KOMINFO digitalent

PRO Professional Academy

Digital Talent Scholarship 2022

Device Based model in TF Lite

Lead a sprint through the Machine Learning Track

KOMINFO digitalent

Agenda

- Training own model
- Running TF model in Android App

KOMINFO digitalent

Objektif Pembelajaran

- Mengerti cara kerja TensorFlow Lite
- Memahami apa itu model quantization
- Dapat menjelaskan cara kerja GPU delegates
- Mengenal Optimization techniques bagi TensorFlow Lite

KOMINFO digitalent

Are your students ML-ready?

KOMINFO digitalent

Recap

PART 1

Machine Learning Models in Mobile and Embedded Systems

KOMINFO digitalent

Features

- Lightweight
- Low-latency
- Privacy
- Improved Power consumption
- Efficient model format
- Pre-trained models

Components in TensorFlow Lite

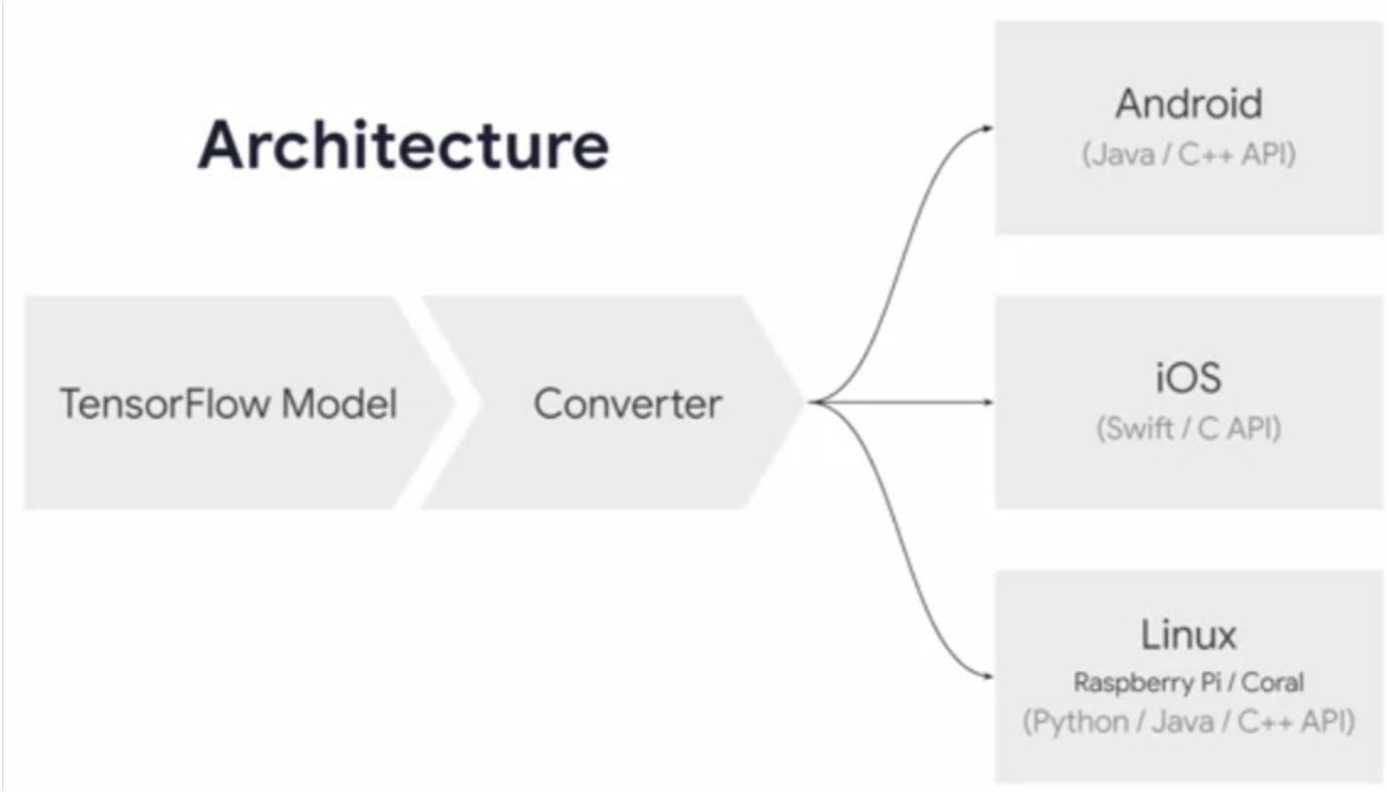
Converter (to TensorFlow Lite format)

