基于移动端的 TensorFlow 相关技术进展

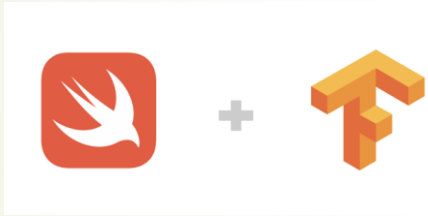
王玉成 IoT GDE
wfing123@gmail.com
2018.12.09

架构迎接未来变化
IAS 2018

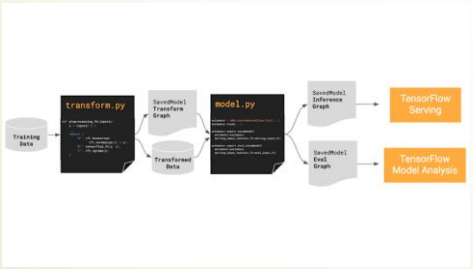
TensorFlow 现状

TensorFlow.js

TensorFlow Lite



TensorFlow Hub



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。

TF Lite的优势

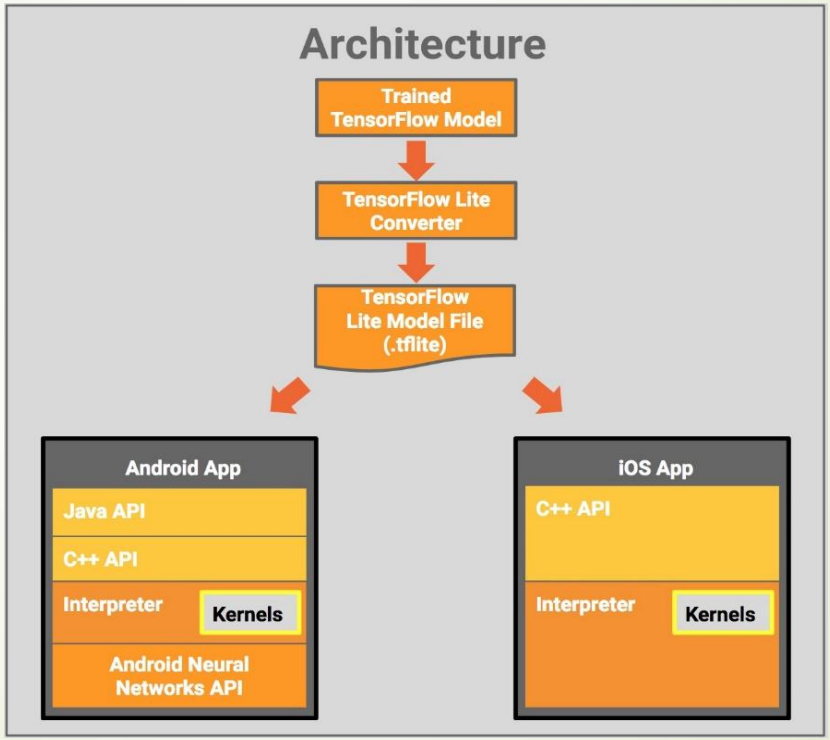
- 低延迟
- 低容量
- 易扩展

使用TensorFlow Lite的产品





TF LITE架构


Architecture
Trained TensorFlow Model
TensorFlow Lite Converter
TensorFlow Lite Model File (.tflite)
Android App
Java API
C++ API
Interpreter
Kernels
Android Neural Networks API
iOS App
C++ API
Interpreter
Kernels

TF LITE特性

- 量化、 浮点运算
- FlatBuffers
- 优化解释器
- 小容量
- 硬件加速接口

基于ANDROID性能基准测试

模型	设备	平均推理时间(std dev)
Mobilenet_1.0_224(float)	Pixel 2	166.5 ms (2.6 ms)
	Pixel xl	122.9 ms (1.8 ms)
Mobilenet_1.0_224 (quant)	Pixel 2	69.5 ms (0.9 ms)
	Pixel xl	78.9 ms (2.2 ms)
NASNet mobile	Pixel 2	273.8 ms (3.5 ms)
	Pixel xl	210.8 ms (4.2 ms)
SqueezeNet	Pixel 2	234.0 ms (2.1 ms)
	Pixel xl	158.0 ms (2.1 ms)
Inception_ResNet_V2	Pixel 2	2846.0 ms (15.0 ms)
	Pixel xl	1973.0 ms (15.0 ms)
Inception_V4	Pixel 2	3180.0 ms (11.7 ms)
	Pixel xl	2262.0 ms (21.0 ms)

