# 掌控人工智能的力量：使用莱迪思sensAI快速上手

**莱迪思半导体白皮书**

**2019年1月**

人工智能（AI）如今无处不在。这项革命性科技正逐渐渗透到更多领域，影响范围之广将远超出你的想象。不管从事什么业务，每家公司似乎都或多或少与AI产生联系。尤其是如今人们想方设法将AI运用到自动驾驶汽车、物联网（IoT）、网络安全、医疗等诸多领域。企业领导者应当深刻了解如何将AI运用到他们的产品之中，如果率先采用AI获得成功，迟迟未行动的后来者将会陷入困境。

然而AI应用种类各异，各有千秋。不同的应用领域要求的AI技术也不尽相同。目前最受关注的应用类别当属嵌入式视觉。这一领域的AI使用所谓的卷积神经网络（CNN），试图模拟人眼的运作方式。在这篇AI白皮书中，我们主要关注视觉应用，当然其中许多概念也适用于其他应用。


**了解详情：**

www.latticesemi.com

**在线联系我们：**

www.latticesemi.com/contact
www.latticesemi.com/buy

# 目录

# 网络边缘AI的要求

AI涉及创造一个工作流程的训练模型。然后该模型在某个应用中对现实世界的情况进行推理。因此，AI应用有两个主要的生命阶段：训练和推理。

训练是在开发过程中完成的，通常在云端进行。推理作为一项持续进行的活动，则是通过部署的设备完成。因为推理涉及的计算问题会非常复杂，目前大部分都是在云端进行。但是做决策的时间通常都十分有限。向云端传输数据然后等待云端做出决策非常耗时。等到做出决策，可能为时已晚。而在本地做决策则能节省那宝贵的几秒钟时间。

这种实时控制的需求适用于需要快速做出决策的诸多领域。例如人员侦测：

智能家居电器

智能音频/视频消费类电子产品

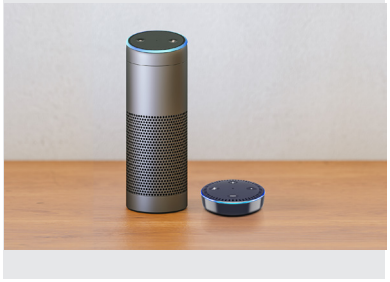
智能门铃

自动售货机

安全摄像头

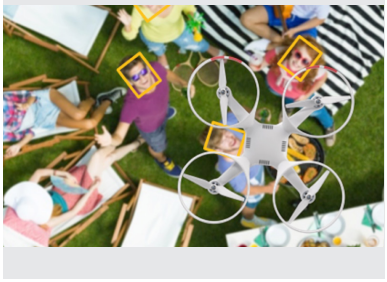
智能门

其他实时在线的应用包括：



智能音箱

零售商店摄像头



自拍无人机

收费站摄像头

机器视觉


汽车后装市场摄像头

在快速决策这种需求的推动下，目前将推理过程从云端转移到“网络边缘”的诉求异常强烈——即在设备上收集数据然后根据AI决策采取行动。这将解决云端不可避免的延迟问题。

本地推理还有两个好处。第一个就是隐私安全。数据从云端来回传输，以及储存在云端，容易被入侵和盗取。但如果数据从未到达设备以外的地方，出现问题的几率就小得多。

另一个好处与网络带宽有关。将视频传送到云端进行实时处理会占用大量的带宽。而在本地做决策则能省下这部分带宽用于其他要求较高的任务。

此外：

- 这类设备通常都是使用电池供电——或者，如果是电源直接供电，两者都有散热限制，从而给设备的持续使用造成限制。而与云端通信的设备需要管理自身的功耗的散热问题。
- AI模型演化速度极快。在训练始末，模型的大小会有极大差异，并且在进入开发阶段以前，可能无法很好地估算所需计算平台的大小。此外，训练过程发生的细微改变就会对整个模型造成重大影响，增加了变数。所有这些使得网络边缘设备硬件大小的估计变得尤为困难。
- 在为特定设备优化模型的过程中，始终伴随着权衡。 这意味着模型在不同的设备中可能以不同的方式运行。
- 最后，网络边缘中的设备通常非常小。这就限制了所有AI推理设备的大小。