- Mengubah Model Tensorflow menjadi bentuk yang lebih efisien dibaca oleh interpreter
- Memperkenalkan optimization untuk meningkatkan kinerja model dan/atau untuk mengurangi size model

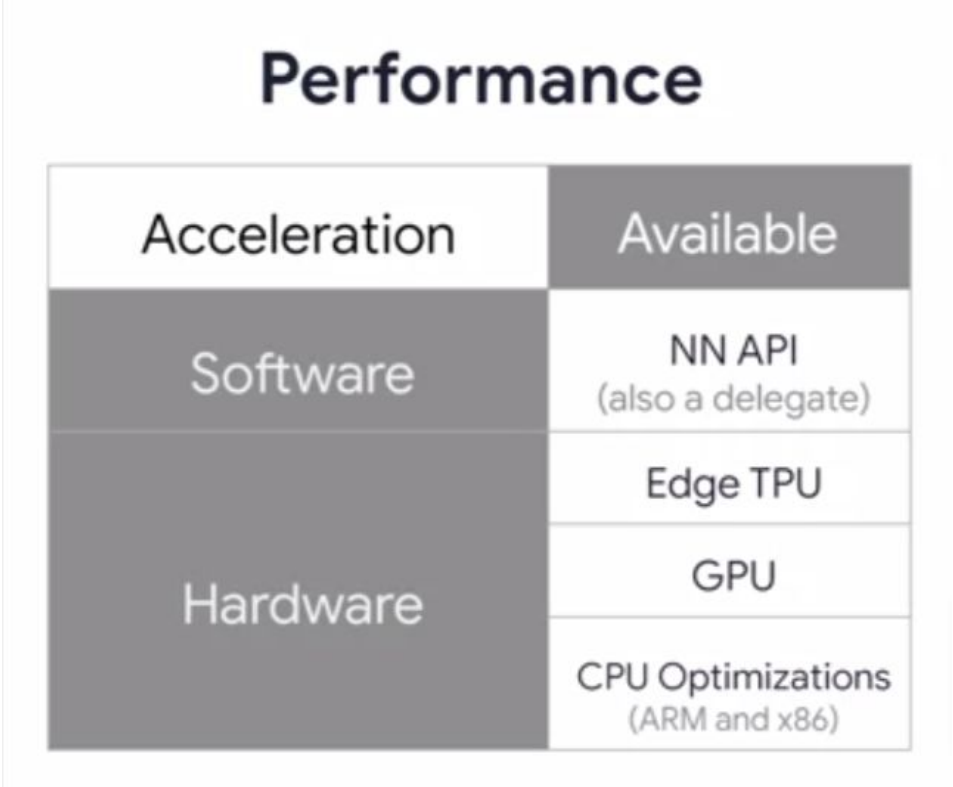
Interpreter (Core)

- Support multiple platforms (Android, iOS, embedded Linux, microcontrollers)
- Platform APIs for accelerated inference

Architecture



Performance



NN API

Android Neural Networks API (NNAPI) adalah Android C API yang dirancang untuk menjalankan operasi komputasi intensif bagi machine learning di perangkat Android. NNAPI dirancang untuk menyediakan lapisan dasar fungsionalitas bagi framework machine learning dengan level lebih tinggi, seperti TensorFlow Lite dan Caffe2, yang membuat dan melatih jaringan neural. API ini tersedia di semua perangkat yang menjalankan Android 8.1 (API level 27) atau yang lebih baru.

