基于 GoAI 的边缘设备全栈人工智能开发

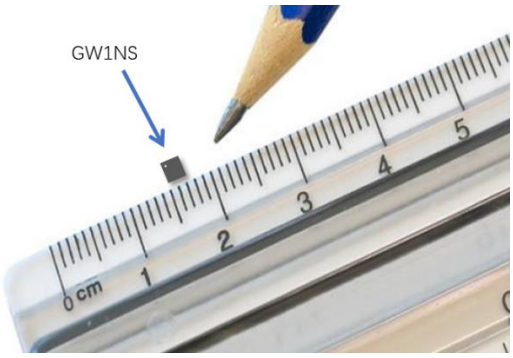
白皮书

WP951-1.1, 2020-09-20

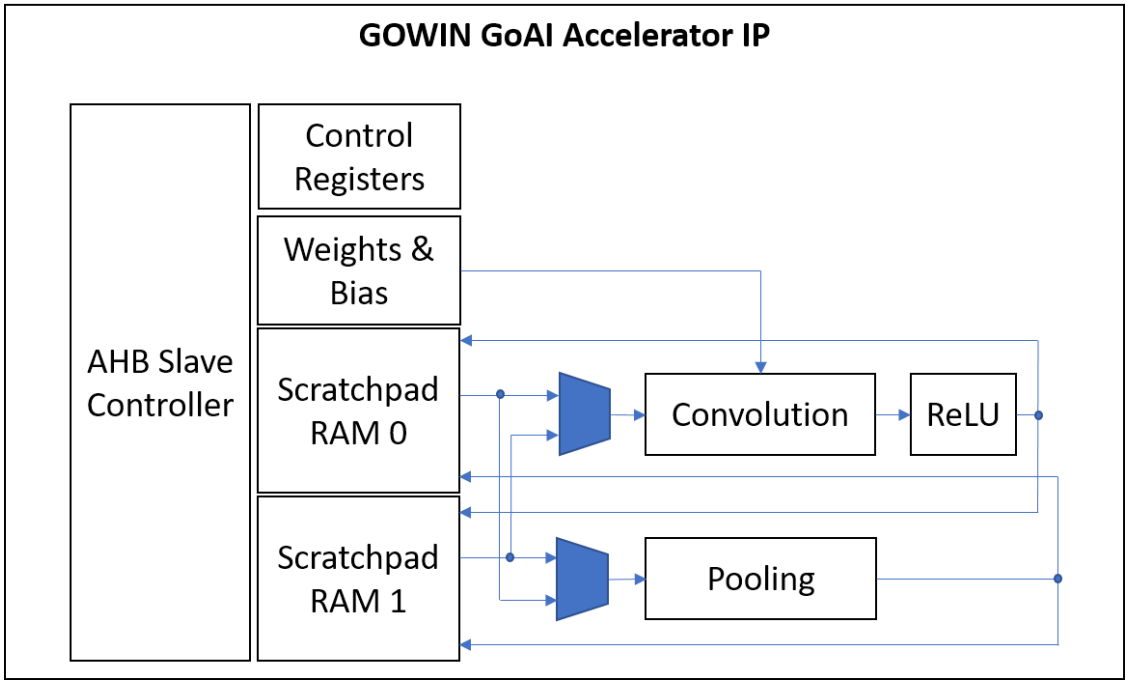
随着自动化、IoT、工业及消费类应用的不断出现，对边缘设备的要求和期待越来越高。边缘推理正逐渐成为边缘设备的通用功能，用于进行本地决策、减少延迟和降低连接节点成本。

随着时间的推移，当前主流解决方案常常难以满足客户对成本、功耗、方案面积以及适应和集成的灵活性等方面的需求。此外，现有的微处理器性能通常也不能满足神经网络繁重的计算需求。客户既要承受市场投放时间的压力同时又期待新技术的创新出现。

低密度 FPGA 可根据神经网络复杂度提供灵活的、可扩展解决方案解决客户对成本、功耗、大小等常见需求。高云半导体 FPGA 低功耗、高性能，提供不同晶圆级、从 1k 到 55k LUTs 不同密度的器件、QFN、 BGA 等封装类型可选、最小封装可达 3.24mm^2 。