由此我们总结出以下关于网络边缘推理的几点重要要求：

**用于网络边缘AI推理的引擎必须：**

- **功耗低**
- **非常灵活**
- **拓展性强**
- **尺寸小**

莱迪思的sensAI能让你开发出完全具备以上四个特征的推理引擎。它包含了硬件平台、软IP、神经网络编译器、开发模块和开发资源，能够助您迅速开发理想中的设计。

## 推理引擎的选择

将推理引擎构建到网络边缘设备中涉及两个方面：开发承载模型运行的硬件平台以及开发模型本身。

理论上来说，模型可以在许多不同的架构上运行。但若要在网络边缘，尤其是在实时在线的应用中运行模型，选择就变少了，因为要考虑到之前提到的功耗、灵活性和扩展性等要求。

- **MCU -** 设计AI模型的最常见做法就是使用处理器，可能是GPU或者DSP，也有可能是微控制器。但是网络边缘设备上的处理器可能就连实现简单的模型也无法处理。这样的设备可能只有低端的微控制器（MCU）。而使用较大的处理器可能会违反设备的功耗和成本要求，因此对于此类设备而言，AI似乎难以实现。

  这正是低功耗FPGA发挥作用的地方。与增强处理器来处理算法的方式不同，莱迪思的ECP5或UltraPlus FPGA可以作为MCU的协处理器，处理MCU无法解决的复杂任务之余，将功耗保持在要求范围内。由于这些莱迪思FPGA能够实现DSP，它们可以提供低端MCU不具备的计算能力。

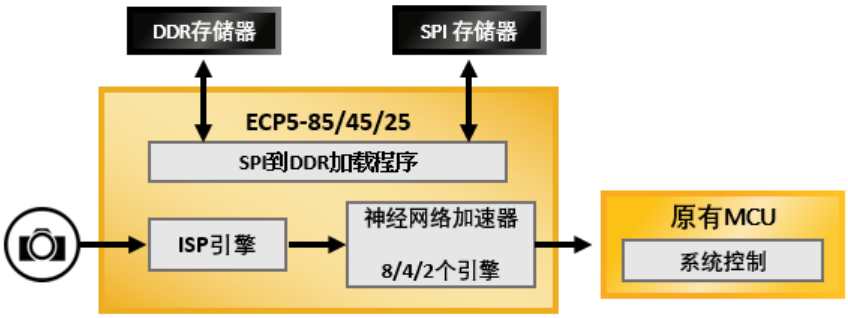



*图1：FPGA作为MCU的协处理器*

- **ASICS和ASSP -** 对于更为成熟、大批量销售的AI模型而言，采用ASIC或特定应用标准产品（ASSP）或许是可行之道。但是由于工作负载较大，它们在实时在线的应用中的功耗太大。

  在此情况下，Lattice FPGA可以充当协处理器，处理包括唤醒关键字的唤醒活动或粗略识别某些视频图像（如识别与人形相似的物体），然后才唤醒ASIC或ASSP，识别更多语音或者确定视频中的目标确实是一个人（或甚至可以识别特定的人）。

  FPGA处理实时在线的部分，这部分的功耗至关重要。然而并非所有的FPGA都能胜任这一角色，因为绝大多数FPGA功耗仍然太高，而莱迪思ECP5和UltraPlus FPGA则拥有必要的低功耗特性。

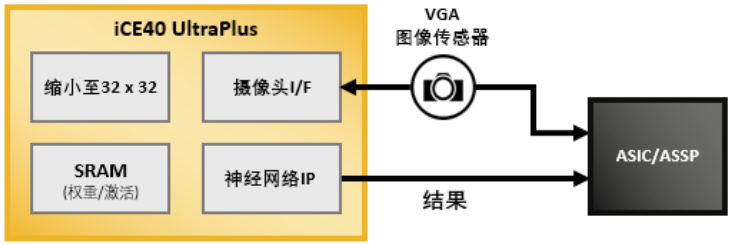



*图2 FPGA作为ASIC/ASSP的协处理器*

- **单独运行的FPGA AI引擎 -** 最后，低功耗FPGA可以作为单独运行的、完整的AI引擎。FPGA中的DSP在这里起了关键作用。即便网络边缘设备没有其他的计算资源，也可以在不超出功耗、成本或电路板尺寸预算的情况下添加AI功能。此外它们还拥有支持快速演进算法所需的灵活性和可扩展性。

