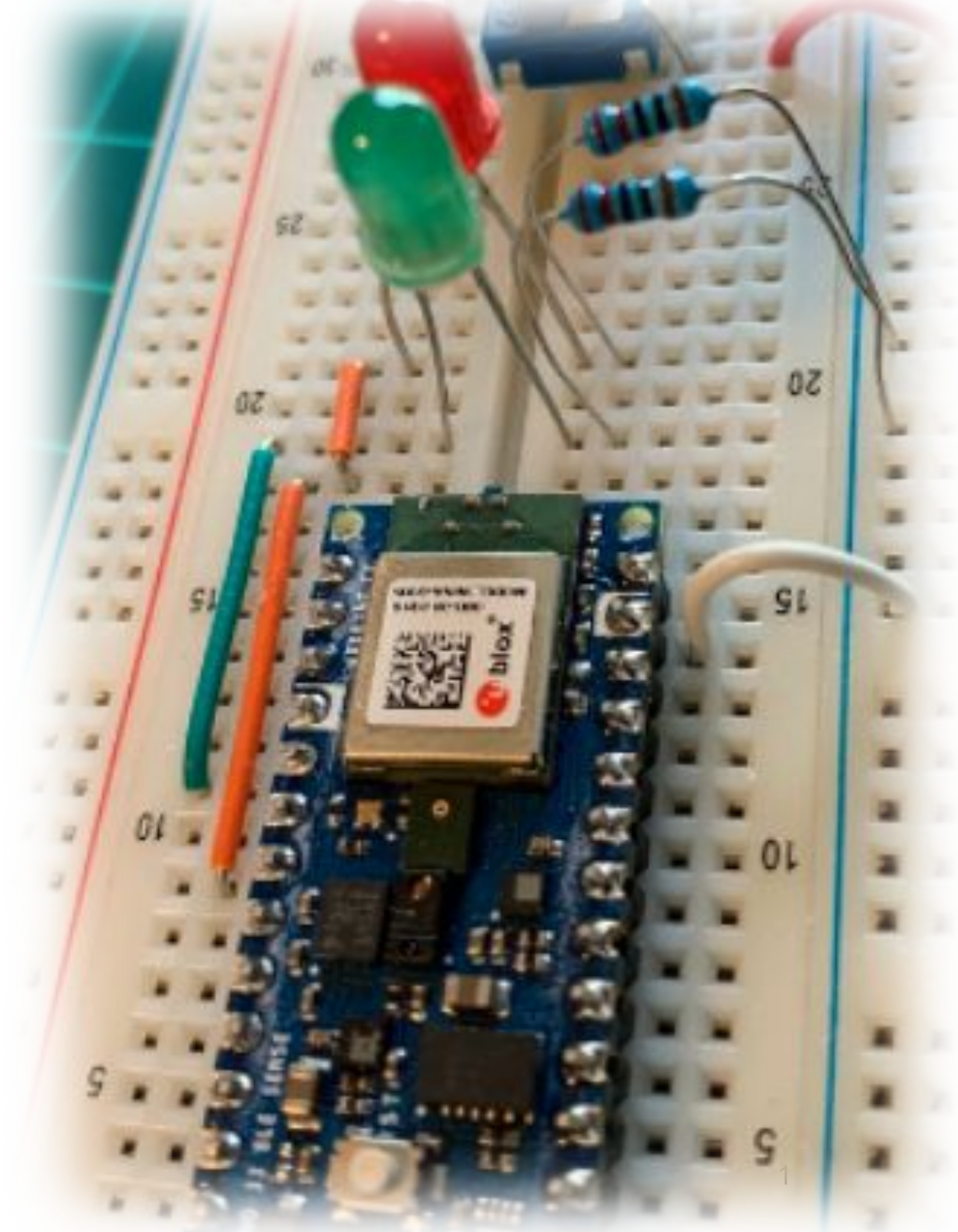
IESTI01 – TinyML

Embedded Machine Learning

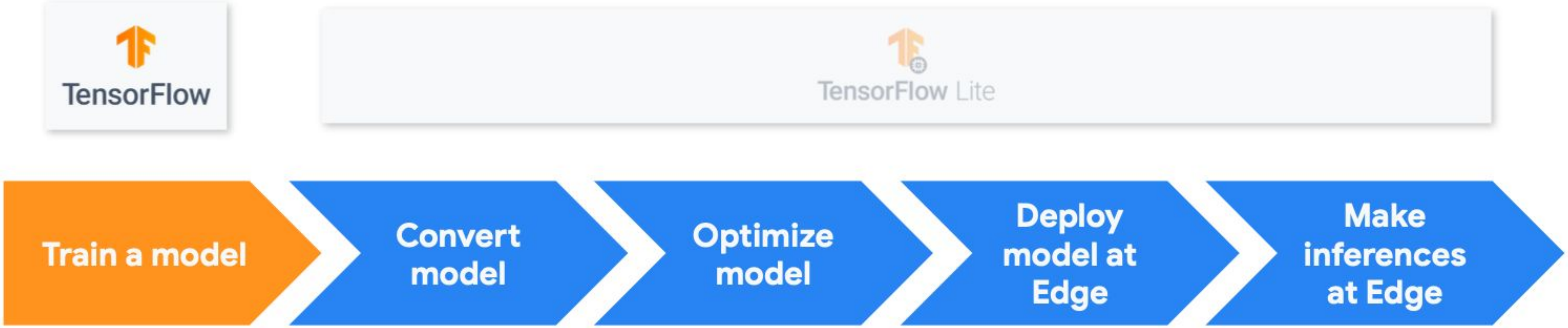
16. Introduction to TensorFlow Lite and TFL-Micro