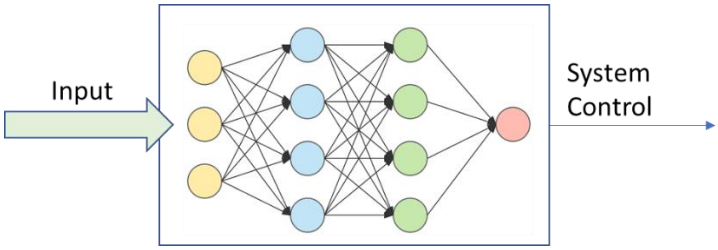


为提高性能、加速开发 AI 边缘解决方案上市时间，高云半导体已开发新的加速 IP 及基于高云 FPGA 硬件平台的整体解决方案“GoAI”。GoAI 将高云半导体 AI 加速 IP 集成到现有的机器学习框架中，与单独使用 Cortex-M 类微控制器相比，其性能提高了 78 倍以上。

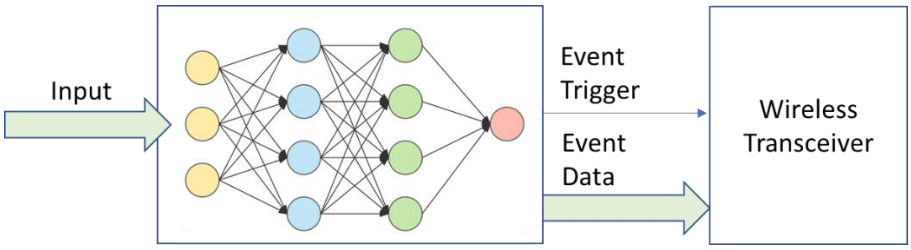


边缘 AI 在系统中的应用

边缘 AI 在系统中通常有两个应用。一个是对没有连接到因特网中的设备进行推理。这些系统通过机器学习检测输入信息并通过这些信息控制与之相连的系统的输出。



另外一个用处是在将数据发送到云端进行进一步处理前先进行预处理。这样做有诸多好处，比如把部分计算任务从云端卸载到边缘之后，减少网络传输进而节省功耗；仅将预处理后的数据发送到云端进而节约成本。



部署边缘 AI

今天的人工智能主要使用以卷积神经网络为中心的机器学习技术。这些网络本质上是许多过滤器或“神经元”的集合，这些神经元具有可学习的权重和偏置常量，可通过训练识别输入的关键特性。这些权重是通过训练计算出来的，通过在未经训练的网络上运行一组具有已知结果的输入，调整权重以识别输入。

训练卷积神经网络通常需要拥有大量的计算能力。但由于它只用于生成用于推断输入特性的权重，所以它通常不需要实时运行。一旦网络经过训练，就可以将权重加载到网络中，以检测输入特性。这种推理通常比训练所需的计算能力少得多。

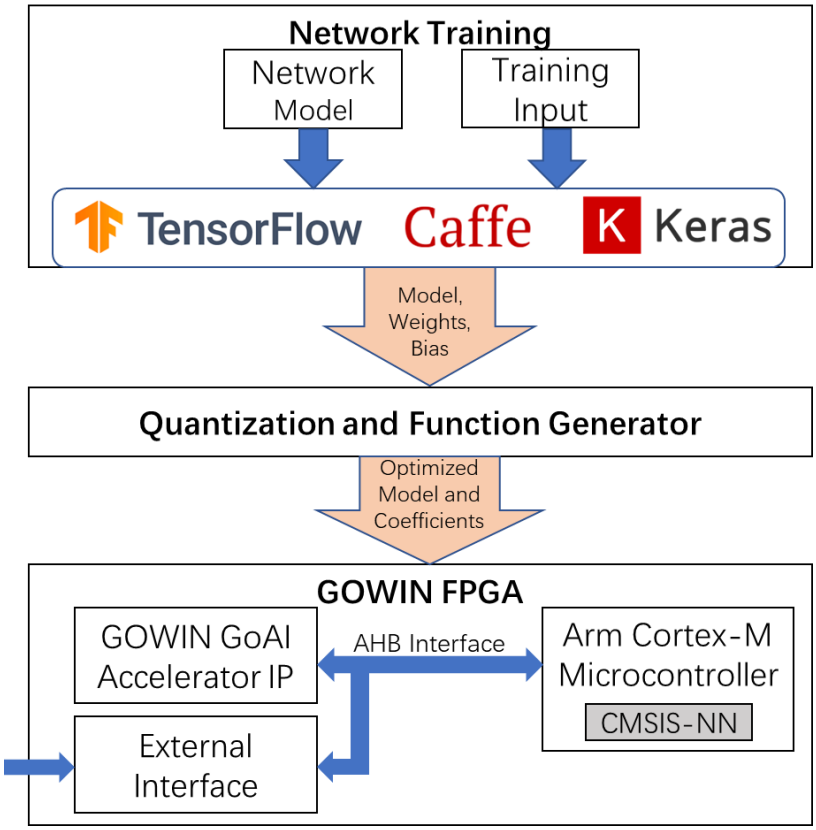
虽然推理所需的计算能力较少，但往往微处理器仍然达不到要求。这是因为每个处理器时钟周期，微控制器处理一条计算指令的频率常常低于 200mhz，即使是很小的机器学习网络，其性能也不足以进行实时检测。此外，许多 AI 相关的应用需要特殊的接口和数据缓存。例如，相机数据通常需要以帧的形式存储在 RAM 中，因为需要同时对图像中的多个像素进行过滤。

