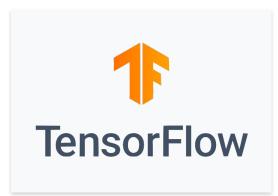
Inference Engine: TF vs. TFLite

ML Workflow















Less memory

Less compute power

Only focused on inference



† TensorFlow

Fixed

TensorFlow Lite

Fixed

High Priority

Not Needed

Application Developer

Topology

Weights

Binary Size

Distributed Compute

Developer Background

Variable

Variable

Unimportant

Needed

ML Researcher

† TensorFlow

Fixed

TensorFlow Lite

Fixed

High Priority

Not Needed

Application Developer

Topology

Weights

Binary Size

Distributed Compute

Developer Background

Variable

Variable

Unimportant

Needed

ML Researcher

Topology

Weights

Binary Size

Distributed Compute

Developer Background



Variable

Variable

Unimportant

Needed

ML Researcher



Fixed

Fixed

High Priority

Not Needed

Application Developer

Topology

Weights

Binary Size

Distributed Compute

Developer Background



Variable

Variable

Unimportant

Needed

ML Researcher



Fixed

Fixed

High Priority

Not Needed

Application Developer

Topology

Weights

Binary Size

Distributed Compute

Developer Background



Variable

Variable

Unimportant

Needed

ML Researcher



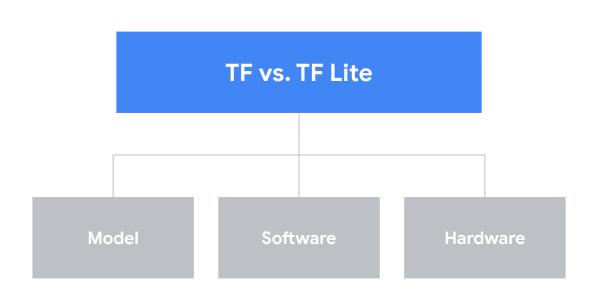
Fixed

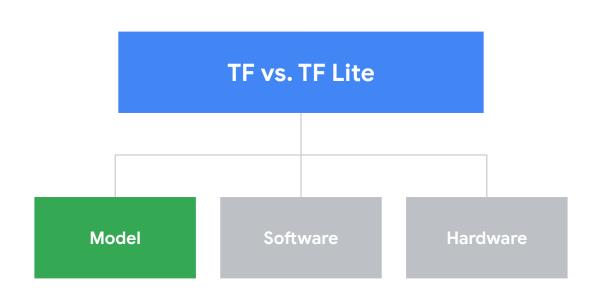
Fixed

High Priority

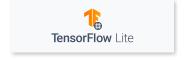
Not Needed

Application Developer









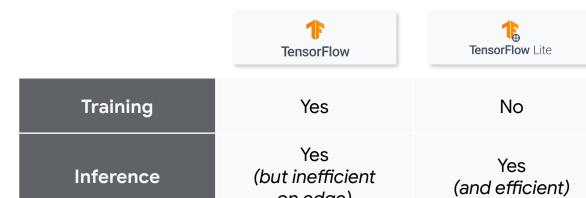


Training

Yes

No

No



on edge)

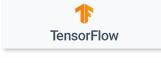
TensorFlow Lite Micro

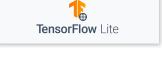
No

Yes

(and even

more efficient)







Training	

How Many Ops

Yes

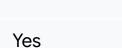
No

Yes



Yes

~130



Inference

(but inefficient

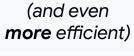
(and efficient)





on edge)

~1400



~50

† TensorFlow	TensorFlow Lite
Yes	No

Training

Inference

How Many Ops

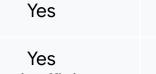
Native Quantization

Tooling + Support



~1400

No









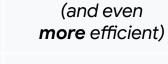


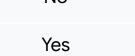


~130

Yes





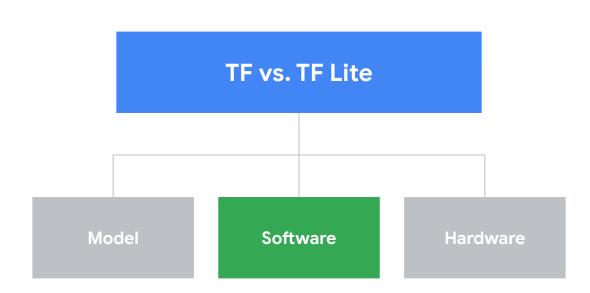


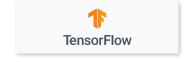
~50

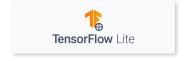
Yes



TensorFlow Lite Micro







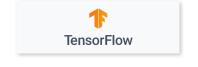


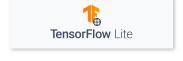
Needs an OS

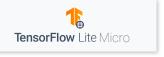
Yes

Yes

No







	Needs	an	0

Yes

Yes

No

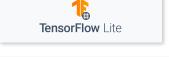
Memory Mapping of Models

No

Yes

Yes







Needs an OS
Memory Mappir

Yes

Yes

pping

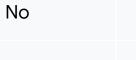
of Models

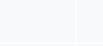
Yes

Yes



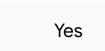
Delegation to





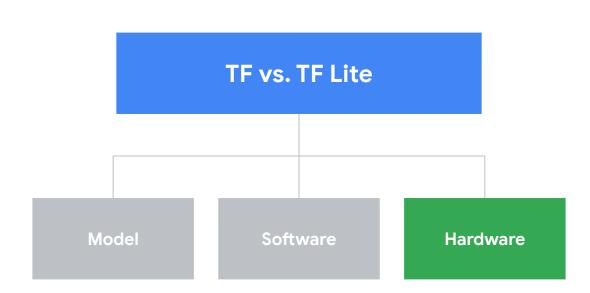




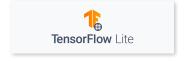


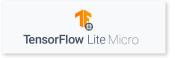










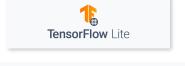


Base Binary Size

3MB+

100KB ~10 KB







Base Binary Size

-..-

100KB

~10 KB

Base Memory Footprint

~5MB

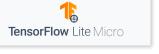
300KB

~10 KB

20KB







Base Binary Size 3MB+

100KB

~10 KB

Base Memory

~5MB

Footprint

Architectures

300KB

20KB

Optimized

X86, TPUs, GPUs

Arm Cortex A, x86