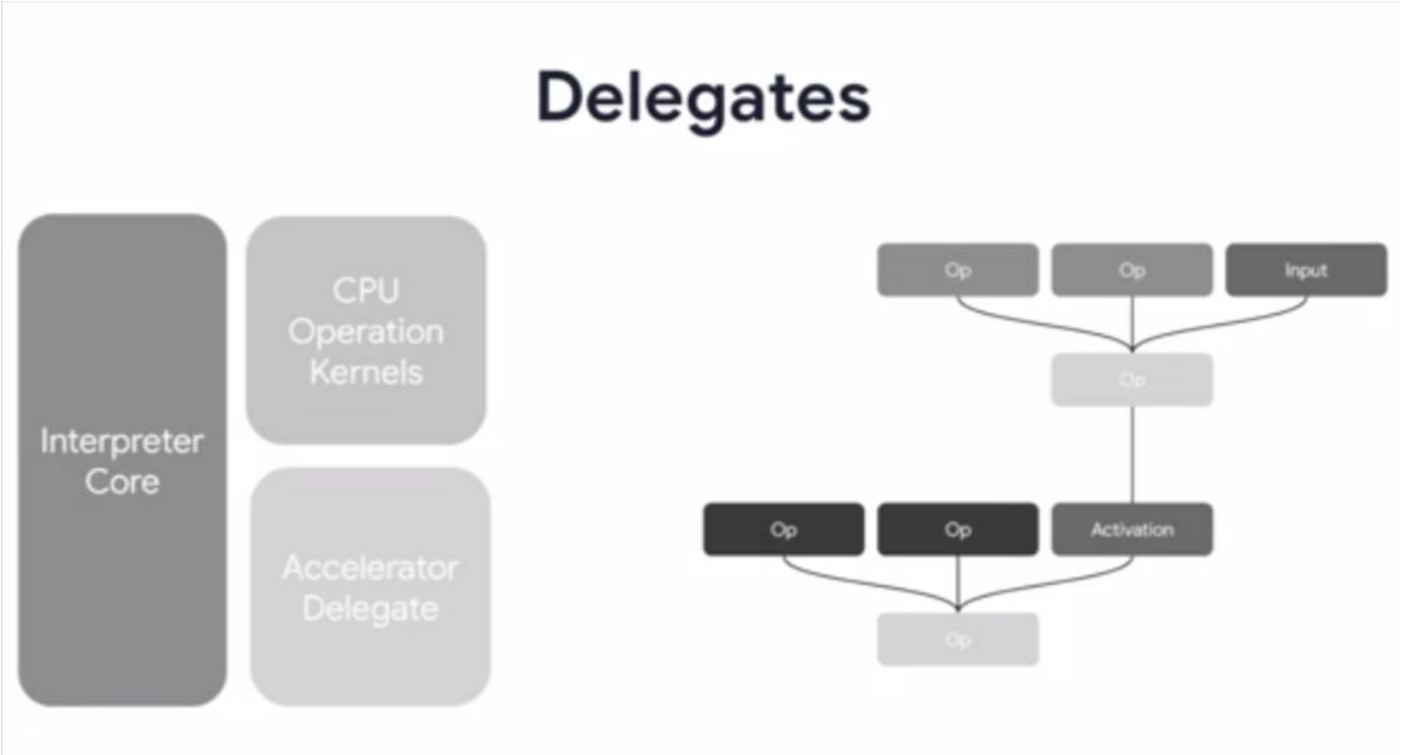
KOMINFO digitalent

[TensorFlow Lite, Experimental GPU Delegate - YouTube]

Delegasi

Delegasi memungkinkan akselerasi perangkat keras model TensorFlow Lite dengan memanfaatkan akselerator di perangkat seperti GPU dan Digital Signal Processor (DSP).

Delegates


Delegates
Interpreter Core
CPU Operation Kernels
Accelerator Delegate
Op
Op
Input
Op
Op
Op
Activation
Op