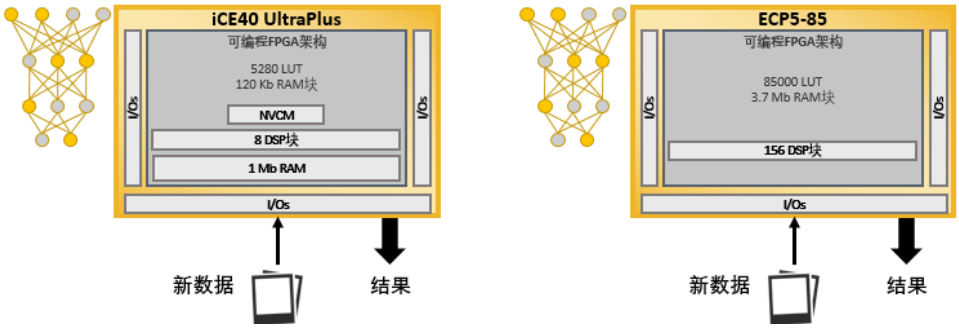



*图3 单独使用FPGA的整合解决方案*

## 在莱迪思FPGA中构建推理引擎

设计AI推理模型的硬件需要不断平衡所需资源数量与性能和功率要求。莱迪思的ECP5和UltraPlus产品系列能轻松实现这种平衡。

ECP5系列拥有三种不同规格的器件，能够运行一到八个推理引擎。它们集成的本地存储器从1 Mb到3.7 Mb不等。功耗最高仅为1 W，尺寸也只有100 $mm^2$。

相比之下，UltraPlus系列的功耗水平低至ECP5系列的千分之一，仅为1 mW。占用的电路板面积仅为5.5 $mm^2$，包括了最多8个乘法器和最多1 Mb的本地存储器。