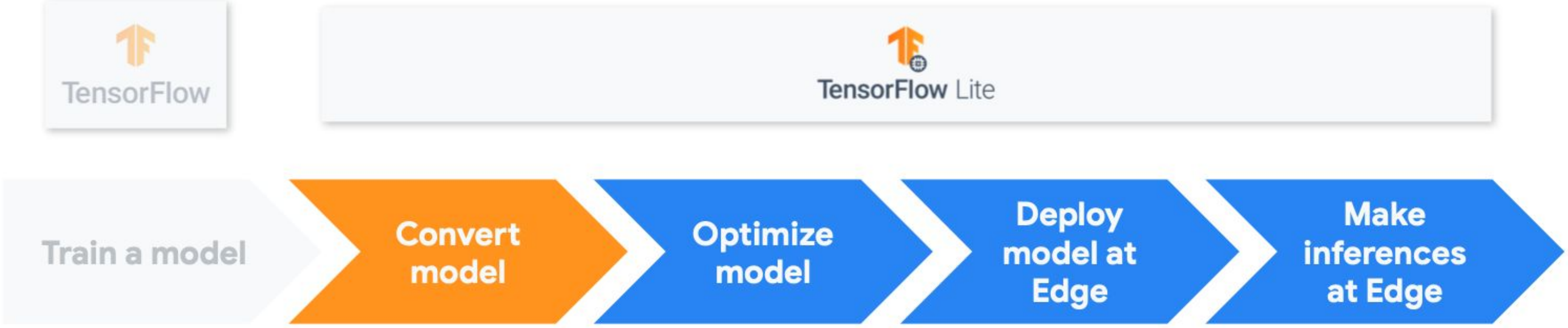
Prof. Marcelo Rovai
UNIFEI

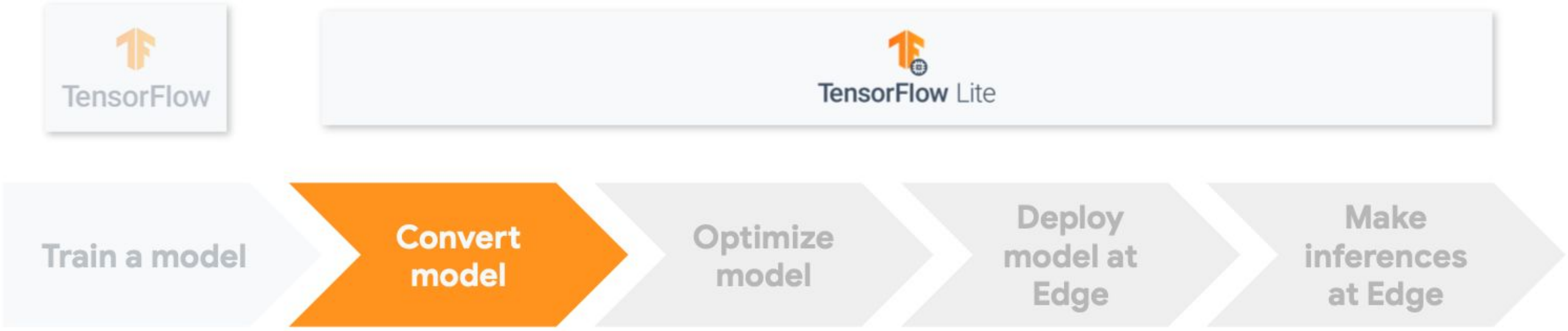


Introduction to TFLite

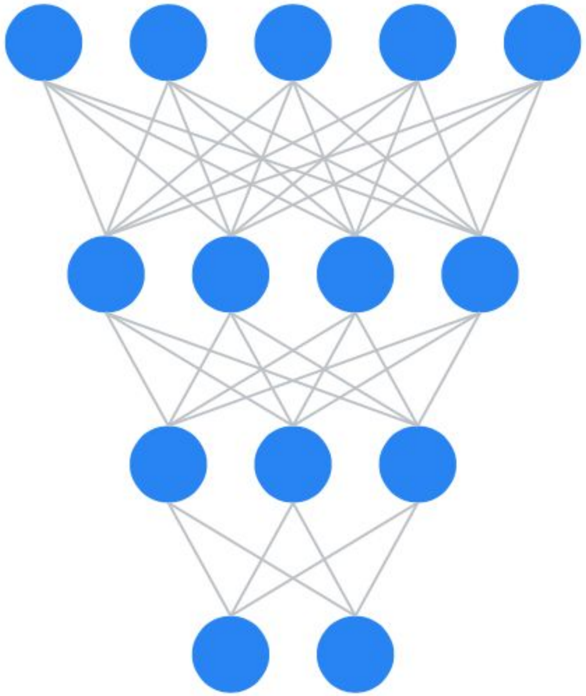
Inference at the Edge


TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge

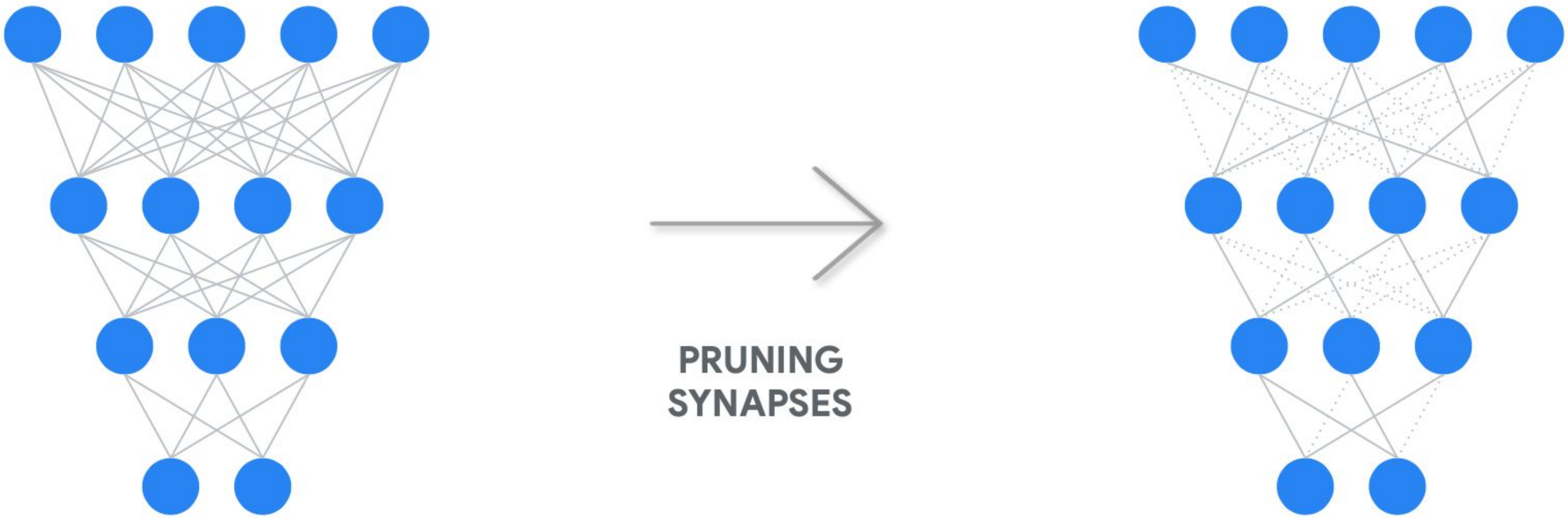

TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge


TensorFlow
TensorFlow Lite
Train a model
Convert model
Optimize model
Deploy model at Edge
Make inferences at Edge

