APKBUS

TensorFlow在物联网硬件端的加速
---Edge TPU

王玉成

APKBUS

1. 原理
2. 应用
3. 现场展示

TensorFlow Lite简介

- TensorFlow Lite是在移动和嵌入式设备上运行机器学习模型的官方解决方案。它支持Android，iOS和其他操作系统上的低延迟和小二进制大小的设备上机器学习推理。

</> 优势

- 性能： TF Lite速度快，没有明显的精度损失
- 低延迟：优化的浮点和定点CPU内核，操作融合等。
- 低容量： 含有依赖，量化，注册等功能。
- 扩展性： 适用于Android, ios, 以及其它的嵌入式物联网设备
- 加速 ：使用内部/外部的加速器进行加速
- 工装：转换，压缩，基准测试和功耗等

TensorFlow 特性

- TensorFlow Lite支持一系列核心运算符，包括量化和浮点运算，它们已针对移动平台进行了调整。它们结合了预融合激活和偏置，以进一步提高性能和量化精度。此外，TensorFlow Lite还支持在模型中使用自定义操作。
- TensorFlow Lite基于FlatBuffers定义了一种新的模型文件格式 。FlatBuffers是一个开源，高效的跨平台序列化库。它类似于ProtoBuffer，但主要区别在于FlatBuffers在访问数据之前不需要对辅助表示进行解析/解包步骤，通常与每个对象的内存分配相结合。此外，FlatBuffers的代码占用空间比ProtoBuffer小一个数量级。
- TensorFlow Lite拥有一个新的移动优化解释器，其主要目标是保持应用程序的精简和快速。解释器使用静态图形排序和自定义（动态性较小）内存分配器来确保最小的负载，初始化和执行延迟
- 尺寸较小：当所有支持的操作符链接时，TensorFlow Lite小于300KB，当仅使用支持InceptionV3和Mobilenet所需的操作符时，小于200KB。 。
- TensorFlow Lite提供了一个利用硬件加速的接口（如果在设备上可用）。它通过Android神经网络API实现，可在Android 8.1（API级别27）及更高版本上使用。

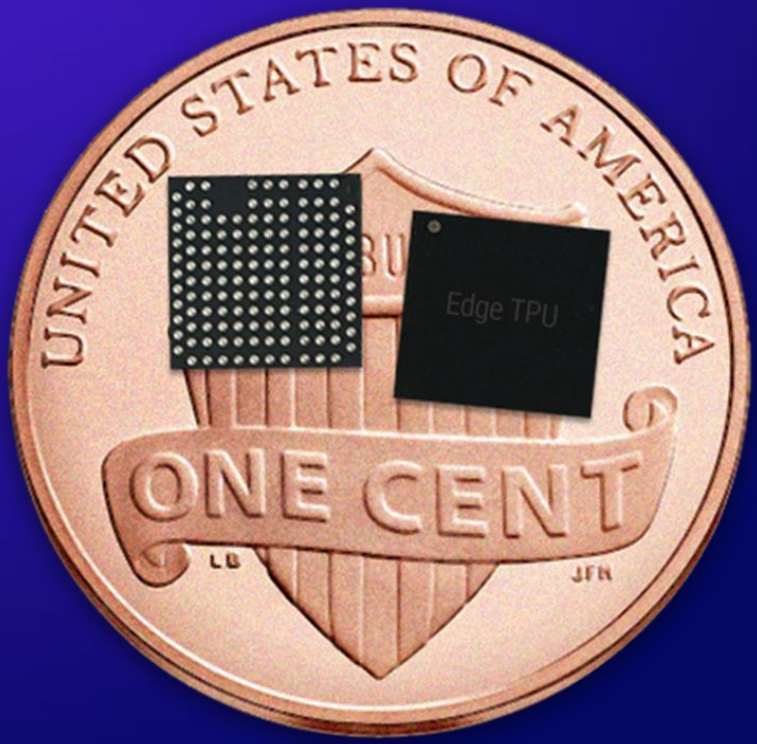
Google内部相关产品





</> Edge TPU简介

- Edge TPU是谷歌专用的ASIC芯片，专为在边缘运行TensorFlow Lite ML模型而设计。在设计Edge TPU时，我们专注于优化“每瓦性能”和“每美元性能”。Edge TPU旨在补充我们的Cloud TPU产品，因此您可以加速云中的ML培训，然后在边缘进行快速的ML推理。您的传感器不仅仅是数据采集器 - 它们可以做出本地，实时，智能的决策。