选择模型

- MobileNets
- Inception-v3
- On Device Smart Reply

训练自己的模型

- 使用Tensorflow训练自定义模型

转换模型格式

- tf.GraphDef（.pb）
  - 表示TensorFlow训练或计算图的protobuf。它包含运算符，张量和变量定义。
- CheckPoint（.ckpt）
  - 来自TensorFlow图的转化变量。由于这不包含图形结构，因此无法自行解释。
- FrozenGraphDef
  - 它的子类GraphDef不包含变量。
- TensorFlow Lite模型（.tflite）
  - 一个序列化的 FlatBuffer，包含用于TensorFlow Lite解释器的TensorFlow Lite运算符和张量，类似于 FrozenGraphDef。

软件开发-生成模型

- V < 1.7

  freeze_graph --input_graph=/tmp/mobilenet_v1_224.pb

  …

  toco --input_file..

  ….

- V >= 1.7

  tflite_convert

  …

软件开发-使用模型

- import tensorflow as tf
- img = tf.placeholder(name="img", dtype=tf.float32, shape=(1, 64, 64, 3))
- val = img + tf.constant([1., 2., 3.]) + tf.constant([1., 4., 4.])
- out = tf.identity(val, name="out")
- with tf.Session() as sess:
- tflite_model = tf.contrib.lite.toco_convert(sess.graph_def, [img], [out])
- open("converteds_model.tflite", "wb").write(tflite_model)

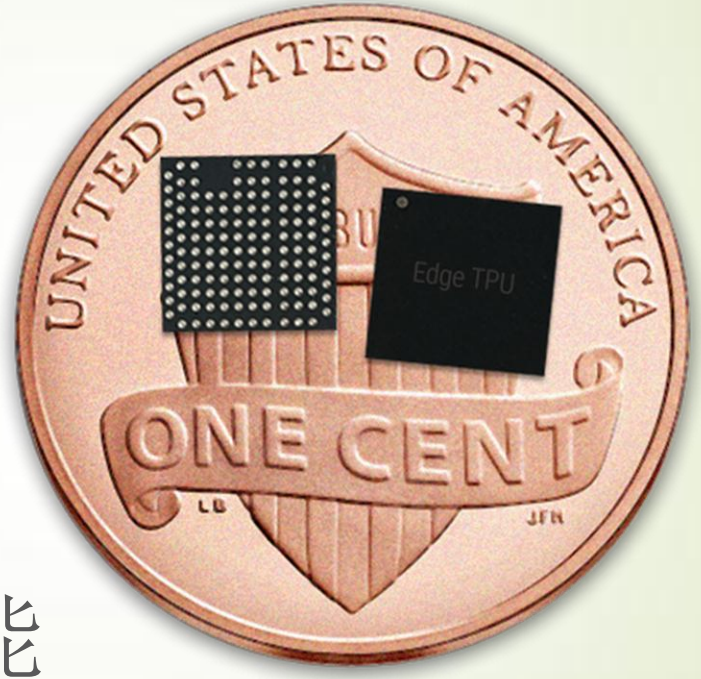
移动端硬件最新进展

- 2018.7.24~26日，Google Cloud Next'18上公布了Edge TPU
- 2025年，数据总量将超过40万亿千兆字节。
- 智能的实时决策,即所谓的“边缘”。

两个产品

- Edge TPU，一种新的硬件芯片，
- Cloud IoT Edge，一种将Google Cloud强大的AI功能扩展到网关和连接设备的软件堆栈。

EDGE TPU



- 每瓦性能
- 每美元性能
- 本地实时智能决策

EDGE TPU特性

- 端到端的AI基础设施
- 小尺寸高性能
- AI硬件、软件和算法的协同设计
- 广泛的应用

EDGE TPU功能

	边缘（设备/节点，网关，服务器）	谷歌云
任务	ML推理	MI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