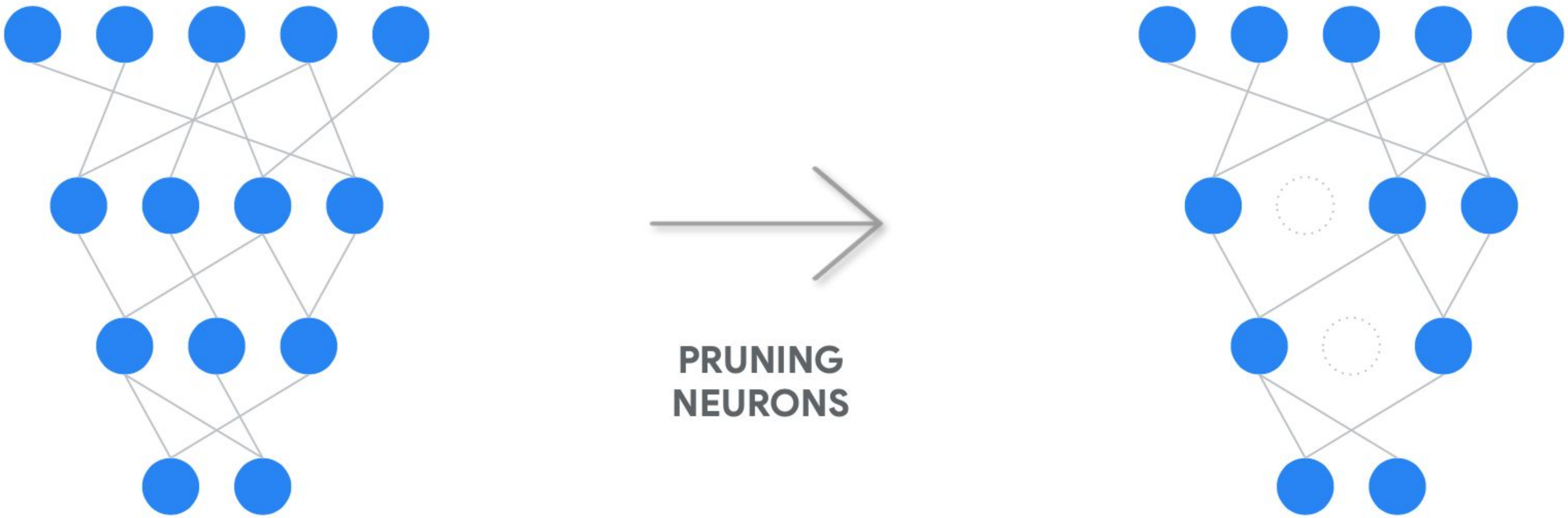
Pruning



Pruning



Pruning



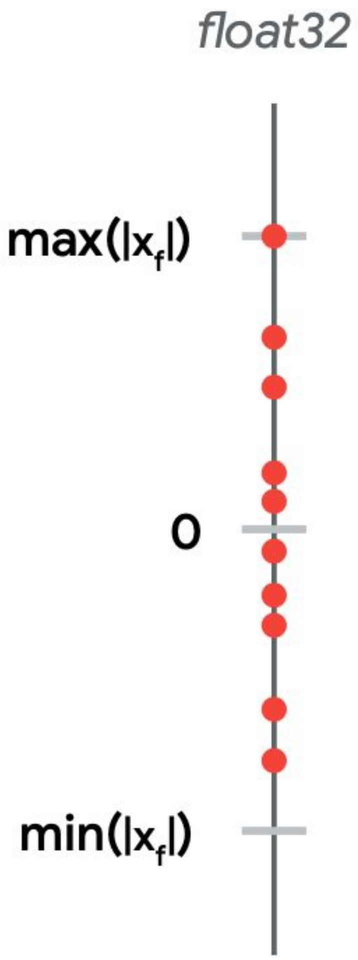
Quantization

Quantization is an optimization that works by **reducing the precision** of the numbers used to represent a model's parameters, which by default are 32-bit floating point numbers. This results in a:

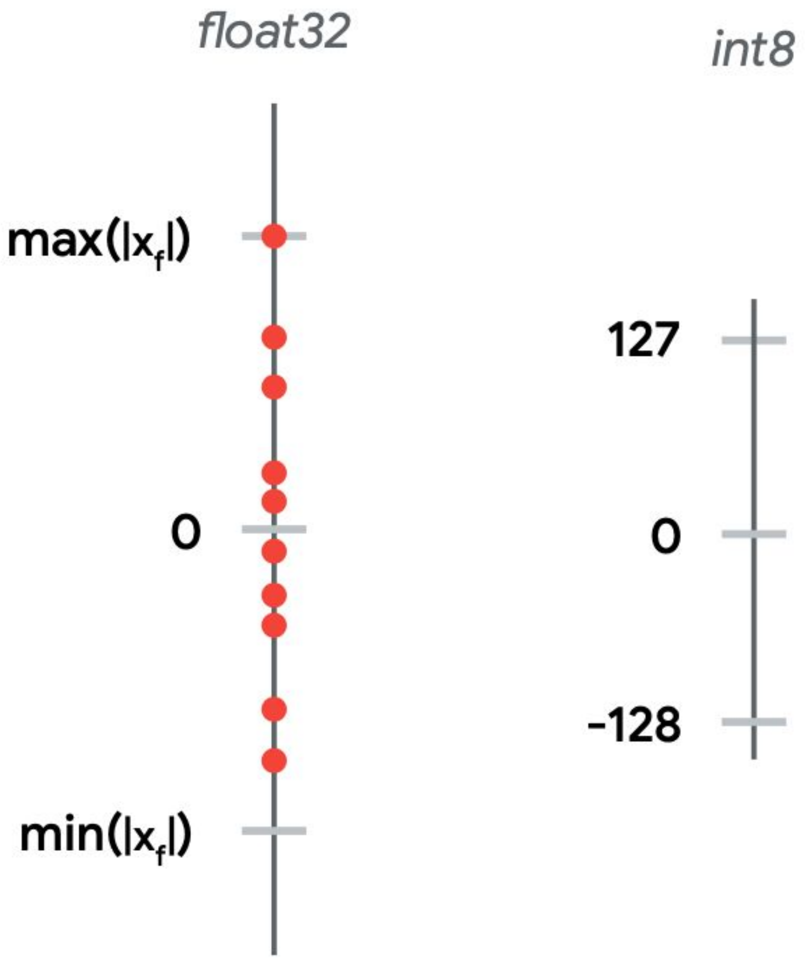
- ✔ **smaller model size**,
- ✔ **better portability (*)** and
- ✔ **faster computation**.