</> Edge TPU功能

	边缘（设备/节点，网关，服务器）	谷歌云
任务	ML推理	Ml训练和推理
软件，服务	云物联网边缘，Android事物	Cloud ML Engine，Kubernetes Engine，计算引擎，Cloud IoT Core
ML框架	TensorFlow Lite，NN API	TensorFlow，scikit-learn，XGBoost，Keras
硬件加速器	边缘TPU，GPU，CPU	云TPU，GPU和CPU

类型	推理加速器
性能示例	Edge TPU使用户能够以高效率的方式在高分辨率视频上以每秒30帧的速度同时执行多帧最先进的AI模型。
NUMERICS	Int8，Int16
IO接口	PCIe，USB

Cloud IoT Edge

- 提高运营可靠性

  由于您的边缘数据可以在本地存储、处理和获取智能，因此您可以在内部构建强大的物联网解决方案，而无需担心间歇性云连接。这对于需要实时处理的视频和音频应用程序或设备无法可靠连接到外部网络或Internet的情况很有用。

- 更快的实时预测

  通过运行设备上的机器学习模型，带有Edge TPU的Cloud IoT Edge可为关键物联网应用提供比通用物联网网关更快的预测 - 同时确保数据隐私和机密性。此外，Cloud IoT Edge和Edge TPU已经过广泛测试，可以本地运行开源参考模型，如MobileNet和Inception V3。

- 提高设备和数据的安全性

  Cloud IoT Edge可以在边缘设备上本地处理和分析图像，视频，手势，声学和运动，而无需将原始数据发送到云，然后等待响应。此本地处理可满足特定于行业的某些合规性需求，并降低数据隐私风险。Cloud IoT Edge使用JSON Web令牌对边缘设备进行身份验证，以便私钥永远不会离开设备。