莱迪思还提供可在这些器件上高效运行的CNN IP以及可用于ECP5系列的CNN加速器。

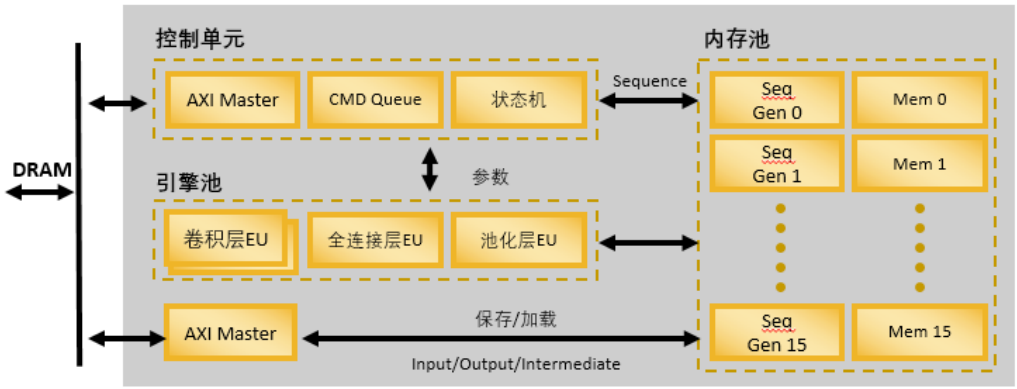



*图4 适用于ECP5系列的CNN加速器*

莱迪思还提供可用于UltraPlus系列的轻量化CNN加速器。

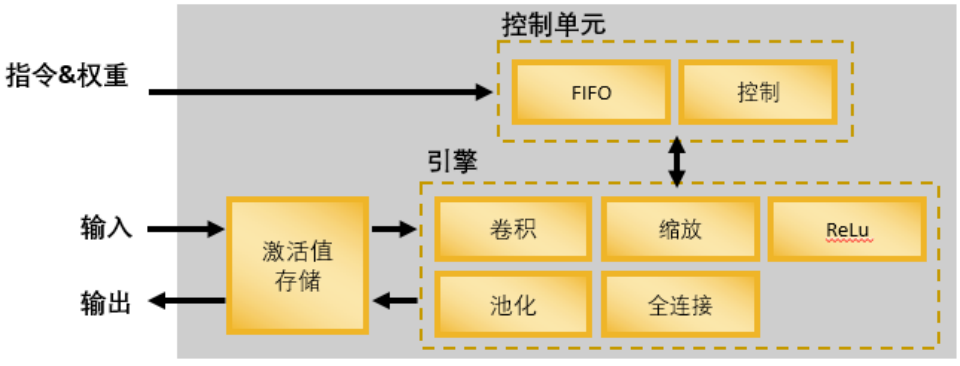
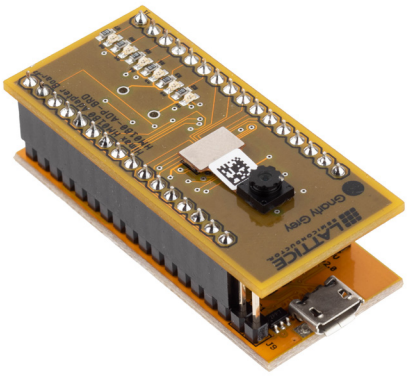


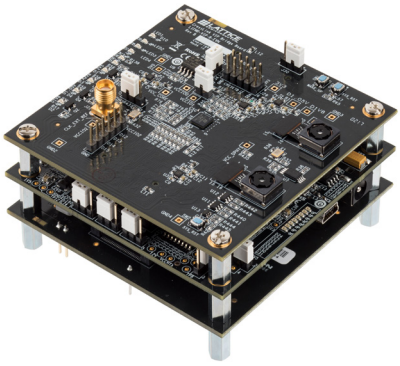

*图5 适用于UltraPlus系列的轻量化CNN加速器*

这里暂且不谈细节；重点在于您无须从头开始设计自己的AI引擎。您可以联系莱迪思获取关于这些IP的更多信息。

最后，您还可以在开发模块上运行并测试这些演示，两个模块分别对应这两种产品系列。Himax HM01B0 UPduino Shield采用了一片UltraPlus FPGA， 尺寸为22x50 $mm^2$。嵌入式视觉开发套件采用了一片ECP5 FPGA，尺寸为80x80 $mm^2$。



Himax HM01B0 UPduino Shield



嵌入式视觉开发套件

*图6 评估AI应用的开发模块*

有了FPGA、软IP和其他处理数据所需的硬件部分，就可以使用Lattice Diamond设计工具进行编译，从而生成位流，在每次上电后对目标设备中的FPGA进行配置。

## 在莱迪思FPGA上构建推理模型

创建推理模型与创建底层运行平台大不相同。它更抽象，涉及更多运算，且不涉及RTL设计。 这一过程主要有两个步骤：创建抽象模型，然后根据所选平台优化模型的实现。

模型训练在专门为此过程设计的框架中进行。最流行的两个框架是Caffe和TensorFlow，但不限于此。

CNN由很多层构成——卷积层，可能还会有池化层和全连接层——每一层都有由前一层的结果馈送的节点。每个结果都在每个节点处加权重，权重多少则由训练过程决定。

训练框架输出的权重通常是浮点数。这是权重最为精确的体现，然而大多数网络边缘设备不具备浮点运算功能。这时我们需要针对特定平台对这个抽象模型进行优化，这项工作由莱迪思的神经网络编译器负责。

编译器可以实现加载和查看从某个CNN框架下载的原始模型。您可以运行性能分析，这对模型优化最关键的方面——量化至关重要。

由于无法处理浮点数，因此需要将它们转换为整数。对浮点数四舍五入也就意味着精度会降低。问题是，什么样的整数精度才能满足您想要的精度？通常使用的最高精度为16位，但是权重和输入可以表示为较小的整数。莱迪思目前支持16、8和1位的设计实现。