(*) A lot of MCUs do not handle Float-Point operations

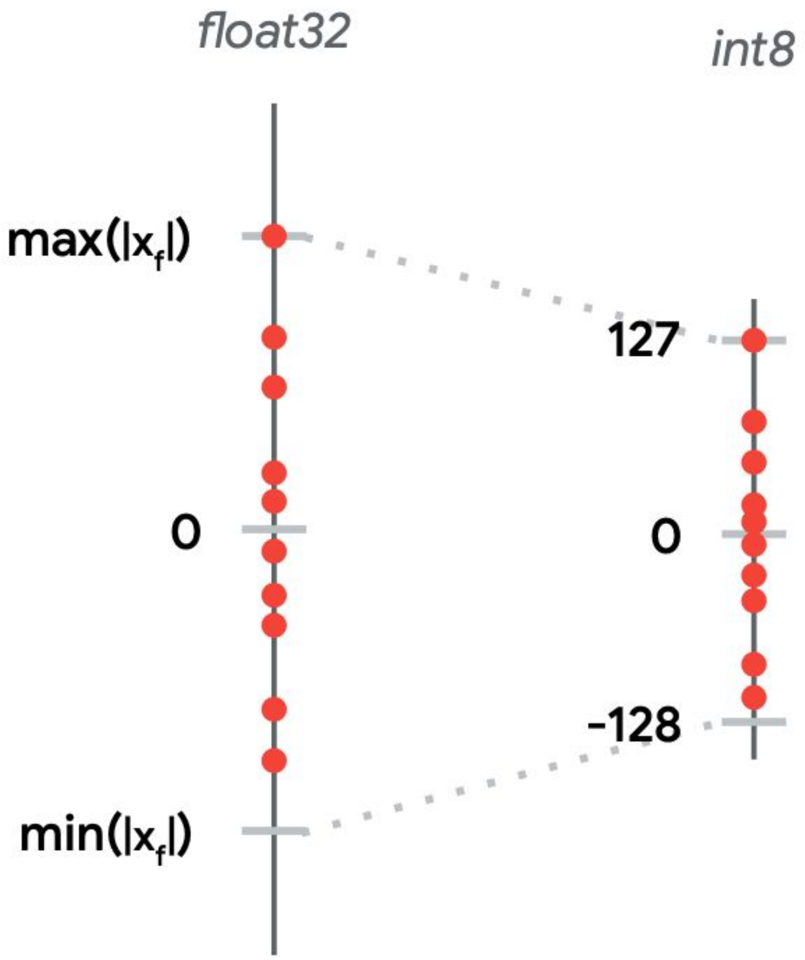
Quantization



Quantization



Quantization



	Floating-point Baseline
MobileNet v1 1.0 224	71.03%
MobileNet v2 1.0 224	70.77%
Resnet v1 50	76.30%

	Floating-point Baseline	Post-training Quantization (PTQ)*
MobileNet v1 1.0 224	71.03%	69.57%
MobileNet v2 1.0 224	70.77%	70.20%
Resnet v1 50	76.30%	75.95%

PTQ is the most common method of quantization. You can also apply QAT (Quantization-aware training), where the parameters are quantized **during** training.

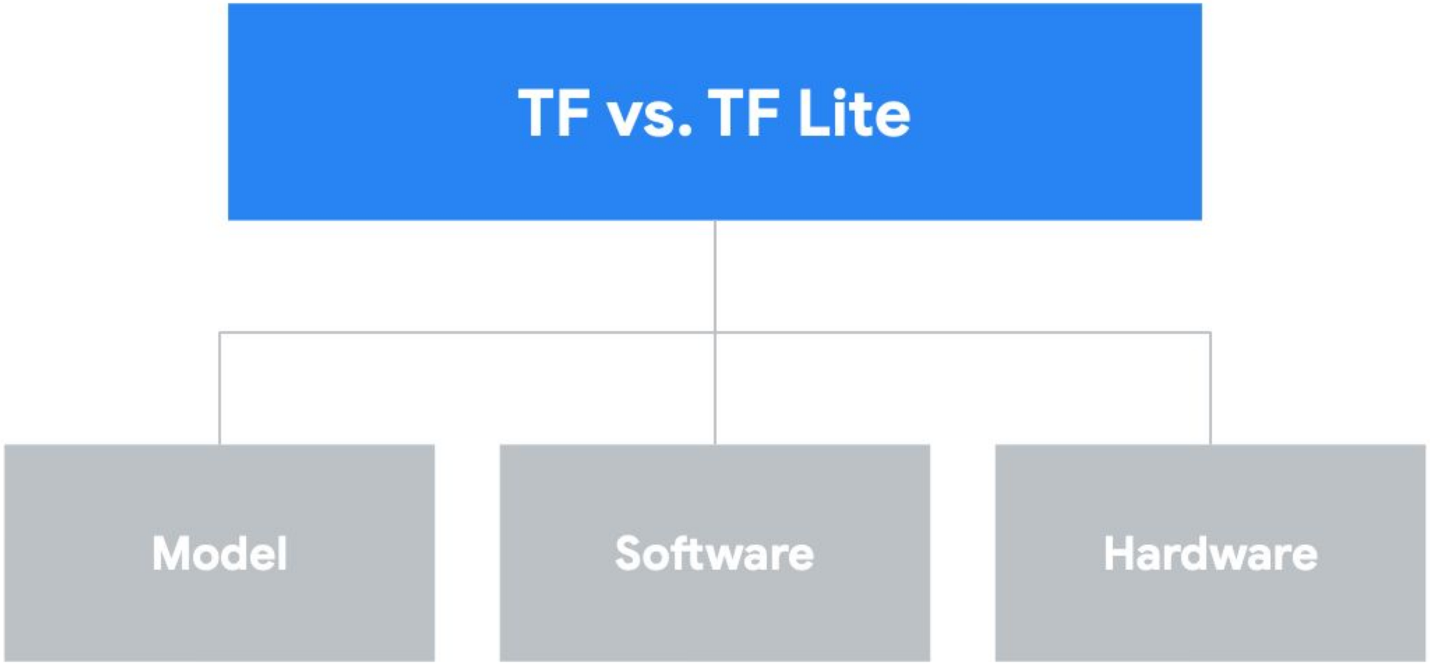
	Floating-point Baseline	Post-training Quantization (PTQ)	Accuracy Drop
MobileNet v1 1.0 224	71.03%	69.57%	▾1.46%
MobileNet v2 1.0 224	70.77%	70.20%	▾0.57%
Resnet v1 50	76.30%	75.95%	▾0.35%