APKBUS

1. 原理
2. 应用
3. 现场展示

</> Coral特性

私密：用户数据全部存储在本地

快速：本地加速AI，推理时间低于2ms

高效：低功耗，嵌入式应用

离线：连接受限的领域进行布署

</> 已发布产品

Dev Board简介

完美支持 TensorFlow Lite

从原型到产品快速量产

稳定运行操作系统

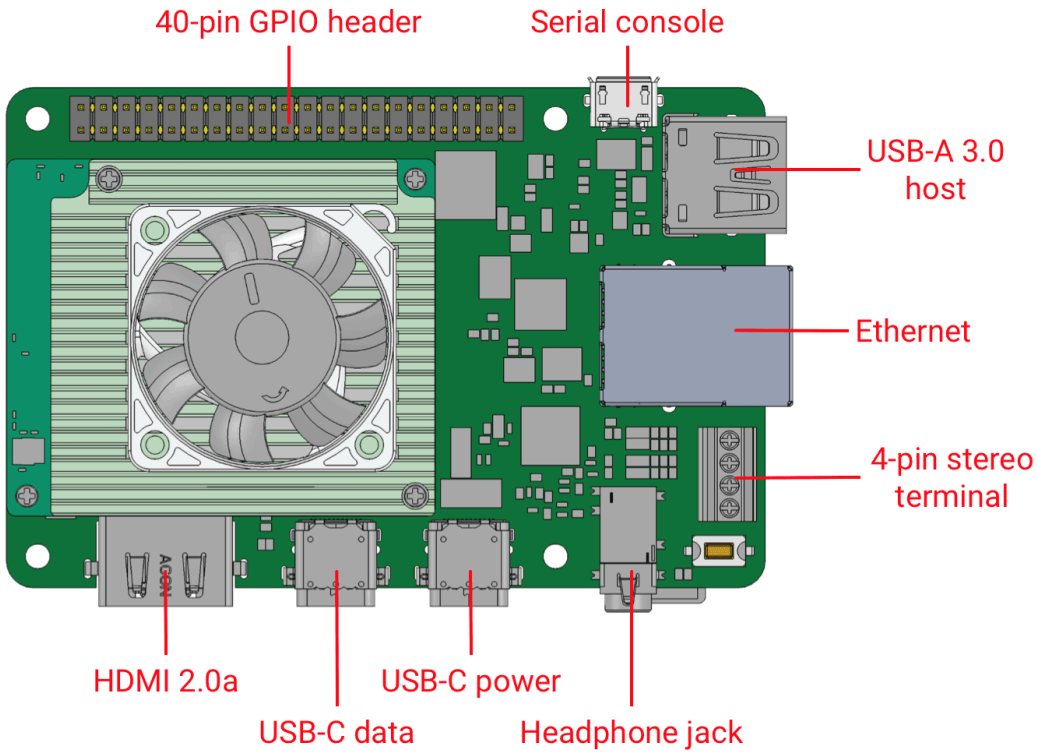
Dev Board性能-核心板

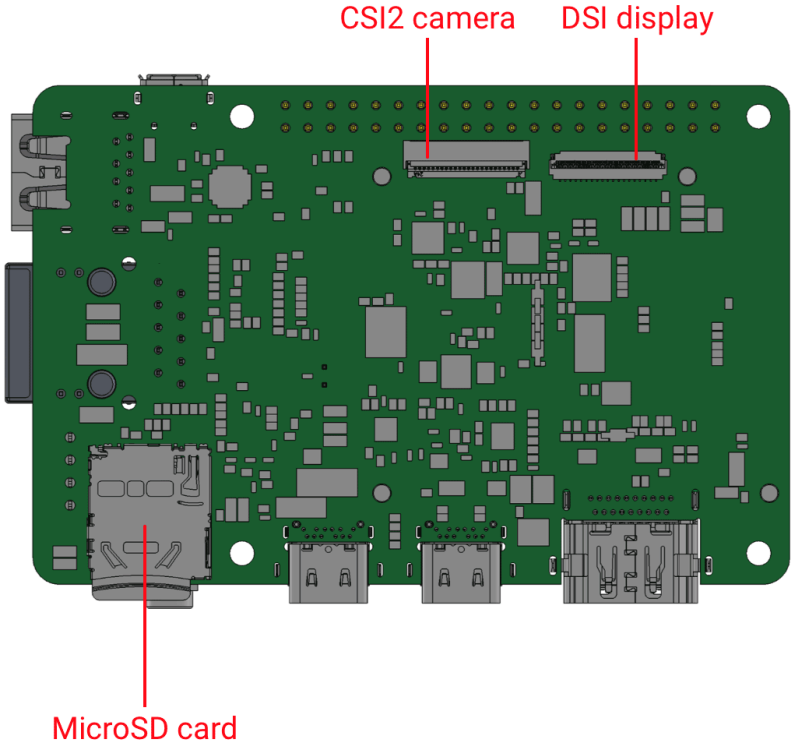
硬件	描述
CPU	NXP i.MX 8M SOC (quad Cortex-A53, Cortex-M4F)
GPU	Integrated GC7000 Lite Graphics
ML accelerator	Google Edge TPU coprocessor
RAM	1 GB LPDDR4
Flash memory	8 GB eMMC
Wireless	Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5GHz) Bluetooth 4.1
Dimensions	48mm x 40mm x 5mm