1位的设计实际是在一位整数域中进行训练以保持精度。显然，更小的数据单元意味着性能更高、硬件尺寸更小以及功耗更低。但是，精度太低就无法准确地推断视野中的物体。

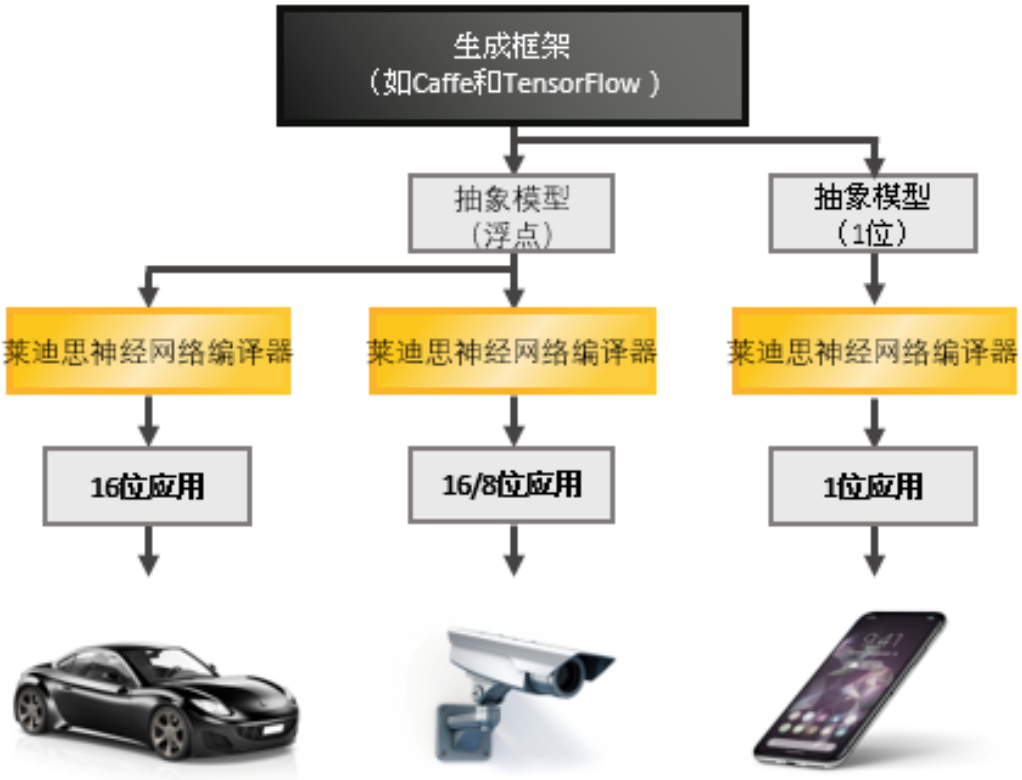



*图7 可以对单个模型进行优化以适用于不同的设备*

神经网络编译器能让您创建代表模型的指令流，然后可以模拟或直接测试这些指令，从而判断在性能、功耗和精度之间是否达到了适度的平衡。测试的标准通常是看一组测试图像（与训练图像不同）中正确处理的图像的百分比。

通常可以通过优化模型来优化运行，包括去掉一些节点以减少资源消耗，然后重新训练模型。这一设计环节可以微调精度，同时保证能在有限的资源下顺利运行。

## 两个检测实例

在以下两个不同的视觉案例中，我们将看到权衡是如何发挥作用的。第一个应用是人脸检测；第二个是人员侦测。我们将看到不同FPGA之间存在的资源差异如何影响到相对应的应用的性能和功耗。

两个示例的输入都源自同一个摄像头，两者都在相同的底层引擎架构中运行。在UltraPlus设计实例中，图像的尺寸缩小后通过8个乘法器进行处理，利用了内部存储器并使用了LED指示灯。