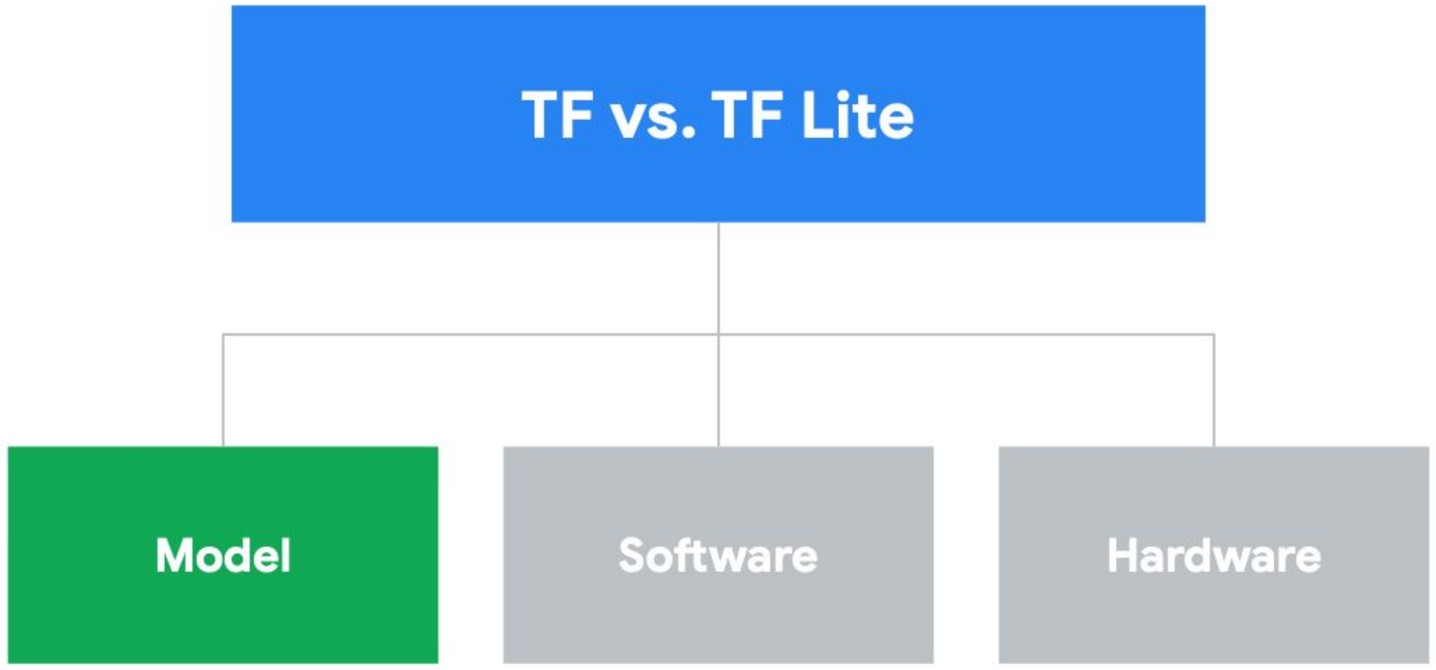
TensorFlow

TensorFlow Lite

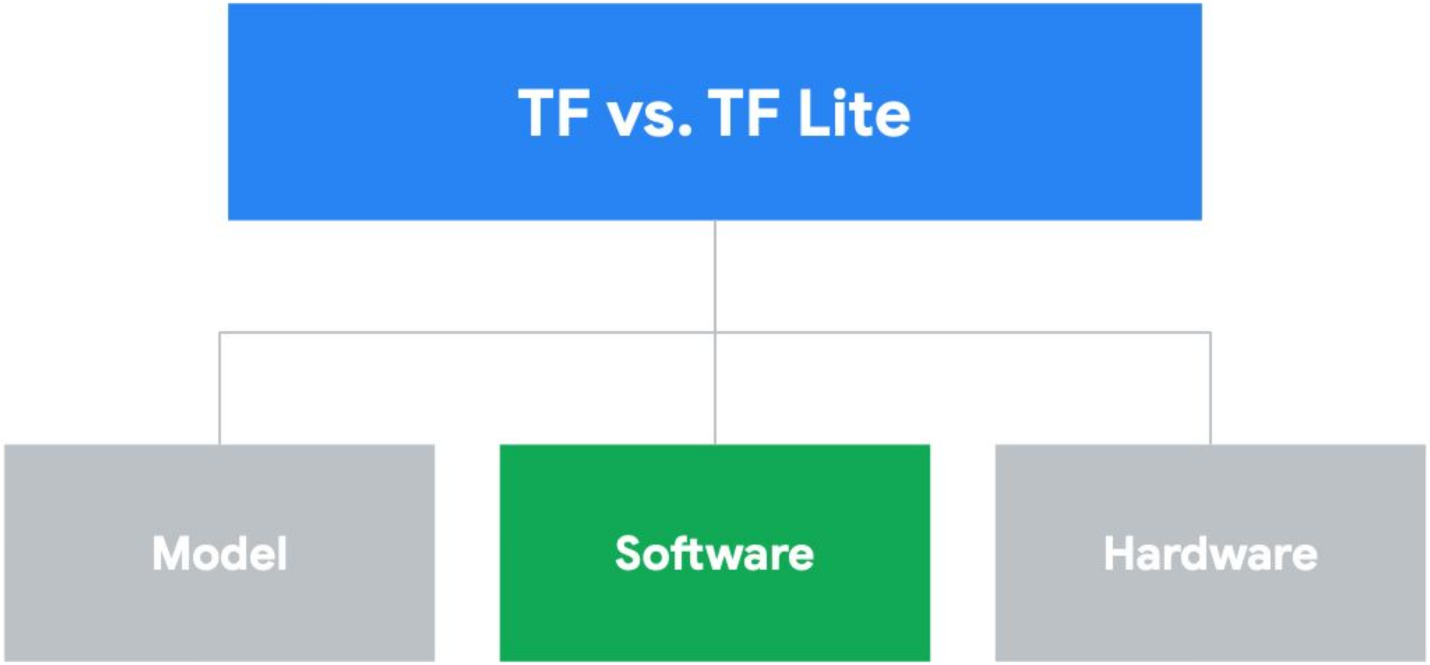
Key Differences

	TensorFlow	TensorFlow Lite
Topology	Variable	Fixed
Weights	Variable	Fixed
Binary Size	Unimportant	High Priority
Distributed Compute	Needed	Not Needed
Developer Background	ML Researcher	Application Developer