Techniques

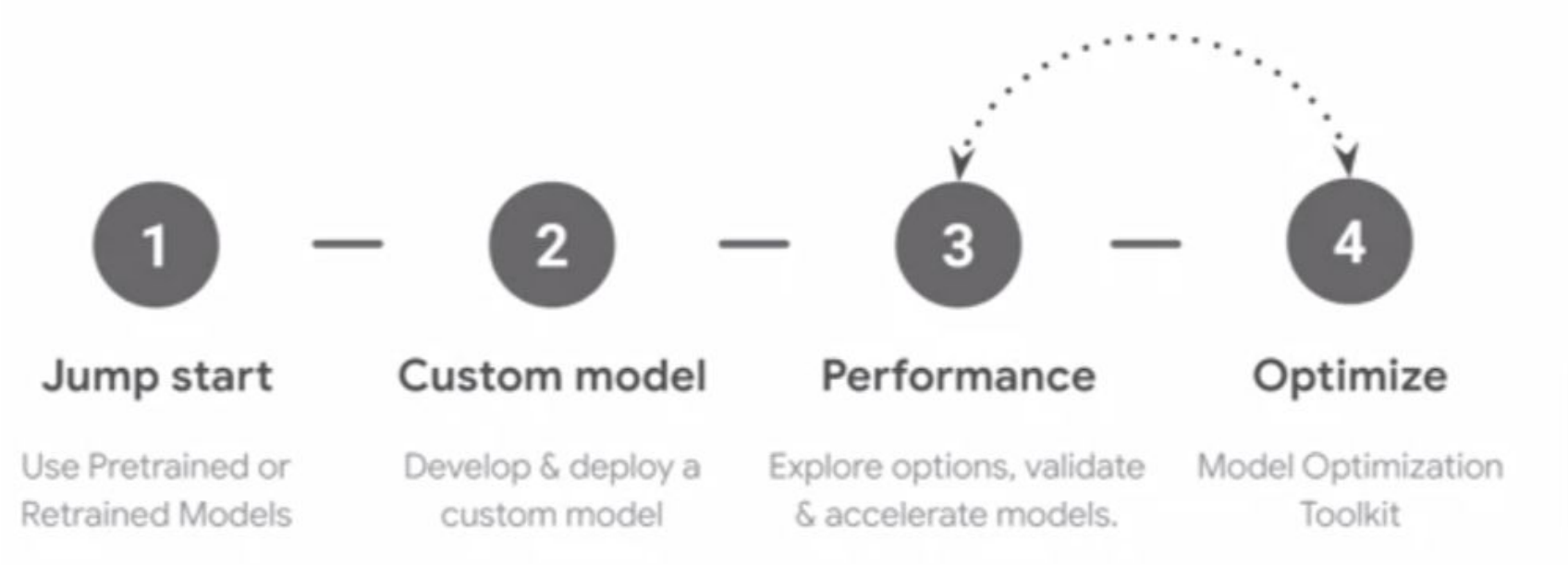
- Quantization, Mengurangi presisi weight dan bias
- Weight pruning, Mengurangi jumlah parameter
- Model topology transform, Mengubah bentuk model
  - Tensor Decomposition
  - Distillation

KOMINFO digitalent

Why Quantize?

- Semua jenis CPU platform disupport
- Mengurangi latency dan inference cost
- Low Memory Footprint
- Allow execution on hardware restricted-to or optimized-for fixed-point operations
- Optimized models for special purpose HW accelerators (TPUs)

