Women Techmakers Algiers

TinyML light-weight ML/DL

Theory and application of the task of image classification.

TinyML workshop →

This workshop is presented for the audience of Women Techmakers Algiers with ❤

February 8th, 2023

A bit about me ?

Final year CS engineering student

- ESI-ALGER (Algiers, Algeria)
- Computer systems
- Masters and state-engineering degrees at preparation

AI R&D research assistant

- LMCS-INFOLOGIC Engineering (Lyon, France)
- Working on predicting diffrent failures in datacenters and cloud systems using AI

Entrepreneurial kiddo

- Ex. dev team leader at ETIC Club
- Candidate for several engineer-entrepreneur trainings

esi ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE

INFOLOGIC ENGINEERING

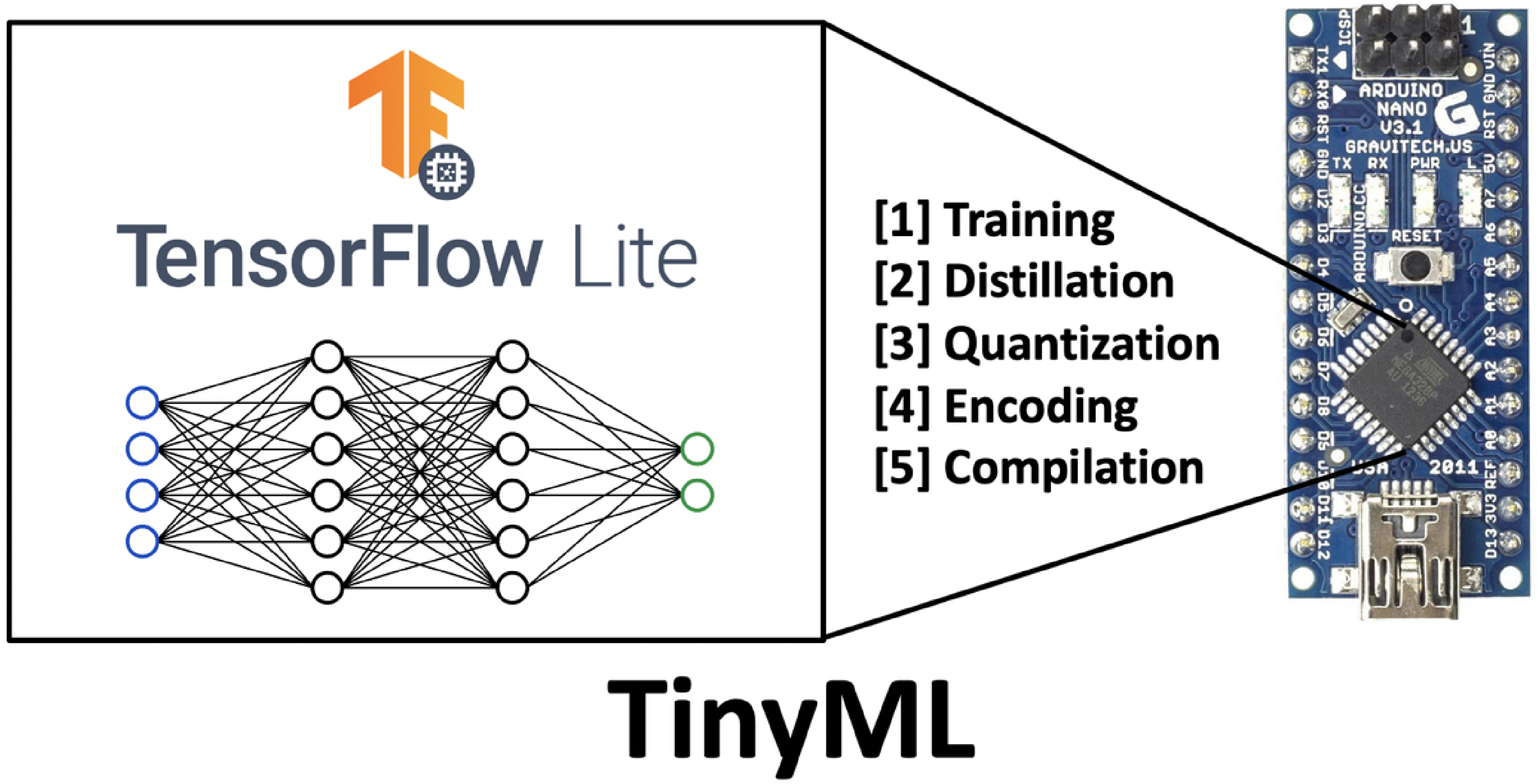
ETIC

I've built a **neural network** what's next then ?