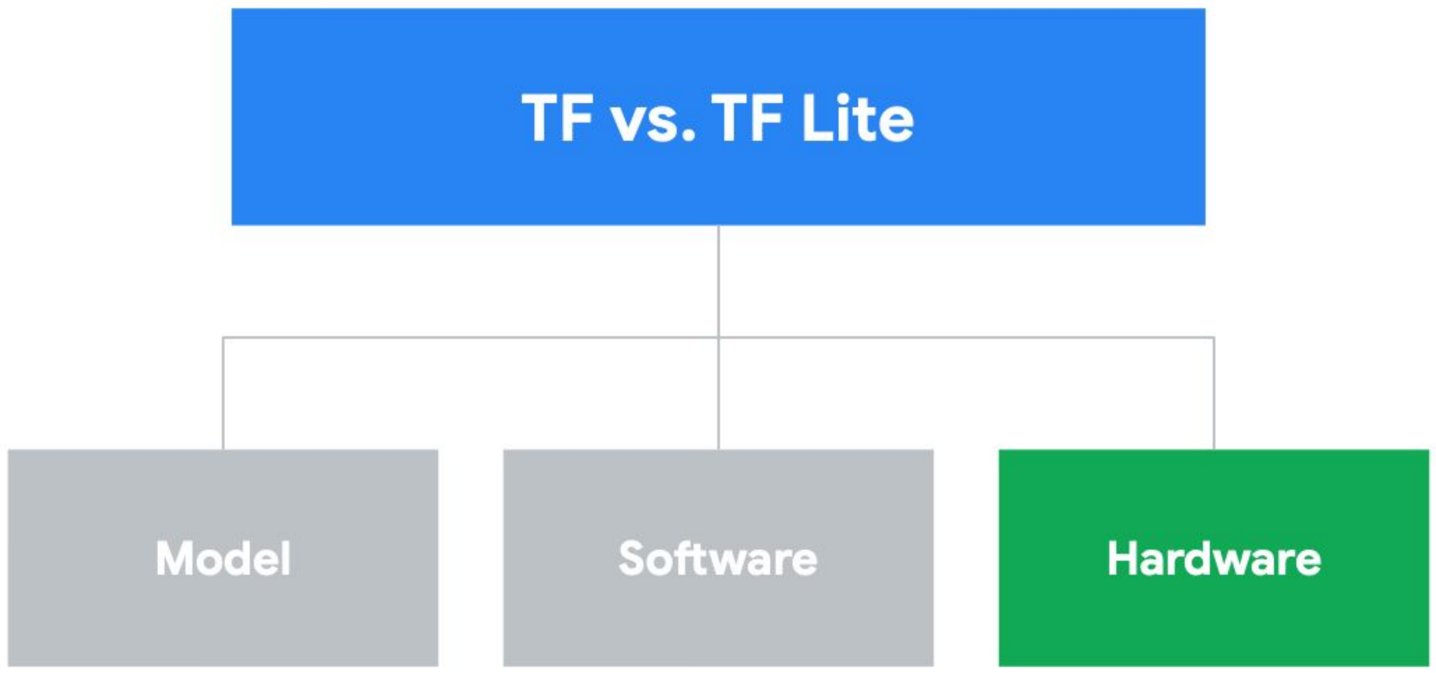

TF vs. TF Lite
Model
Software
Hardware


TF vs. TF Lite
Model
Software
Hardware

	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Training	Yes	No	No
Inference	Yes *(but inefficient on edge)*	Yes *(and efficient)*	Yes *(and even **more** efficient)*
How Many Ops	~1400	~130	~50
Native Quantization Tooling + Support	No	Yes	Yes


TF vs. TF Lite
Model
Software
Hardware

	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Needs an OS	Yes	Yes	No
Memory Mapping of Models	No	Yes	Yes
Delegation to accelerators	Yes	Yes	No

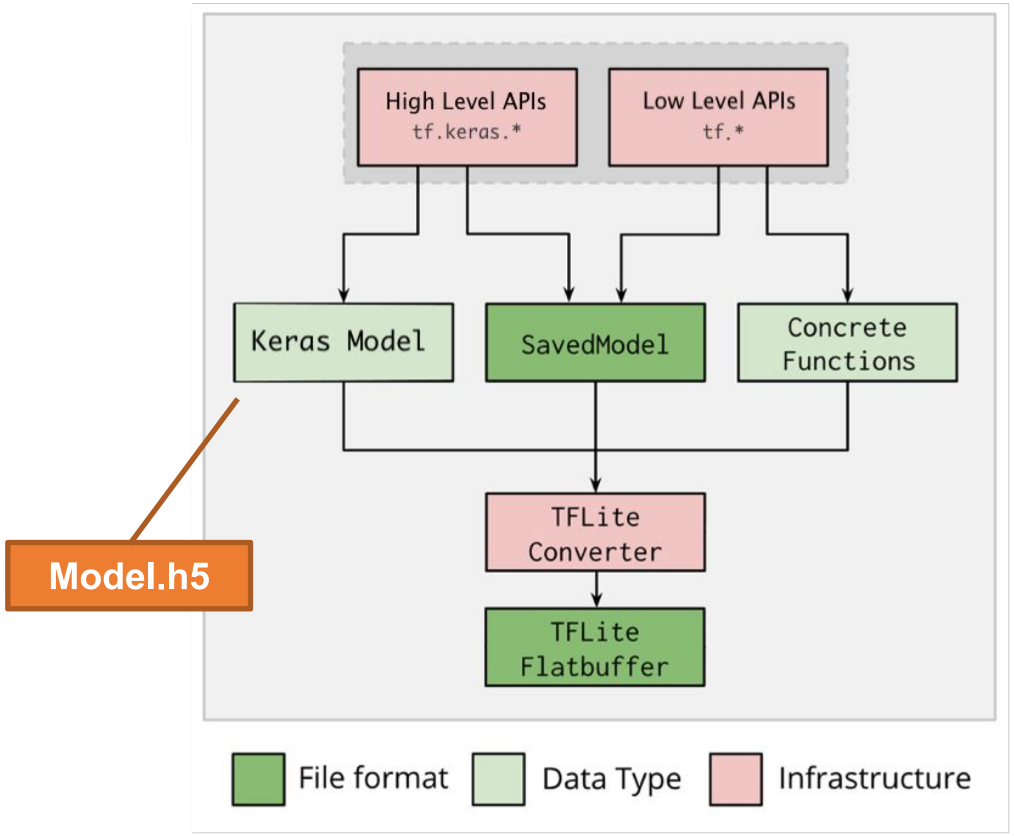

TF vs. TF Lite
Model
Software
Hardware

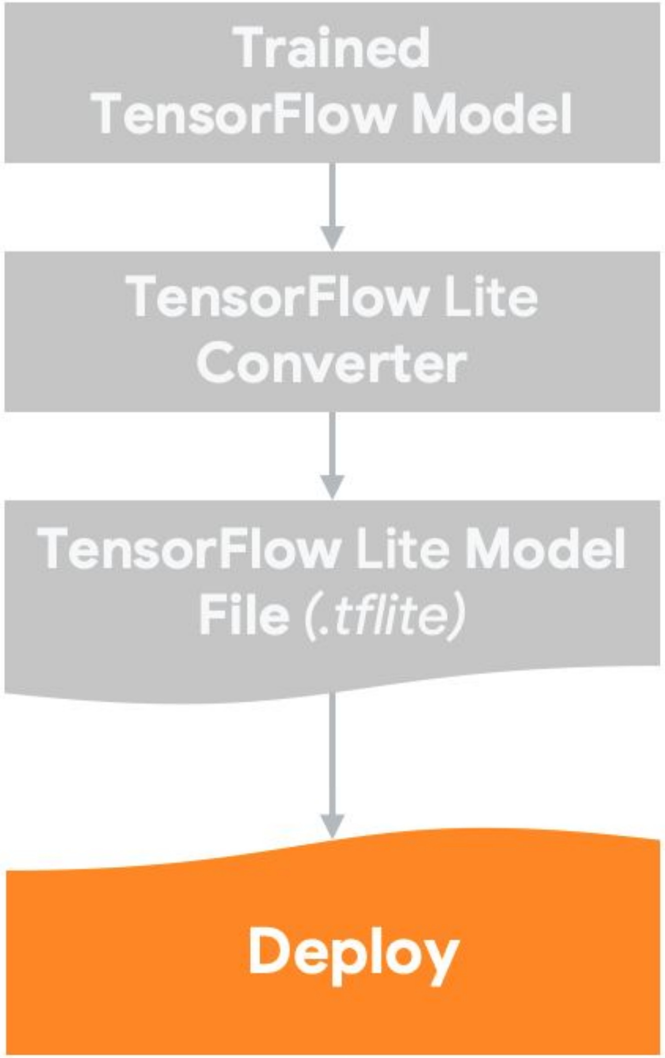
	TensorFlow	TensorFlow Lite	TensorFlow Lite Micro
Base Binary Size	3MB+	100KB	~10 KB
Base Memory Footprint	~5MB	300KB	20KB
Optimized Architectures	X86, TPUs, GPUs	Arm Cortex A, x86	Arm Cortex M, DSPs, MCUs