Dev Board性能-扩展板

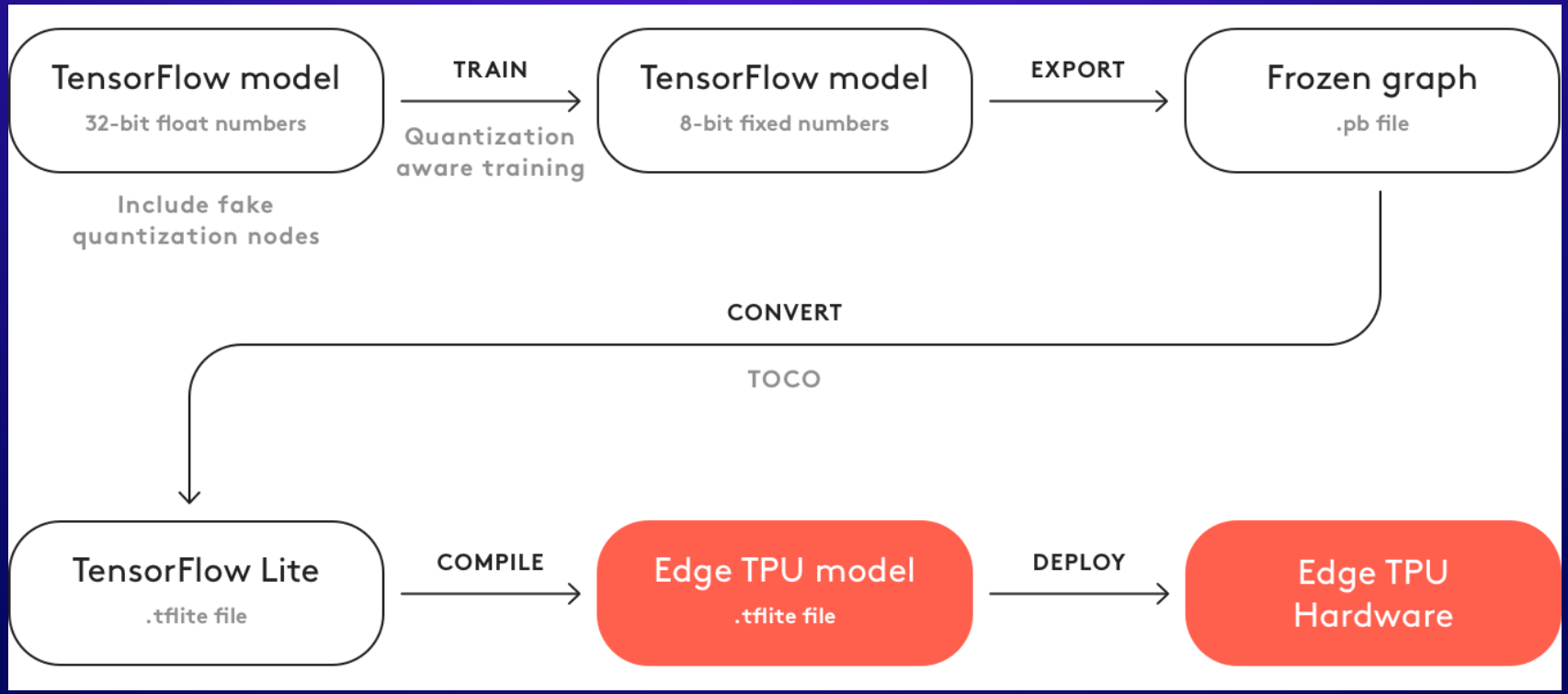
硬件	描述
Flash memory	MicroSD slot
USB	Type-C OTG Type-C power Type-A 3.0 host Micro-B serial console
LAN	Gigabit Ethernet port
Audio	3.5mm audio jack (CTIA compliant) Digital PDM microphone (x2) 2.54mm 4-pin terminal for stereo speakers
Video	HDMI 2.0a (full size) 39-pin FFC connector for MIPI-DSI display (4-lane) 24-pin FFC connector for MIPI-CSI2 camera (4-lane)
GPIO	3.3V power rail 40 - 255 ohms programmable impedance ~82 mA max current
Power	5V DC (USB Type-C)
Dimensions	88 mm x 60 mm x 24mm

