

Inference Engine: TF vs. TFLite



ML Workflow

this course

next course



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro



TensorFlow



TensorFlow Lite



TensorFlow





Less memory

Less compute power

Only focused on *inference*



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

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

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

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

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

```
graph TD; A[TF vs. TF Lite] --> B[Model]; A --> C[Software]; A --> D[Hardware];
```

Model

Software

Hardware




TF vs. TF Lite

```
graph TD; A[TF vs. TF Lite] --> B[Model]; A --> C[Software]; A --> D[Hardware];
```

Model

Software

Hardware

	 TensorFlow	 TensorFlow Lite	 TensorFlow Lite Micro
Training	Yes	No	No



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro

Training

Yes

No

No

Inference

Yes
*(but inefficient
on edge)*

Yes
(and efficient)

Yes
*(and even
more efficient)*



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



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

```
graph TD; A[TF vs. TF Lite] --> B[Model]; A --> C[Software]; A --> D[Hardware];
```

Model

Software

Hardware



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro

Needs an OS

Yes

Yes

No



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro

Needs an OS

Yes

Yes

No

Memory Mapping
of Models

No

Yes

Yes



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

```
graph TD; A[TF vs. TF Lite] --> B[Model]; A --> C[Software]; A --> D[Hardware];
```

Model

Software

Hardware



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro

Base Binary Size

3MB+

100KB

~10 KB



TensorFlow



TensorFlow Lite



TensorFlow Lite Micro

Base Binary Size

3MB+

100KB

~10 KB

Base Memory
Footprint

~5MB

300KB

20KB



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