Optimization and Quantization

Minimizing compression loss

TensorFlow **Workflow**





Converting

Size: 2.1Mb

```
converter = tf.lite.TFLiteConverter.from_keras_model(model)
```

TF Model

```
tflite_model = converter.convert()
```

```
INFO:tensorflow:Assets written to: /tmp/tmprqr8kgp4/assets
```

```
# Save .tflite model
open("/content/cifar10.tflite","wb").write(tflite_model)
```

TFLite Model

```
673324
```

Size: .63Mb

Converting from a saved model

Size: 2.1Mb

```
[81]  1 model_path = '/content/cifar_10_model.h5'
```

TF Model

```
[82]  1 model_cifar10 = tf.keras.models.load_model(model_path)
```

```
[83]  1 converter = tf.lite.TFLiteConverter.from_keras_model(model_cifar10)
```

```
[84]  1 tflite_model = converter.convert()

INFO:tensorflow:Assets written to: /tmp/tmp6fwji5s_/assets
INFO:tensorflow:Assets written to: /tmp/tmp6fwji5s_/assets
```

Save tflite model

Size: .63Mb

```
[85]  1 open("/content/cifar10.tflite","wb").write(tflite_model)

673324
```

TFLite Model

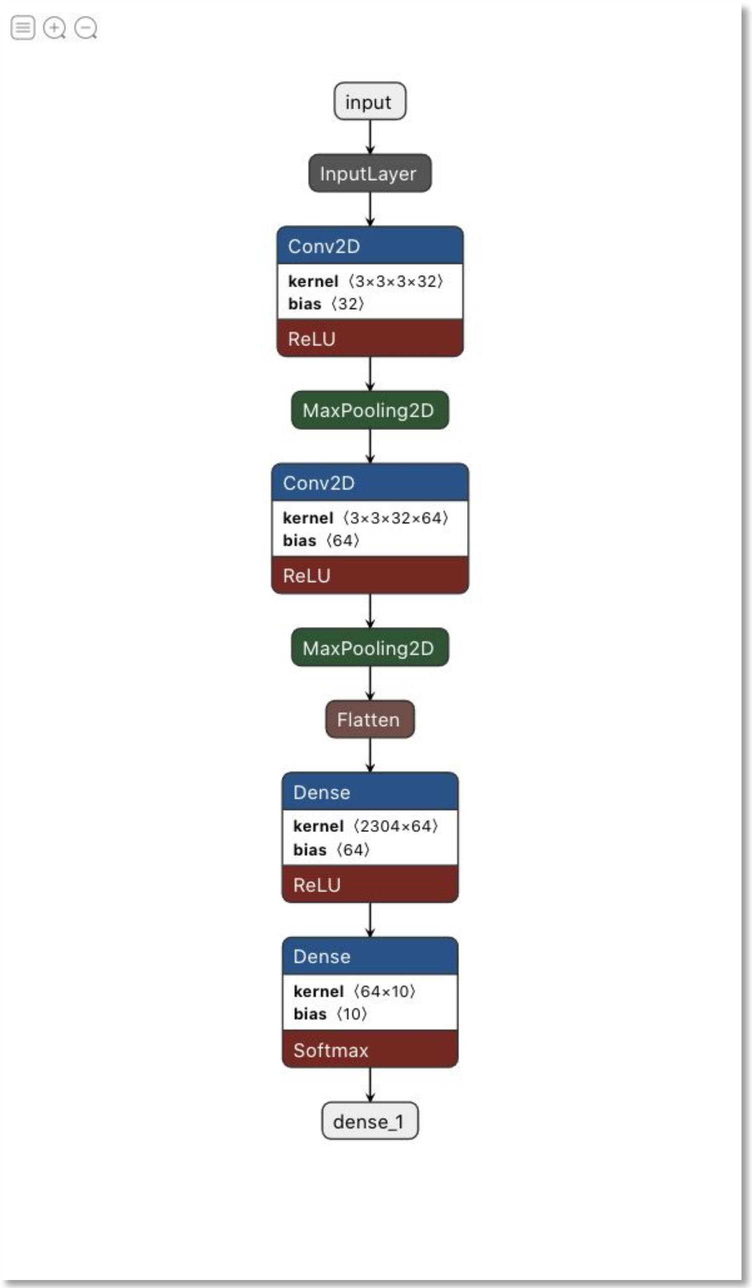
Dynamic range quantization

The simplest form of post-training quantization statically quantizes only the weights from floating point to integer, which has 8-bits of precision:

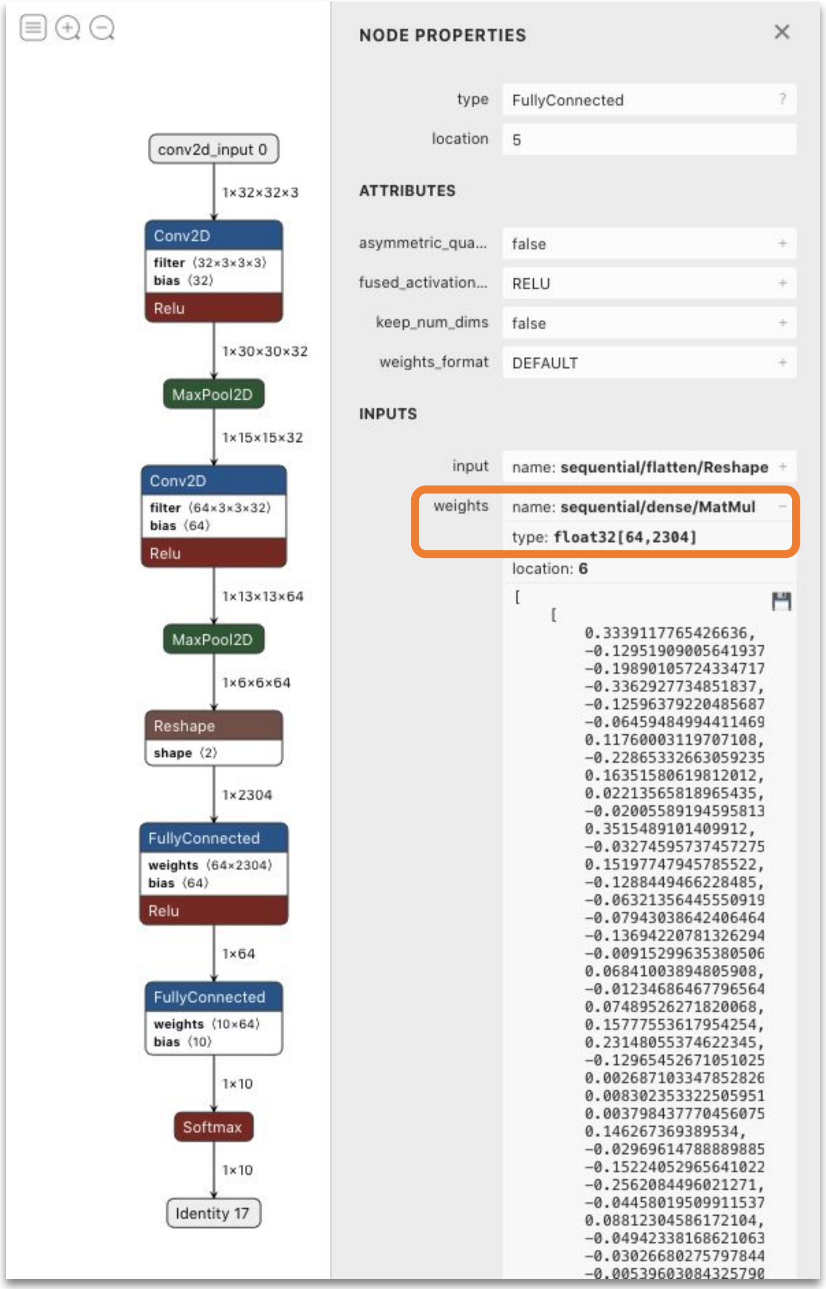
```
[74]  converter = tf.lite.TFLiteConverter.from_keras_model(model)
      converter.optimizations = [tf.lite.Optimize.DEFAULT]
      tflite_quant_model = converter.convert()

INFO:tensorflow:Assets written to: /tmp/tmpyyiq46sj/assets
INFO:tensorflow:Assets written to: /tmp/tmpyyiq46sj/assets
```

Size: .18Mb

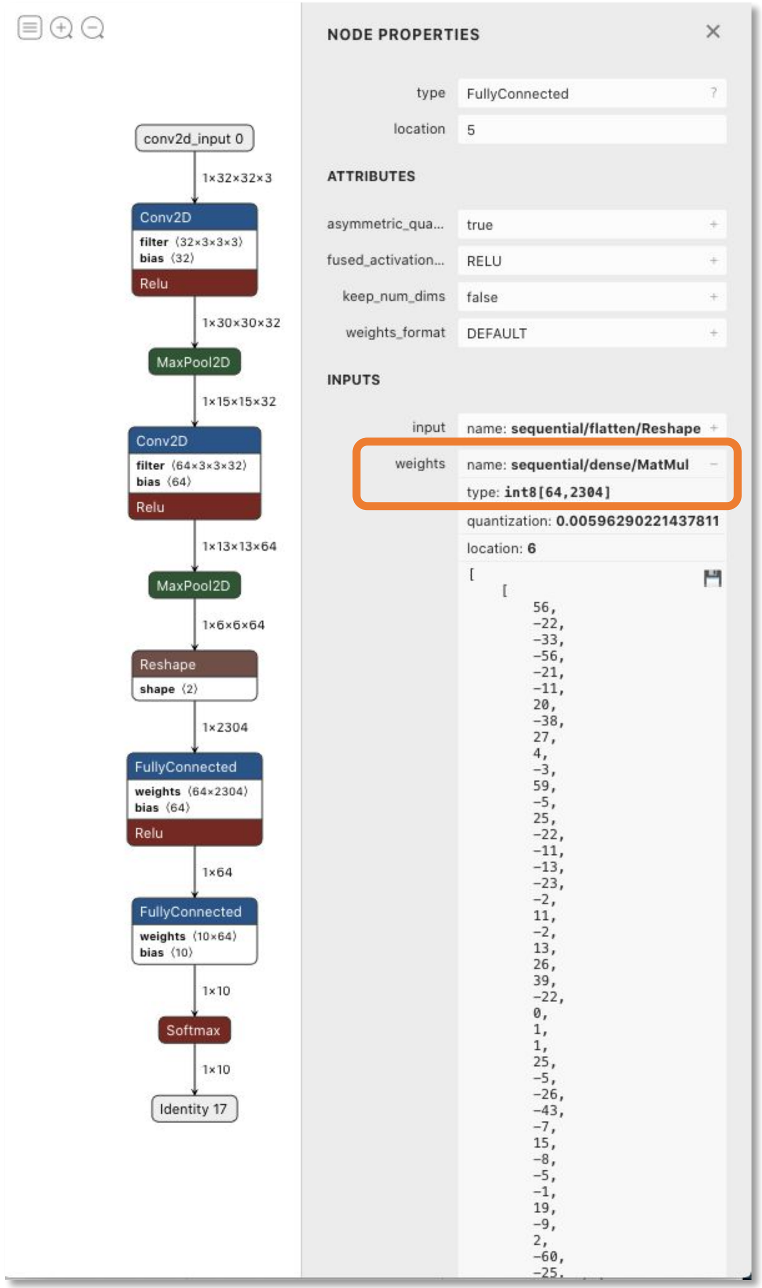
```
[75]  # Save .tflite model
      open("/content/cifar10_quant.tflite","wb").write(tflite_quant_model)

177232
```



Cifar_10.h5



Cifar_10.tflite



Cifar_quant_10.tflite

Image Classification (inference) using TF-Lite

Code Time!

CNN_Cifar_10_TFLite.ipynb

Reading Material

Main references

- Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning
- Professional Certificate in Tiny Machine Learning (TinyML) – edX/Harvard
- Introduction to Embedded Machine Learning (Coursera)
- Text Book: "TinyML" by Pete Warden, Daniel Situnayake

I want to thank Shawn Hymel and Edge Impulse, Pete Warden and Laurence Moroney from Google, and especially Harvard professor Vijay Janapa Reddi, Ph.D. student Brian Plancher and their staff for preparing the excellent material on TinyML that is the basis of this course at UNIFEI.

The IESTI01 course is part of the TinyML4D, an initiative to make TinyML education available to everyone globally.

Thanks

And stay safe!

UNIFEI