Dev Board 接口

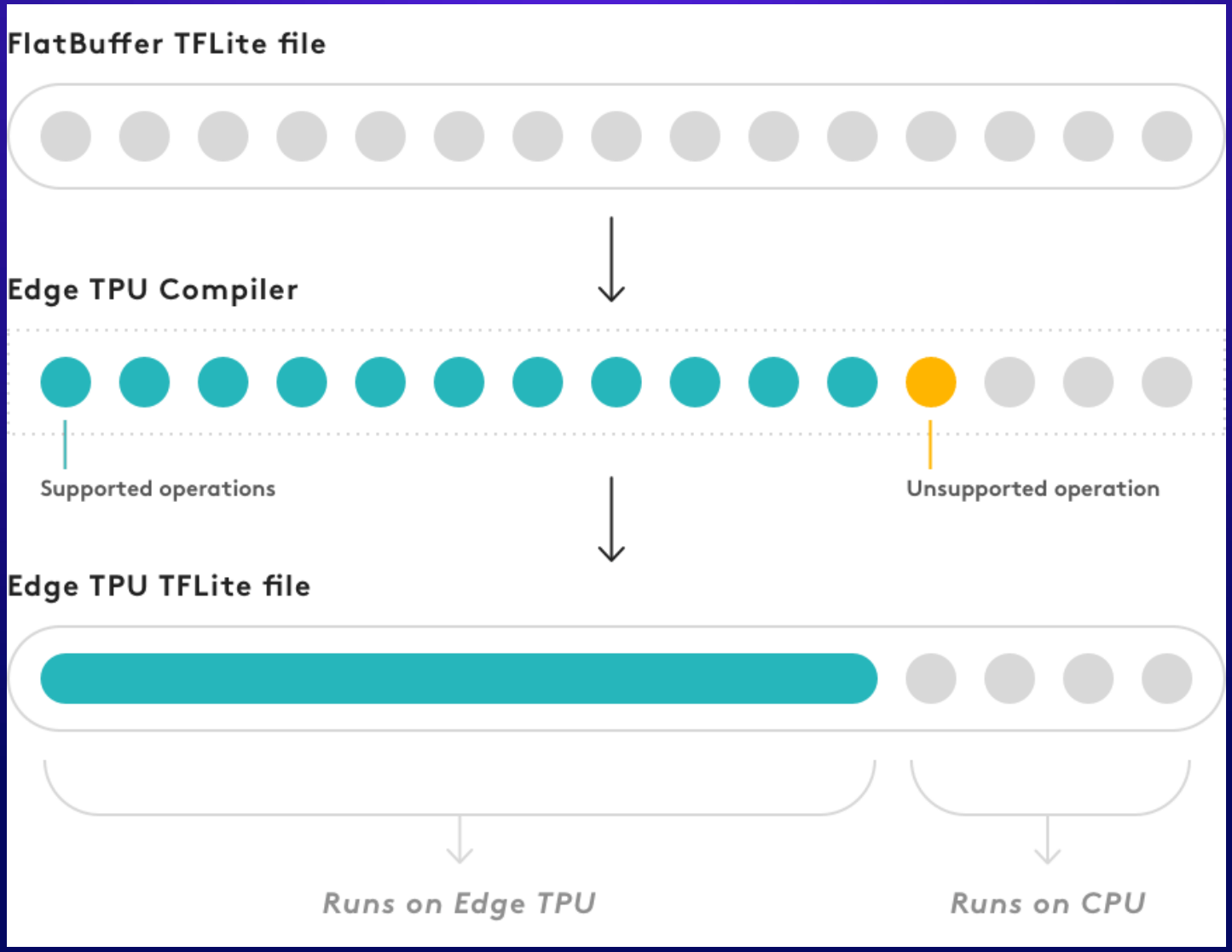

40-pin GPIO header
Serial console
USB-A 3.0 host
Ethernet
4-pin stereo terminal
HDMI 2.0a
USB-C data
USB-C power
Headphone jack


CSI2 camera
DSI display
MicroSD card

生成模型流程


TensorFlow model
32-bit float numbers
Include fake quantization nodes
TRAIN
Quantization aware training
TensorFlow model
8-bit fixed numbers
EXPORT
Frozen graph
.pb file
CONVERT
TOCO
TensorFlow Lite
.tflite file
COMPILE
Edge TPU model
.tflite file
DEPLOY
Edge TPU Hardware

Edge TPU模型运行概况



APKBUS

1. 原理
2. 应用
3. 现场展示

谢谢大家!

欢迎关注安卓巴士公众号

www.apkbus.com