DEPLOYEMENT !

- Option 1 : Cloud computing by deploying the model to the cloud and making API calls. big problem : network latency, storage and computing are costly.

- Option 2 : TinyML frameworks and solutions (MobileNet, TFlite, ...etc)



Pros & Cons

Cloud

- Network latency
- Private data sharing
- Costly ressources (RAM/CPU, Storage)

Edge

- Limited computing power
- Battery consumption
- Limited app size

(TinyML for Edge)

- Not necessarly lower quality but not suitable for large data

Pros & Cons

Cloud

- Suitable for large and complex models
- Suitable for models requiring large data

Edge

- Suitable for real-time ML tasks
- Privacy-preserving

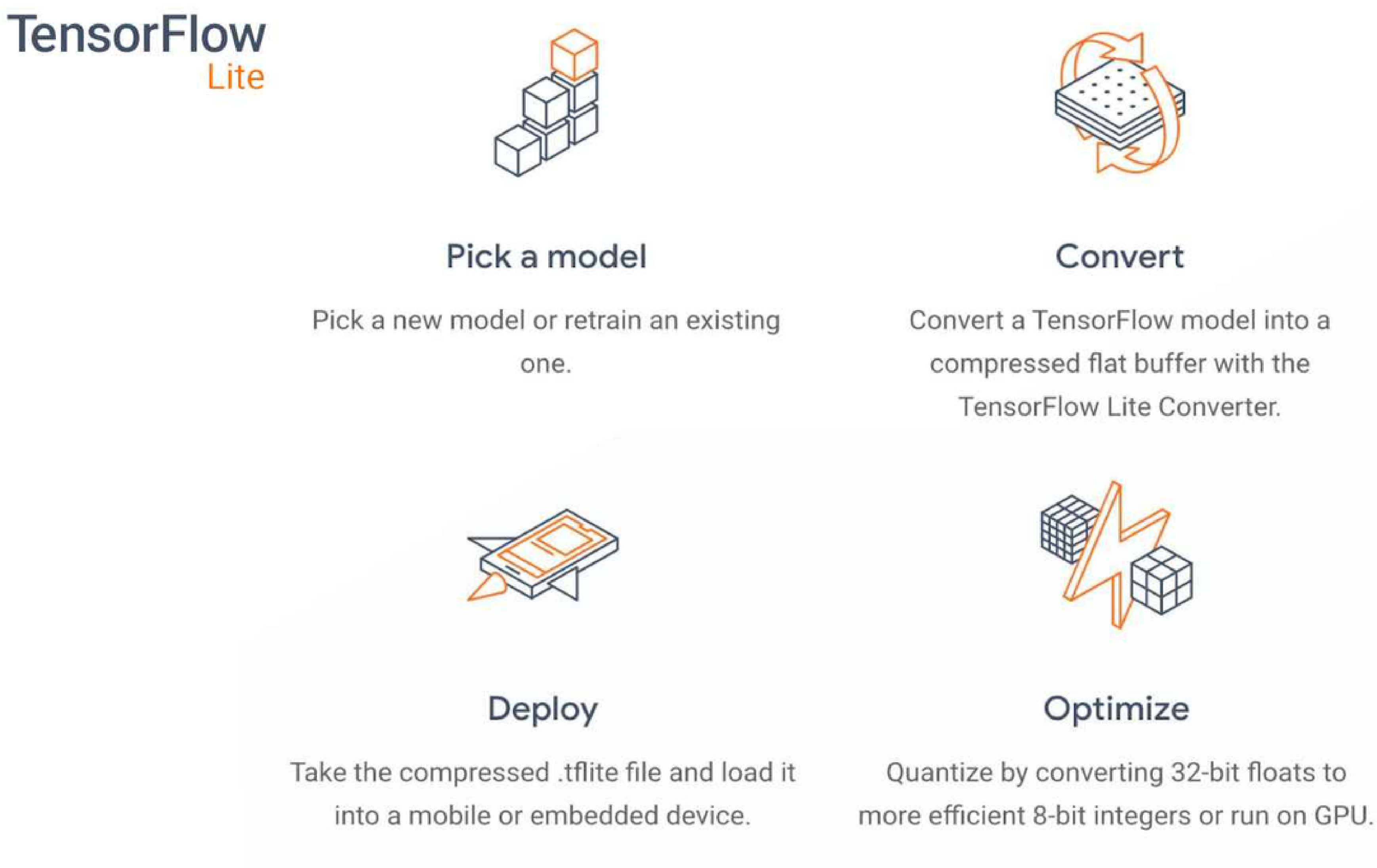
(TinyML for Edge)

- Suitable for deploying models on limited-ressources devices

What's Tensorflow lite anyways ?