然而聚焦边缘的 FPGA 就可以很容易解决这些问题。FPGA 同时拥有数据并行和流水并行计算，这使其能够满足实时性能的同时可以在数 10 兆赫时更有效地运行系统。此外，FPGA 具有灵活的接口，便于连接相机、麦克风、生物传感器和其他输入等。FPGA 可配置内存也使其可以缓冲和保留中间数据或层数据。

虽然 FPGA 是实现边缘 AI 的一个很好的选择，但是需要一个强大的软件栈来简化开发

和部署。目前神经网络建模软件有多个选择，比如 Tensorflow，Caffe 和 Keras，但这些网络通常是使用浮点计算，用于软件的培训和测试，当试图在边缘部署高性价比解决方案时，会遇到各种各样的挑战。

因此，常用的部署工具，如用于微控制器的 Tensorflow Lite 和 Arm CMSIS-NN，使用优化过程将训练后的权重数据从浮点截断并量化为 8 位定点，使其更适用边缘硬件。然而提升（运算）性能依然重要，以致常常需要使用专门用于流水处理卷积和累积层数据的加速器。这些加速器可以使用 ASIC 或 FPGA 进行设计以进一步提高实时性能。



系统案例

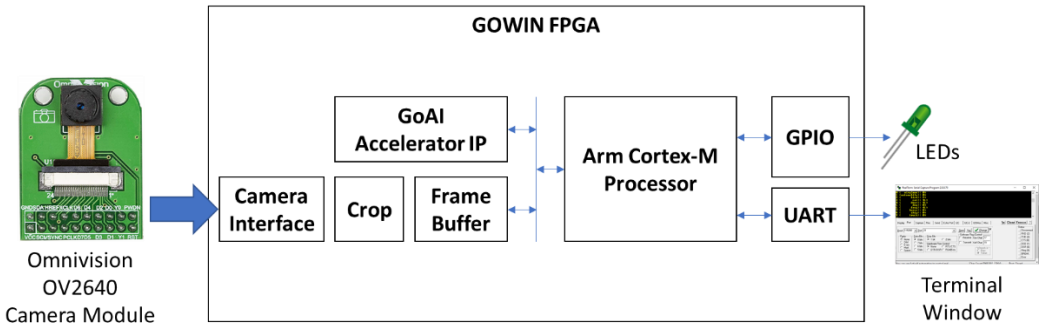
这里以 GoAI 平台对 CIFAR10 数据集进行图像检测为例，贯穿从模型训练到硬件设计的整个开发流程。将 GoAI 加速器与运行在 CMISIS-NN 中同一网络的 Arm Cortex-M 微控制器进行比较。这里的 CIFAR10 数据集是由 10 个分类对象组成的公共数据集，用于检测机器学习系统的各种性能属性。

首先，在 Caffe 中训练网络。本例中使用有不同数量过滤器的三个卷积层进行测试。对网络进行训练后，得到权值和偏置常量，并在 Caffe 中对训练后的网络进行测试，通过提供不同的输入来确认输出是否符合预期。

然后使用脚本工具截断权值和偏置常量并进行量化、编译网络、在 ARM Cortex-M1 和 M3 处理器上使用 CMISIS-NN 函数调用。

优化后的网络部署在 ARM Cortex-M1 处理器上，该处理器带有连接到 AHB 总线的摄像头接口和帧缓冲区。神经网络大约需要 10 秒的时间来处理一张相机图像。