Steps dalam optimizing

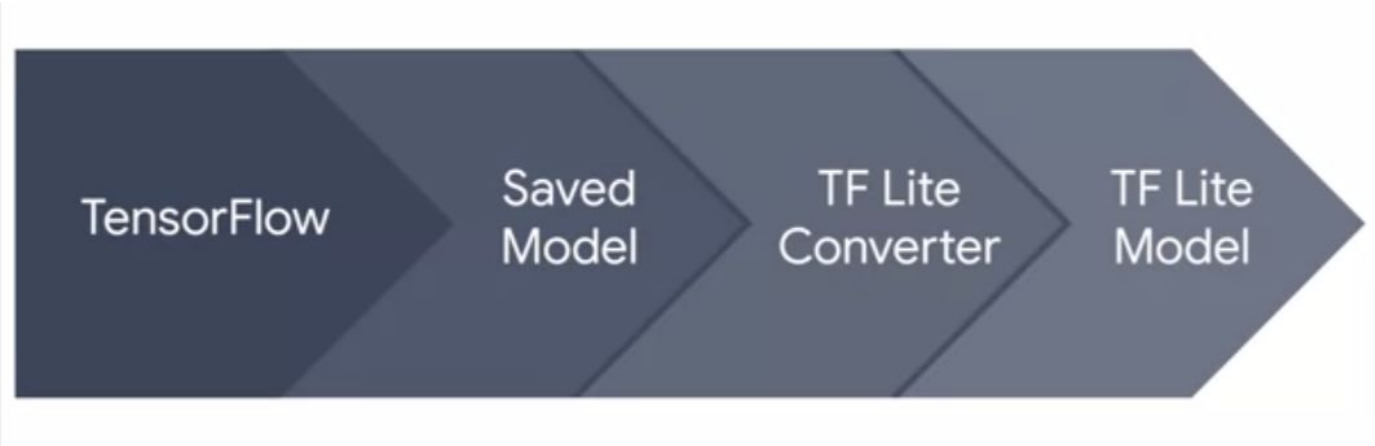


KOMINFO digitalent

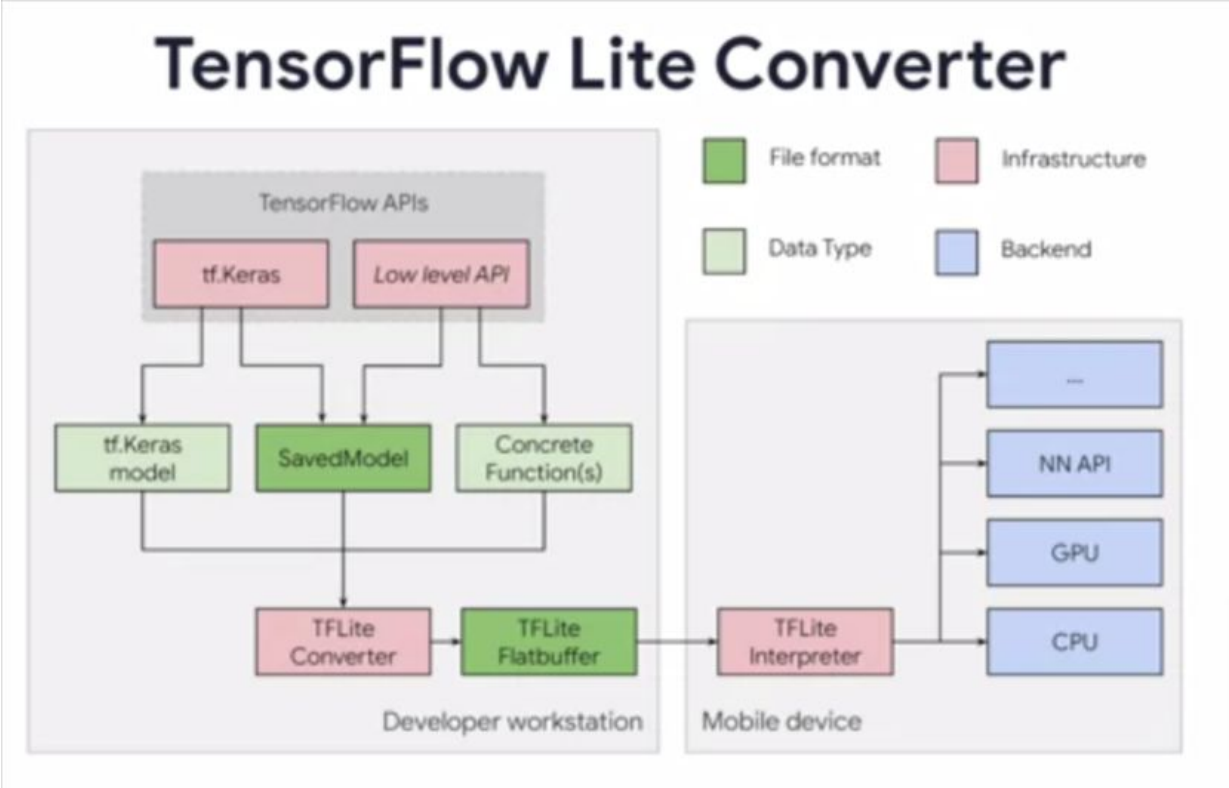
PART 2

Taking a look at the saved model format

Saving, converting, and optimizing a model



Saving, converting, and optimizing a model



Parameters for conversion

KOMINFO digitalent

SavedModel

- Standar untuk serialize sebuah TensorFlow model
- Akan ada MetaGraph, sebuah data flow graph, untuk menyimpan metadata yang membantu prediksi model.
- Ada snapshot dari trained model (dengan weights dan computation model)
- Tidak memerlukan building-code model required
- Supports model versioning