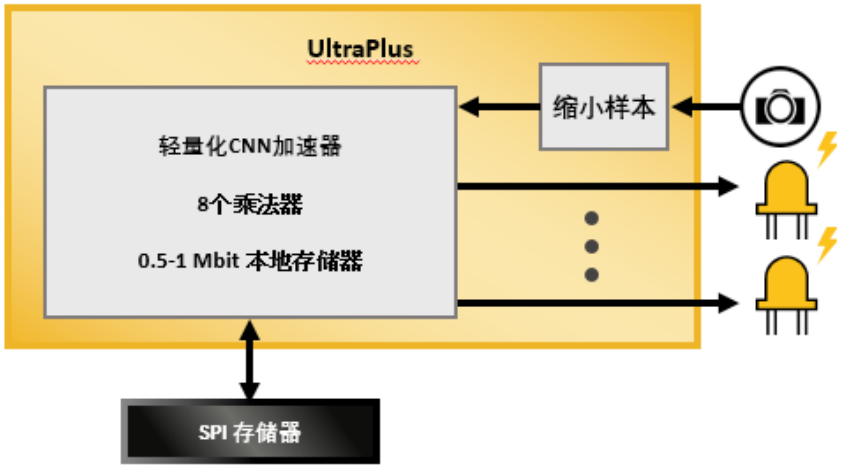



*图8 UltraPlus平台用于人脸检测和人员侦测应用*

ECP5系列资源更多，提供了一个计算能力更强的平台。摄像头捕捉的图像在发送到CNN之前在图像信号处理器（ISP）中进行预处理。处理结果与原始图像在标记引擎上比对，从而将文本或注释覆盖在原始图像上。