GoAI 加速器连接到 AHB 总线，用于处理网络。Cortex-M1 仍先用于将图像数据传递给加速器，下载权重和偏置常量并对加速器进行配置。使用 GoAI 加速器，神经网络只需要大约 0.5 秒的时间进行处理，延时高低主要与通过 UART 发送的结果有关。



对 Arm Cortex-M3 处理器和加速器进行进一步的分析显示单独使用 Arm Cortex-M3 处理器与使用 GoAI 加速器相比，使用 GoAI 加速器性能提高了约 78 倍。

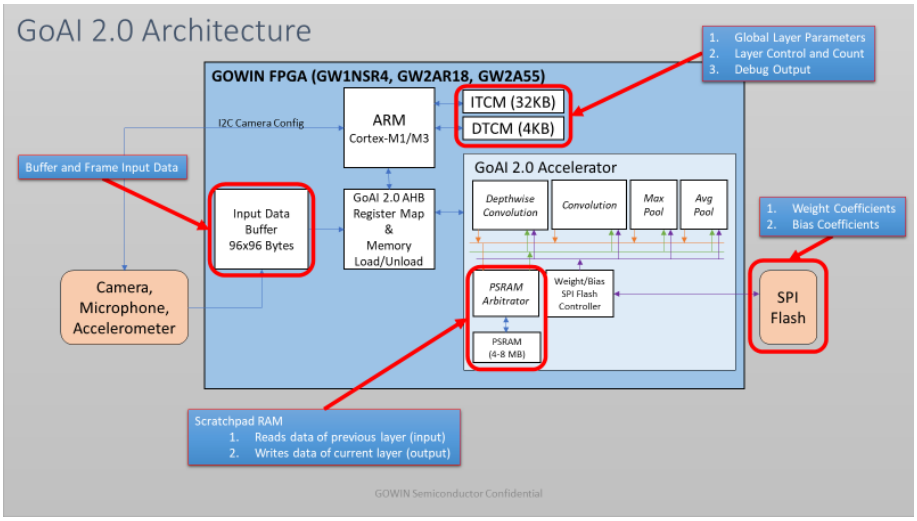
GoAI 2.0

GoAI 2.0 专注于：

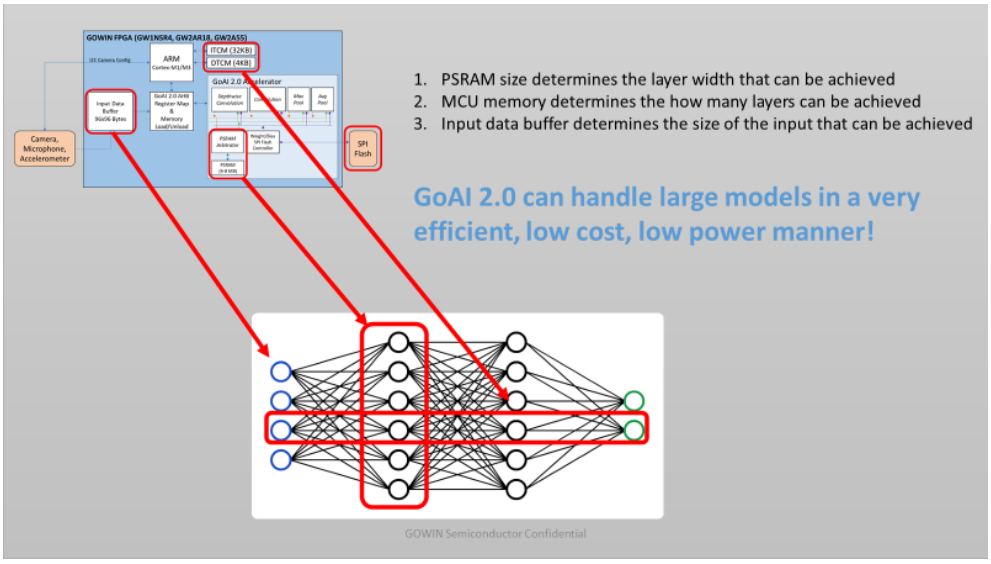
1. FPGA 加速器与 TensorFlow 和 TensorFlow Lite 的集成。
2. 面向采用 6x6mm QFN 封装内嵌 Cortex-M3 硬核处理器的 GOWIN GW1NSR-4C uSoC FPGA。
3. 软件编译和部署软件开发工具包（SDK）。
4. 结构灵活以支持层数多、层深度大的各种模型。