- Production-ready
- Cross-plateforme
- ML deployement framework
- Embedded devices & mobiles

android iOS arm



TensorFlow Lite

Pick a model

Pick a new model or retrain an existing one.

Convert

Convert a TensorFlow model into a compressed flat buffer with the TensorFlow Lite Converter.

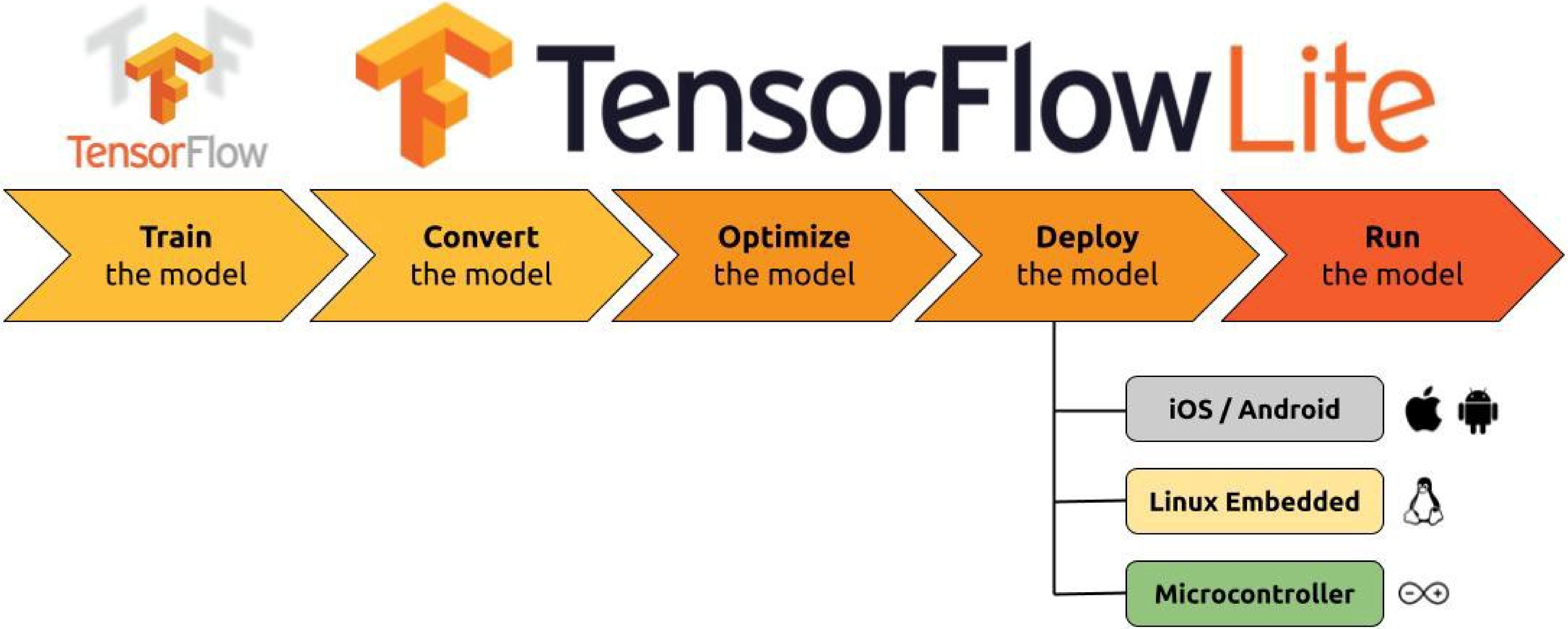
Deploy

Take the compressed .tflite file and load it into a mobile or embedded device.