DEMO Example 1

DEMO Example 2

Post-training quantization

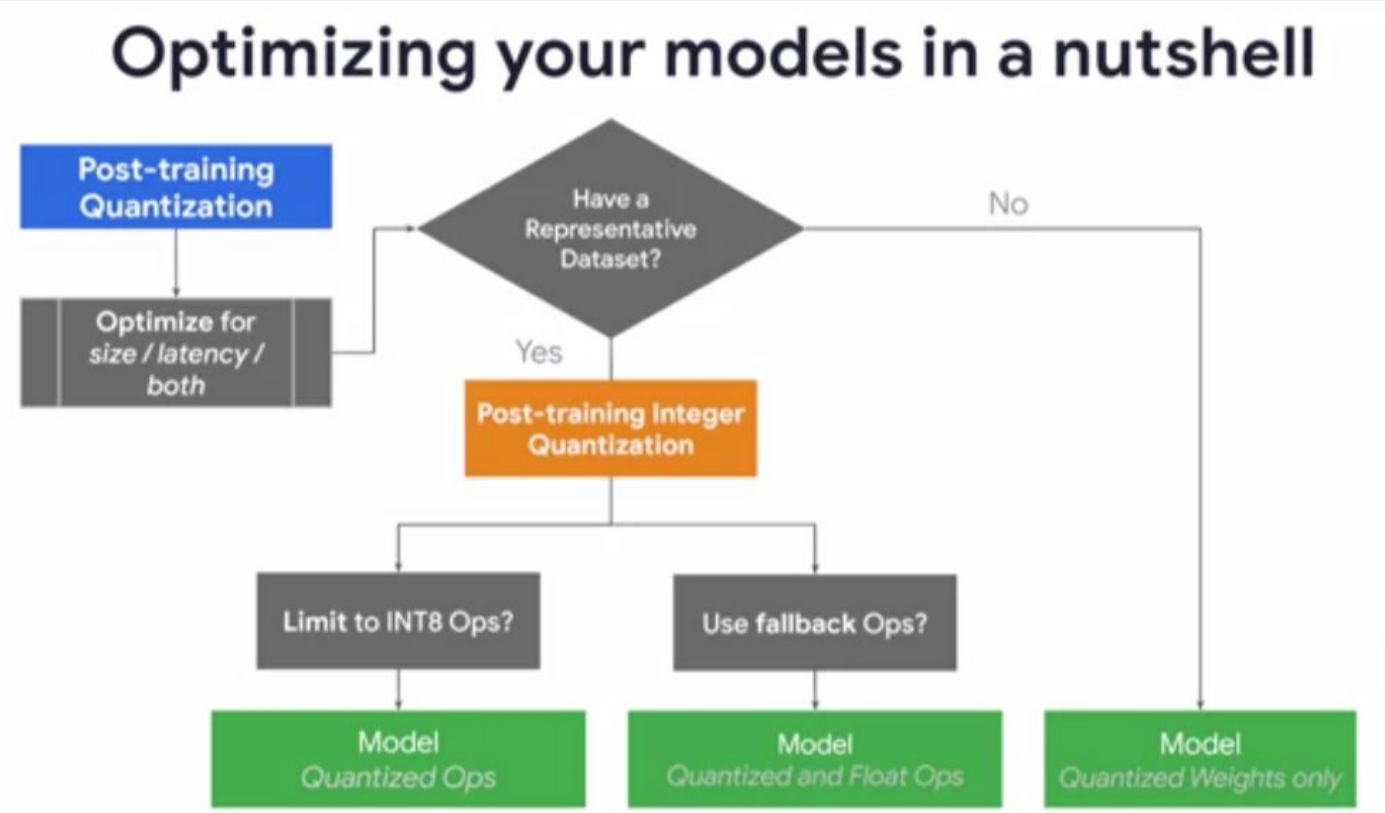
- Mengurangi precision dengan 3 kali lipat lower latency
- Sedikit degradation pada akurasi model
- Optimization modes
  - Default (Both size dan Latency)
  - Size
  - Latency
- Efficiently represents an arbitrary magnitude of ranges
- Quantization target specification

Post-training integer quantization

KOMINFO digitalent

DEMO Post-training quantization

Paths in Optimization



[Quantization in deep learning | Deep Learning Tutorial 49 (Tensorflow, Keras & Python) - YouTube]

KOMINFO digitalent

[Inside TensorFlow: TF Model Optimization Toolkit (Quantization and Pruning) - YouTube]