- Cloud 训练 Edge TPU执行
- 边缘设备连接到云
- 基于TensorFlow Lite的Edge ML

CLOUD IOT EDGE

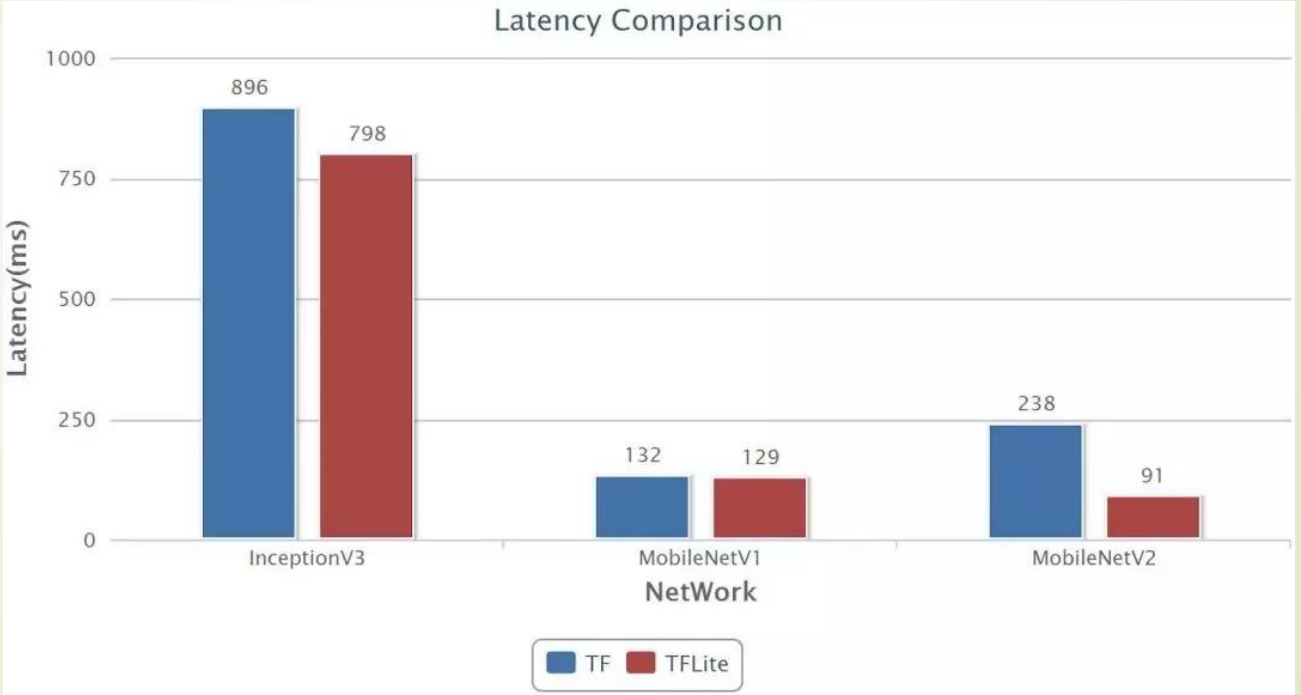
- 提高运营可靠性
- 更快的实时预测
- 提高设备和数据的安全性

TF Lite落地场景-中非木薯

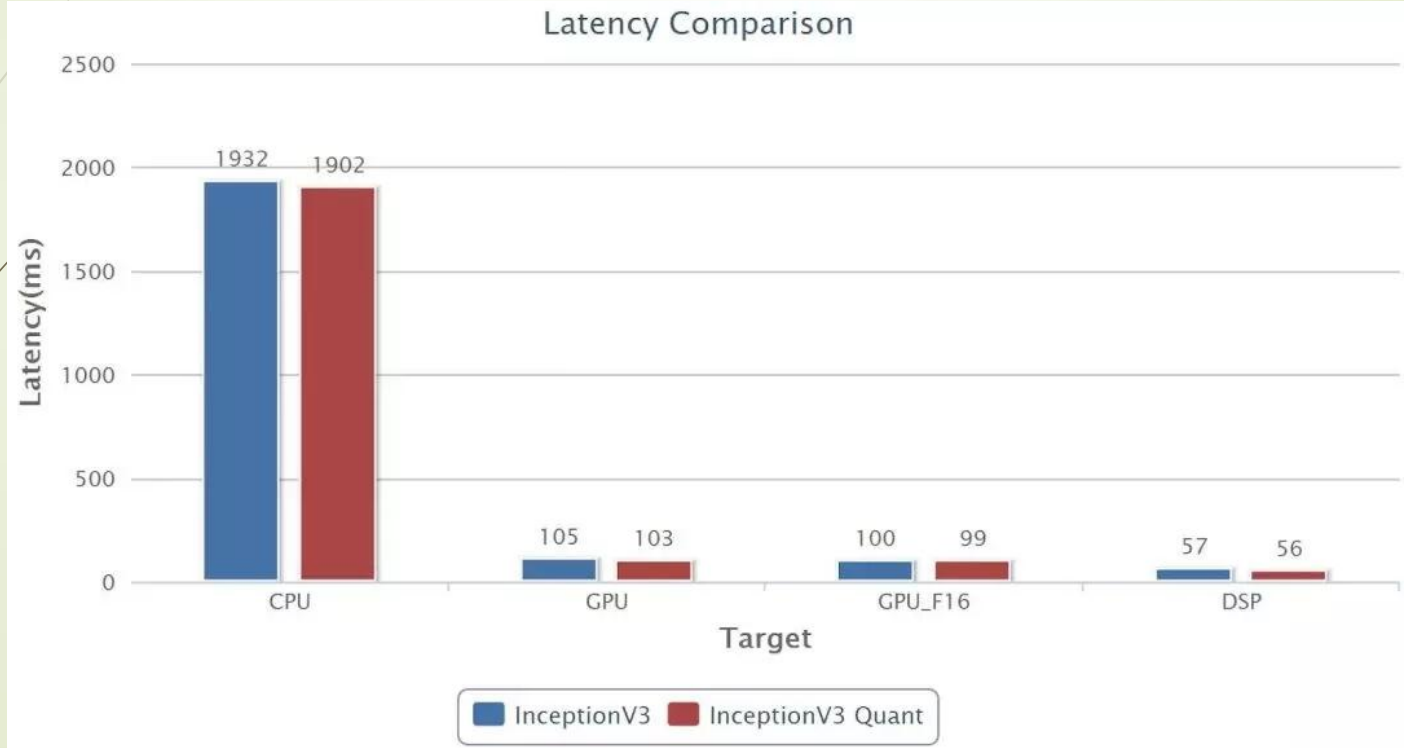