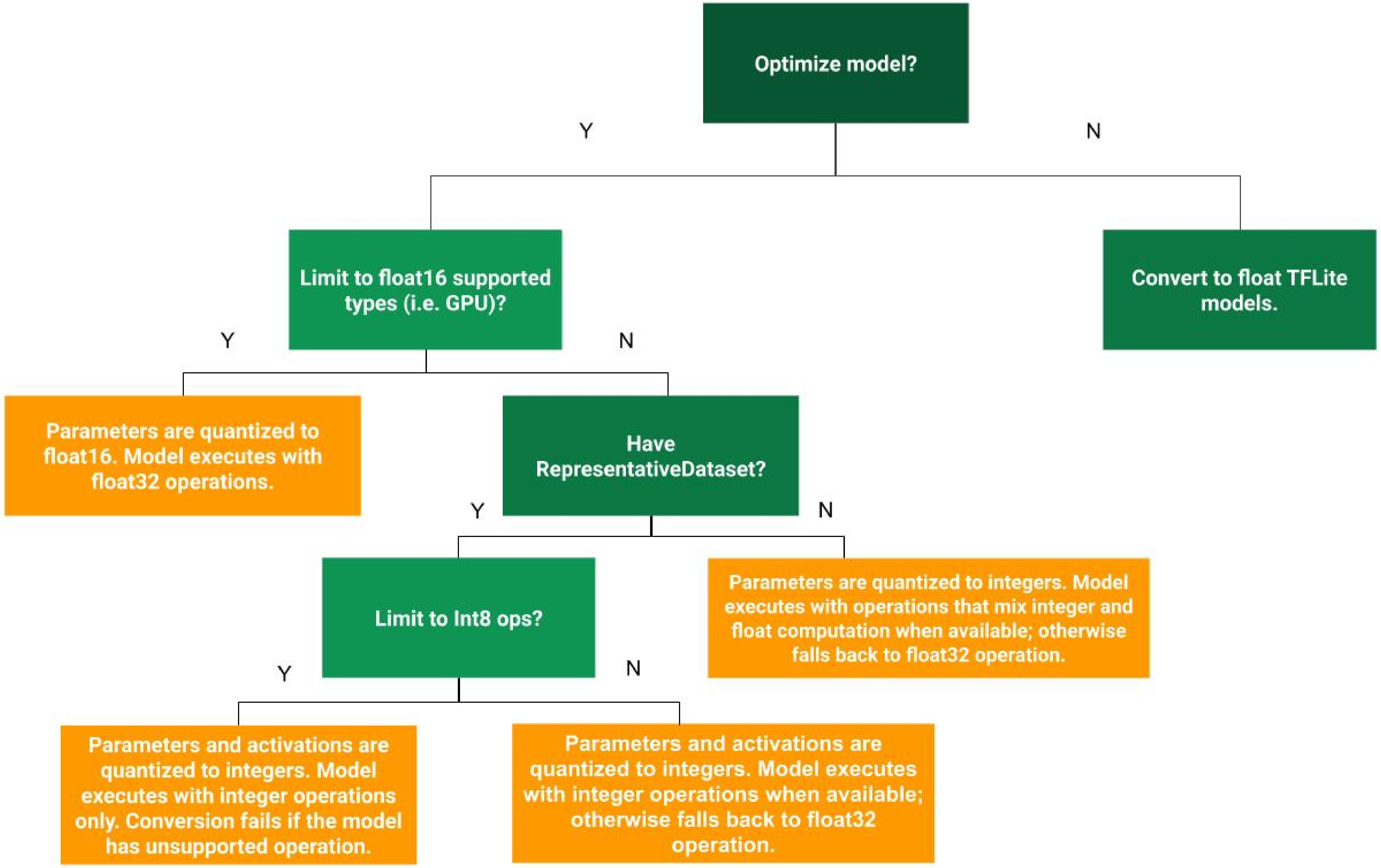
Optimize

Quantize by converting 32-bit floats to more efficient 8-bit integers or run on GPU.

Basic steps of TFlite



Most important TFlite concept : Quantazation



Optimize model?

Convert to float TFLite models.

Limit to float16 supported types (i.e. GPU)?

Parameters are quantized to float16. Model executes with float32 operations.

Have RepresentativeDataset?

Parameters are quantized to integers. Model executes with operations that mix integer and float computation when available; otherwise falls back to float32 operation.

Limit to Int8 ops?

Parameters and activations are quantized to integers. Model executes with integer operations only. Conversion fails if the model has unsupported operation.

Parameters and activations are quantized to integers. Model executes with integer operations when available; otherwise falls back to float32 operation.

Objectives of the workshop

- □ Memory refreshing on tensorflow
- □ Converting tensorflow model to TFLite
- □ Different compressing techniques for embedded devices
- □ Comparing accuracy before and after TFlite
- □ End-to-end implementation for MNIST Fashion

- □ Live Demo of a more complex pretrained classfication task

Thanks!

Questions ?

CODE TIME

FELLAS !

😎