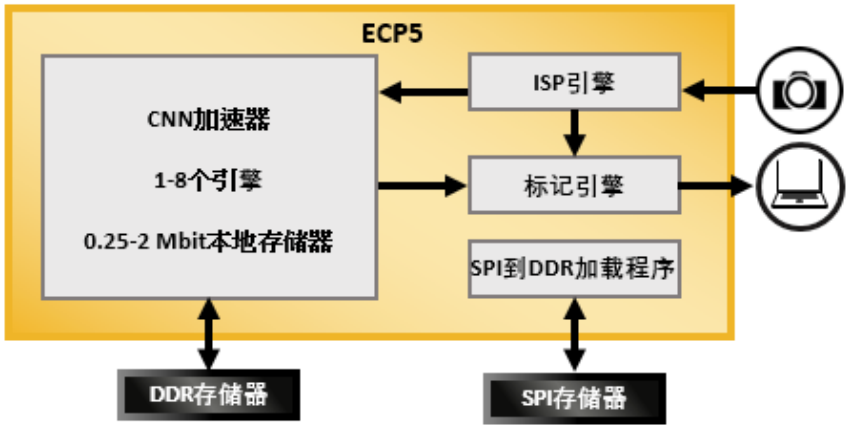



*图9 ECP5平台用于人脸检测和人员侦测应用*

我们可以使用一系列图表来衡量两种应用的性能、功耗和占用面积情况。对于每个应用，我们做了两组示例：一组输入较少，一组输入较多。

图7表示了人脸检测应用的结果。两组分别采用了32x32输入和90x90输入的情况。

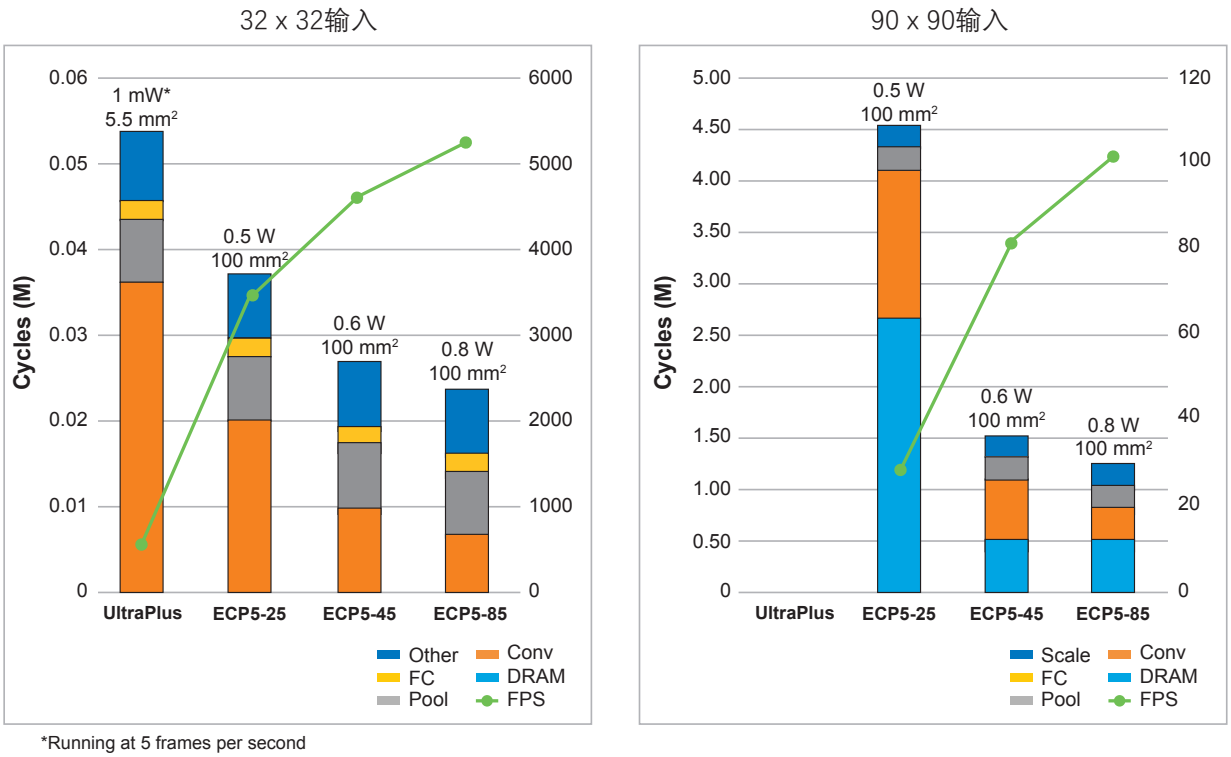


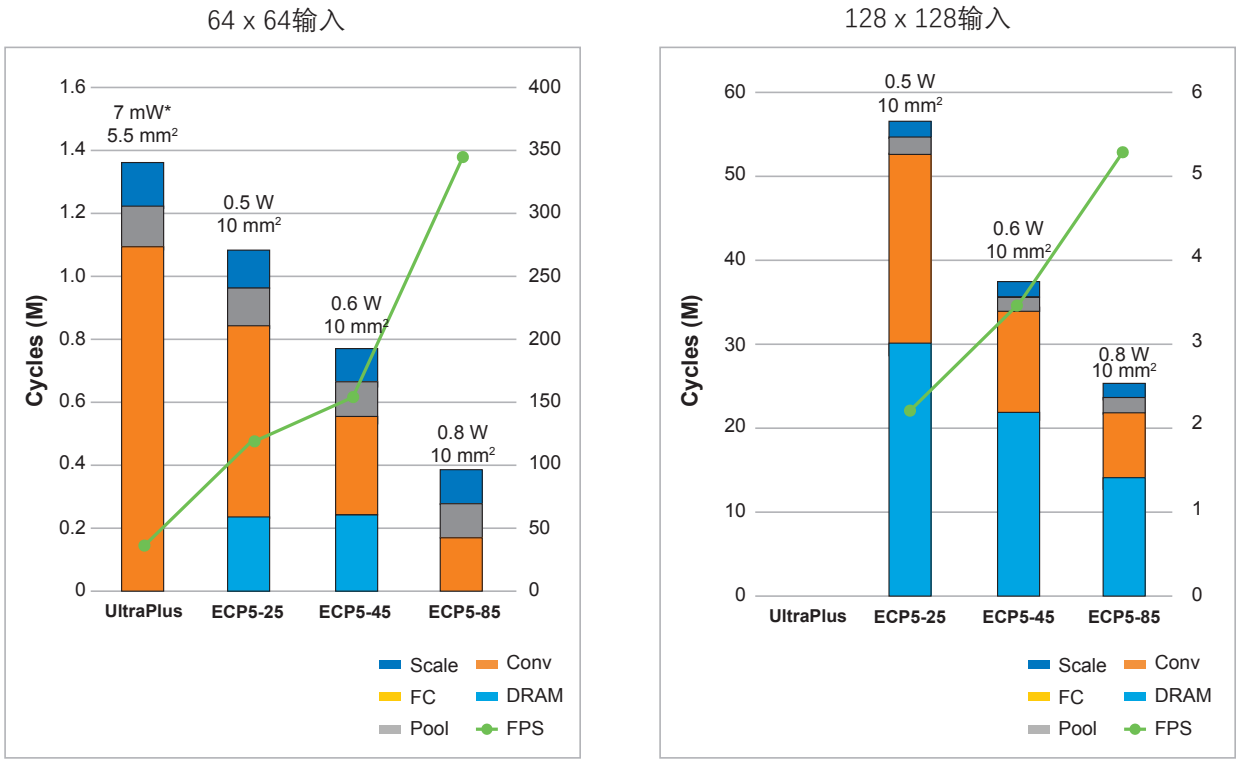

*图10 在UltraPlus和ECP5 FPGA上实现简单和复杂的人脸检测应用时的性能、功耗和占用面积*

左侧的轴代表处理一张图片需要的周期数量以及这些周期的分配情况。右侧的轴代表在各器件（绿线）上实现的每秒帧数（fps）。最后，每种情况下还标注了功耗和占用面积。

左侧的32x32输入示例中，橙色部分代表卷积层上运行的周期。在四个示例中，UltraPlus的乘法器数量最少；其他三片ECP5 FPGA的乘法器数量依次递增。随着乘法器数量的增加，卷积层所需的周期数减少。

90x90输入的示例位于右侧，得到的结果完全不同。在每个柱形图的底部有大面积的蓝色区域。这是由于设计更为复杂，使用了除器件内部存储空间以外的更多存储器。由于需要占用外部DRAM，性能就有所损失。需要注意的是，这种设计无法使用较小的UltraPlus器件。

人员侦测应用的情况类似。两组分别采用了64x64输入和128x128输入的情况。




*Running at 5 frames per second

*图11 在UltraPlus和ECP5 FPGA上实现简单和复杂的人脸检测应用时的性能、功耗和占用面积*

同样，较多的乘法器会减少卷积层的负担，而依赖DRAM则会影响性能。

