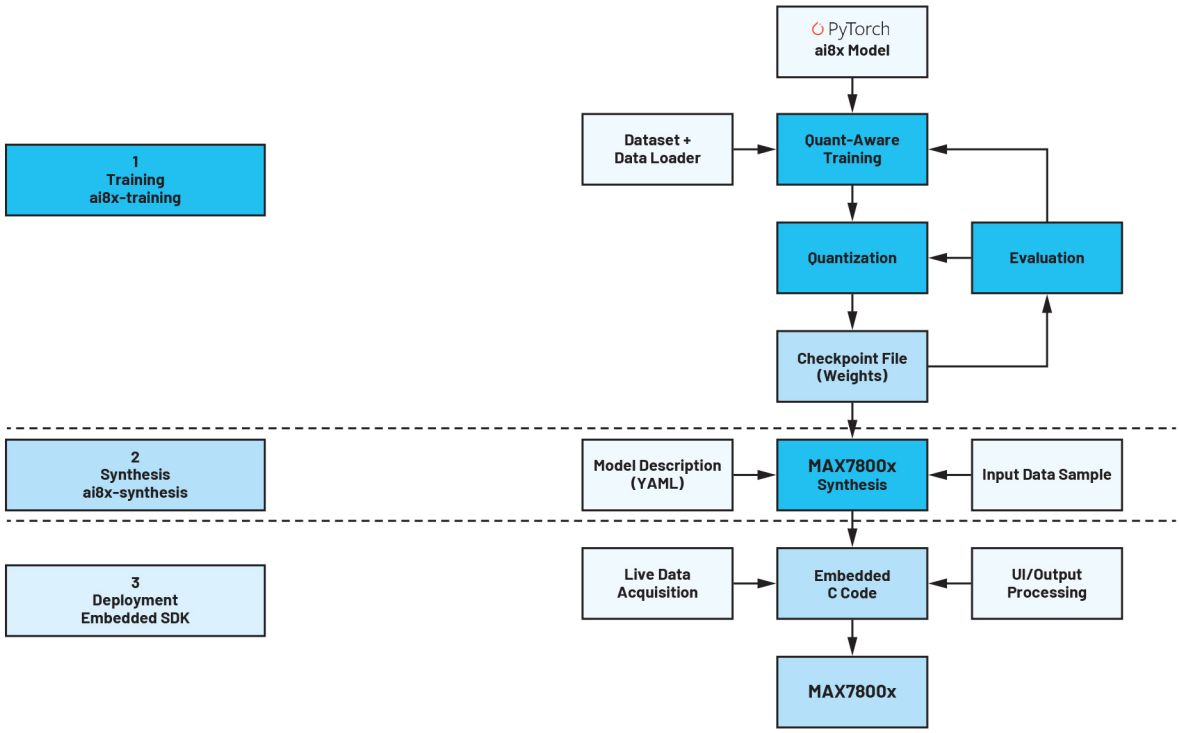

图 3. AI 开发过程

到C/C++项目中。微控制器的开发环境可使用Eclipse IDE和GNU工具链等开源工具。ADI公司提供的软件开发套件(Maxim Micros SDK (Windows))也已经包含了所有开发必需的组件和配置，包括外设驱动以及示例说明，帮助用户简化应用开发过程。

成功通过编译和链接的项目可以在目标硬件上进行评估。ADI开发了两种不同的硬件平台可供选用：图4为MAX78000EVKIT，图5为MAX78000FTHR，一个稍小的评估板。每个评估板都配有一个VGA摄像头和一个麦克风。

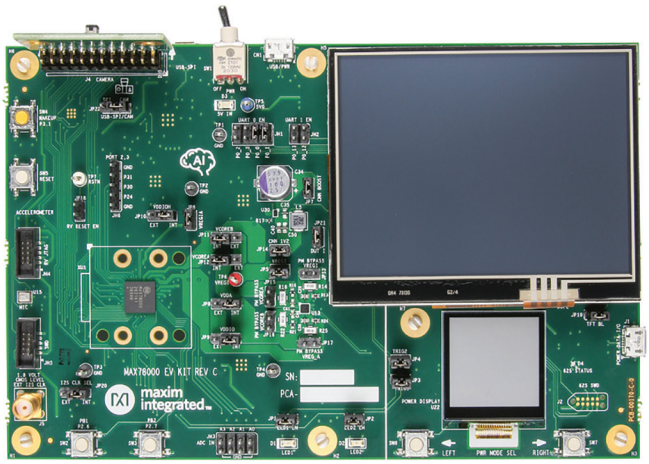

图 4. MAX78000 评估套件

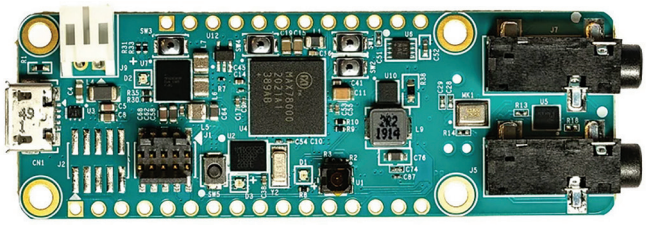

图 5. MAX78000FTHR 评估套件

结论

以前，AI应用必须以昂贵的服务器农场或FPGA为载体，并消耗大量能源。现在，借助带专用CNN加速器的MAX78000系列微控制器，AI应用依靠单组电池供电就可以长时间运行。MAX78000系列微控制器在能效和功耗方面的性能突破大大降低了边缘AI的实现难度，使得新型边缘AI应用的惊人潜力得以释放。欲了解更多信息，请访问超低功耗人工智能(AI) MCU。

参考资料

“研讨会2 - 边缘人工智能：Maxim Integrated MAX78000 AI加速器的实践介绍”。ADI公司。

视频系列：了解人工智能。ADI公司。

PyTorch徽标经知识共享版权归属-相同方式共享 4.0 国际公共许可证许可使用。

作者简介

Ole Dreessen是ADI公司的现场应用工程师，于2014年加入ADI公司，此前曾在Avnet Memec和Macnica任职，负责支持通信产品和高性能微处理器。Ole在微控制器和安全方面拥有广泛的专业知识，拥有丰富的会议主讲经验。在业余时间，他是混沌计算机俱乐部的活跃成员，主要研究逆向工程和嵌入式安全等概念。

©2023 Analog Devices, Inc. 保留所有权利。商标和注册商标属各自所有人所有。

“超越一切可能”是ADI公司的商标。