TensorFlow Lite + 有道翻译王 2.0 Pro

OCR（光学字符识别）
NMT（神经机器翻译）
ASR（自动语音识别）
TTS（语音合成）

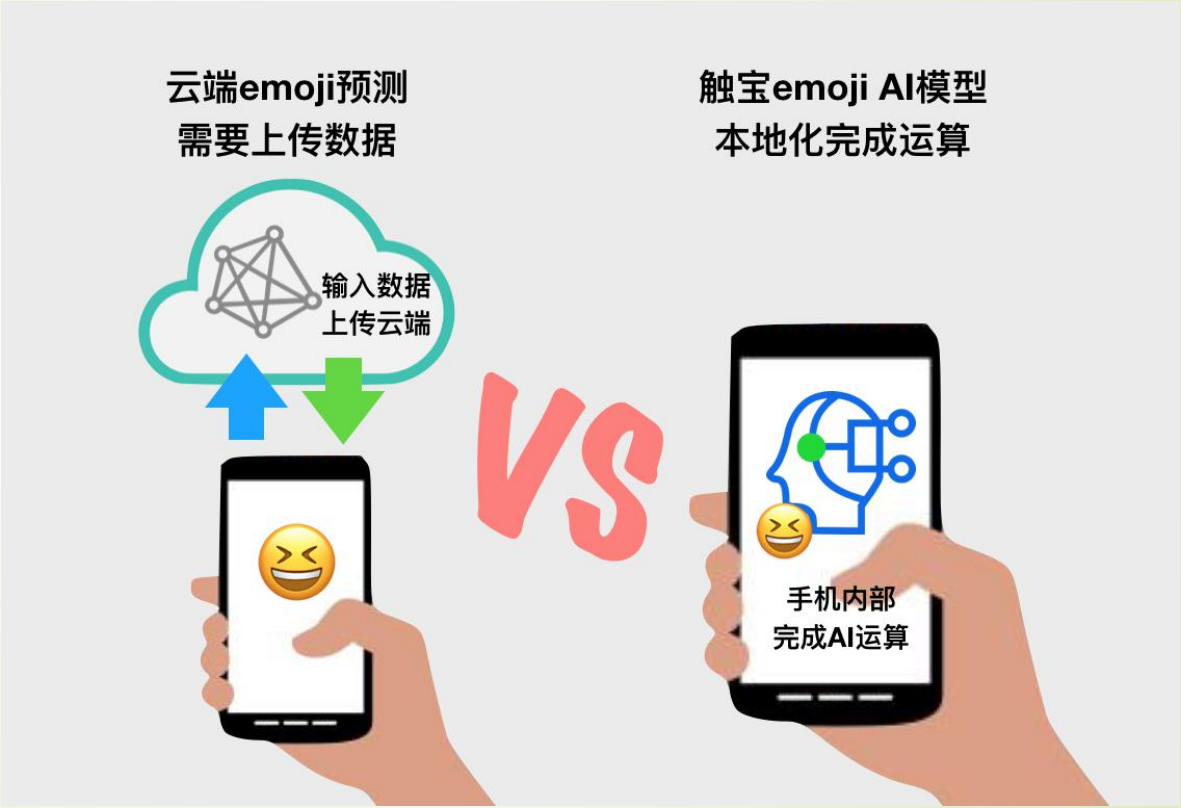
TensorFlowLite + 有道翻译王 2.0 Pro



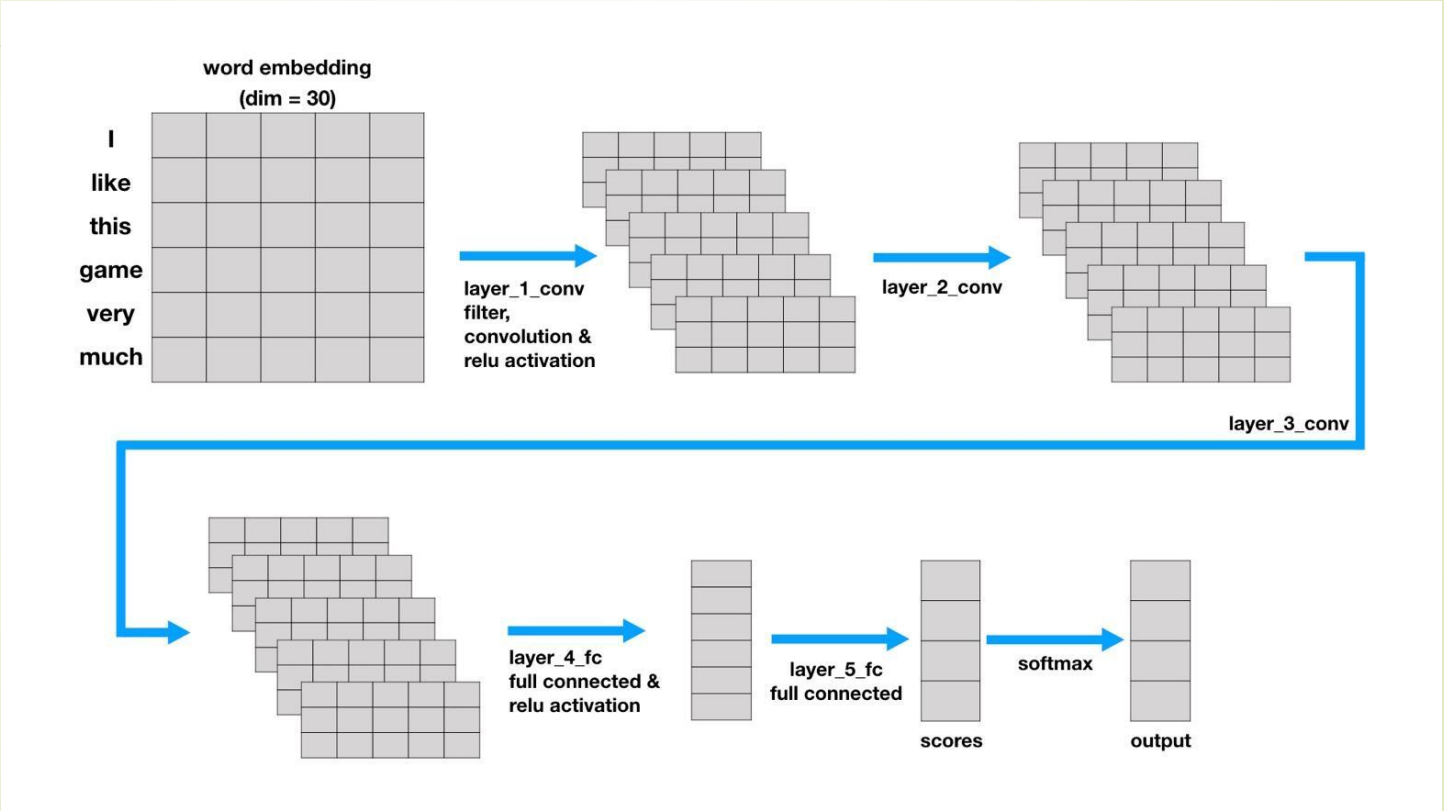
TensorFlowLite + 有道翻译王 2.0 Pro



TF Lite落地场景-触宝



TF Lite落地场景-触宝



TF Lite落地场景-触宝

Hope you have a great day

- 基于关键词 “great day ”, 预测结果: 🎊, ❤️
- 基于机器学习的AI模型, 预测结果: 🙏, ❤️

Any thing I can do, honey

- 基于关键词 “honey”, 预测结果: 🍯, ❤️
- 基于机器学习的AI模型, 预测结果: 🤔, 🍯

Thank you

Q&A