图9总结了各类情形下的性能。它包括了对图像中最小可识别对象或特征的度量，用视野范围的百分比表示。在这里使用更多输入能够为较小的目标提供更多分辨率。

| | | 器件尺寸/功耗/性能 | | | |
|---|---|---|---|---|---|
| **网络** | **最小对象占比** | **UltraPlus 1-7 mW² 5.5 mm²** | **ECP5-25 0.5 W 100 mm²** | **ECP5-45 0.8 W 100 mm²** | **ECP5-85 0.8 W 100 mm²** |
| 人脸检测 32x32输入 | 50% | 465 | 3360 | 4511 | 5251 |
| 人脸检测 90x90输入 | 20% | -- | 28 | 82 | 101 |
| 人员侦测 64x64输入 | 20% | 18 | 115 | 161 | 338 |
| 人员侦测 128x128输入 | 10% | -- | 2.3 | 3.5 | 5.4 |

*图12 两个应用示例在四片FPGA上的性能总结*

## 小结

总之，使用莱迪思sensAI产品提供的资源，您就可以在莱迪思FPGA上轻松实现要求低功耗、具有灵活性和可扩展性的网络边缘AI推理设计。它可以提供成功部署AI算法所需的关键要素：

- 神经网络编译器
- 神经引擎软IP
- Diamond设计软件
- 开发板
- 参考设计

想要从莱迪思获得更多信息，请访问www.latticesemi.com，开始在您的设计中运用AI的力量。

**了解详情：**

www.latticesemi.com

**在线联系我们：**

www.latticesemi.com/contact
www.latticesemi.com/buy