GoAI 2.0 平台使用标准 TensorFlow 开发环境，允许训练和测试任何模型。最终训练模型使用 TFLiteConverter 或 TocoConverter 将模型解析并量化为*.tflite flatbuffers 文件。然后使用 GoAI 2.0 SDK 解析该 flatbuffers 文件，提取模型系数、层参数以及模型函数。

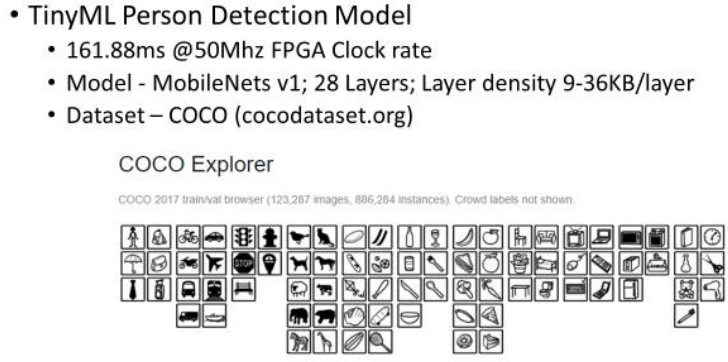
从 flatbuffers 文件中提取出所有必要信息后，GoAI 2.0 SDK 加载系数到外部 SPI 闪存中，加载 C 代码到 Cortex-M3 的嵌入式闪存中，加载比特流到 GW1NSR-4C 器件或其他支持 GoAI 的高云 FPGA 中。



GoAI 2.0 平台结构使其支持的层深尽可能大、卷积和池化层数尽可能多，因为 GW1NSR-4C 中集成了 PSRAM，可以用来存储层相关参数。GW1NSR-4C 包含 8MB 的 PSRAM，其中 4MB 为输入层缓冲，另外 4MB 为输出层缓冲。这意味着一个层输入和层输出的大小最大可达 4MB。Cortex-M3 硬核处理器内的 ITCM 嵌入式闪存大小为 32KB，其只需存储每层的控制环路和滤波器参数。外部 SPI 闪存则存储每层的权重和偏置系数，并可根据所需模型大小进行调整。



使用 Mobilenet v1.025 和 COCO 数据集对 GoAI 2.0 平台进行测试。Mobilenet 是一个具有 28 层的相当大的卷积神经网络。使用 GoAI 2.0，该模型的推理延迟仅为 162 毫秒。



结论

结合成本、功耗、尺寸及市场投放时间，在合理的预算内试图有效部署边缘 AI 会遇到各种各样的挑战。无论是对于连接或未连接的设备，边缘 AI 都变得越来越重要。边缘 AI 需要加速器及完整的软件开发流程用来进行实时处理及集成到通用机器学习模型开发软件中。GOWIN 的 GoAI 加速器和软件解决方案栈为解决性能和市场环境的限制提供了一个理想的解决方案。

技术支持与反馈

高云半导体提供全方位技术支持，在使用过程中如有任何疑问或建议，可直接与公司联系：

网址：www.gowinsemi.com.cn

E-mail：support@gowinsemi.com

Tel：00 86 0755 82620391

版本信息

日期	版本	说明
2019/09/16	1.0	初始版本。
2020/10/10	1.1	新增“GoAI 2.0”介绍。

免责声明

本文档并未授予任何知识产权的许可，并未以明示或暗示，或以禁止发言或其它方式授予任何知识产权许可。除高云半导体在其产品的销售条款和条件中声明的责任之外，高云半导体概不承担任何法律或非法律责任。高云半导体对高云半导体产品的销售和／或使用不作任何明示或暗示的担保，包括对产品的特定用途适用性、适销性或对任何专利权、版权或其它知识产权的侵权责任等，均不作担保。高云半导体对文档中包含的文字、图片及其它内容的准确性和完整性不承担任何法律或非法律责任，高云半导体保留修改文档中任何内容的权利，恕不另行通知。高云半导体不承诺对这些文档进行适时的更新。


GOWIN高云
智慧逻辑 定制未来