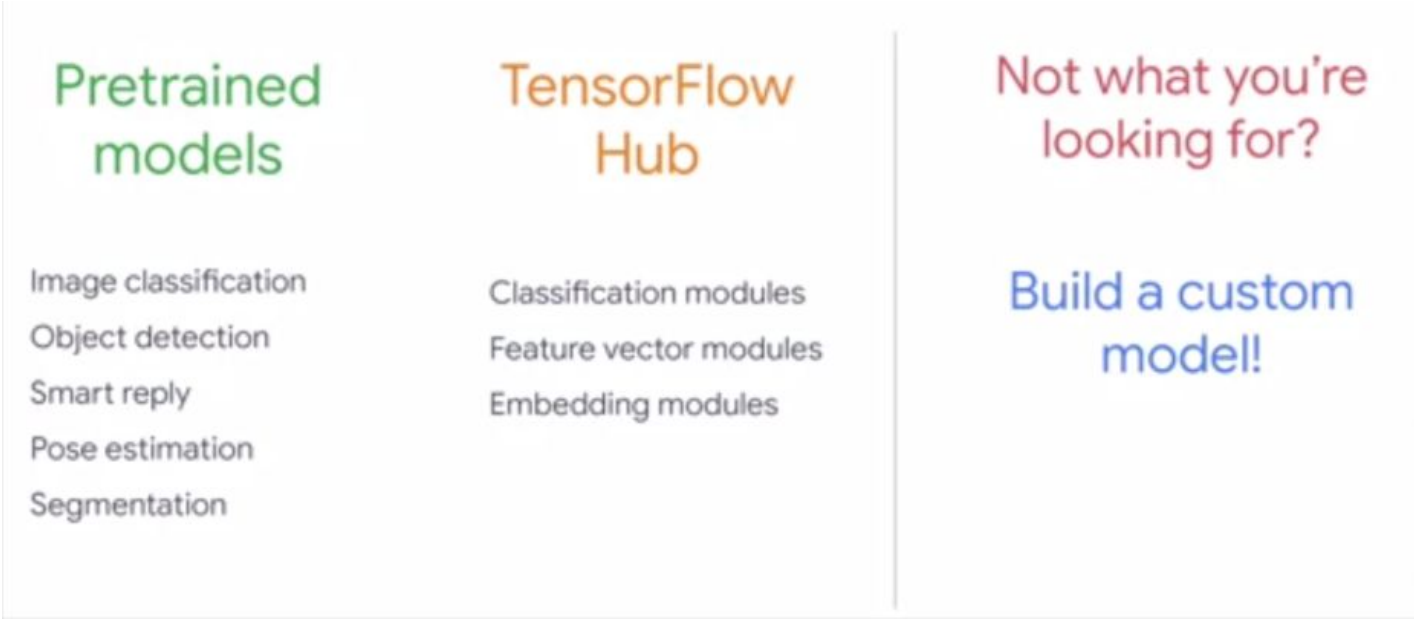
KOMINFO digitalent

DEMO TFLite Interpreter in Python

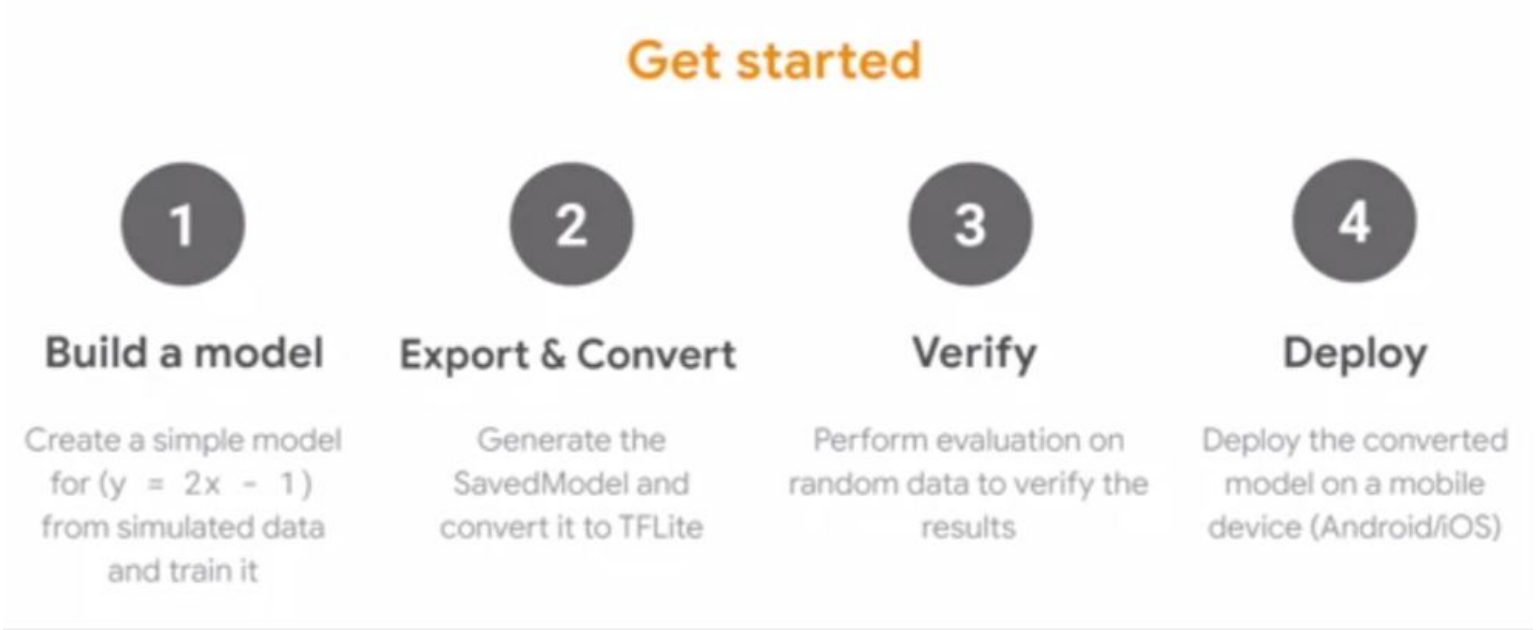
PART 3

First primer on running models on mobile devices

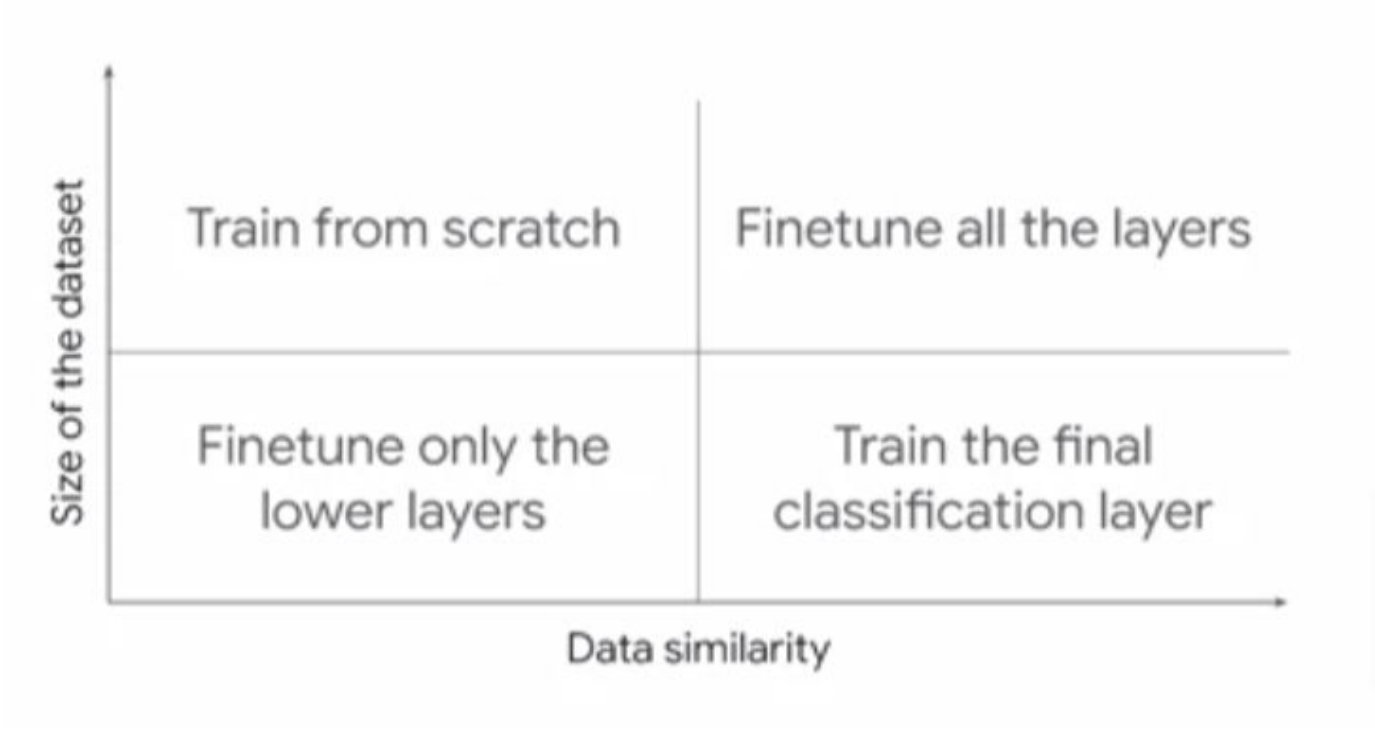
Running Models



Getting a basic model running



Transfer learning



Transfer learning

DEMO Converting a model to TFLite

KOMINFO digitalent

DEMO Transfer Learning with TFLite

KOMINFO digitalent

Q & A

KOMINFO digitalent

Thank You