
OP 1
OP 2
A
TensorFlow

Effective TensorFlow for Non-Experts

廖宝华

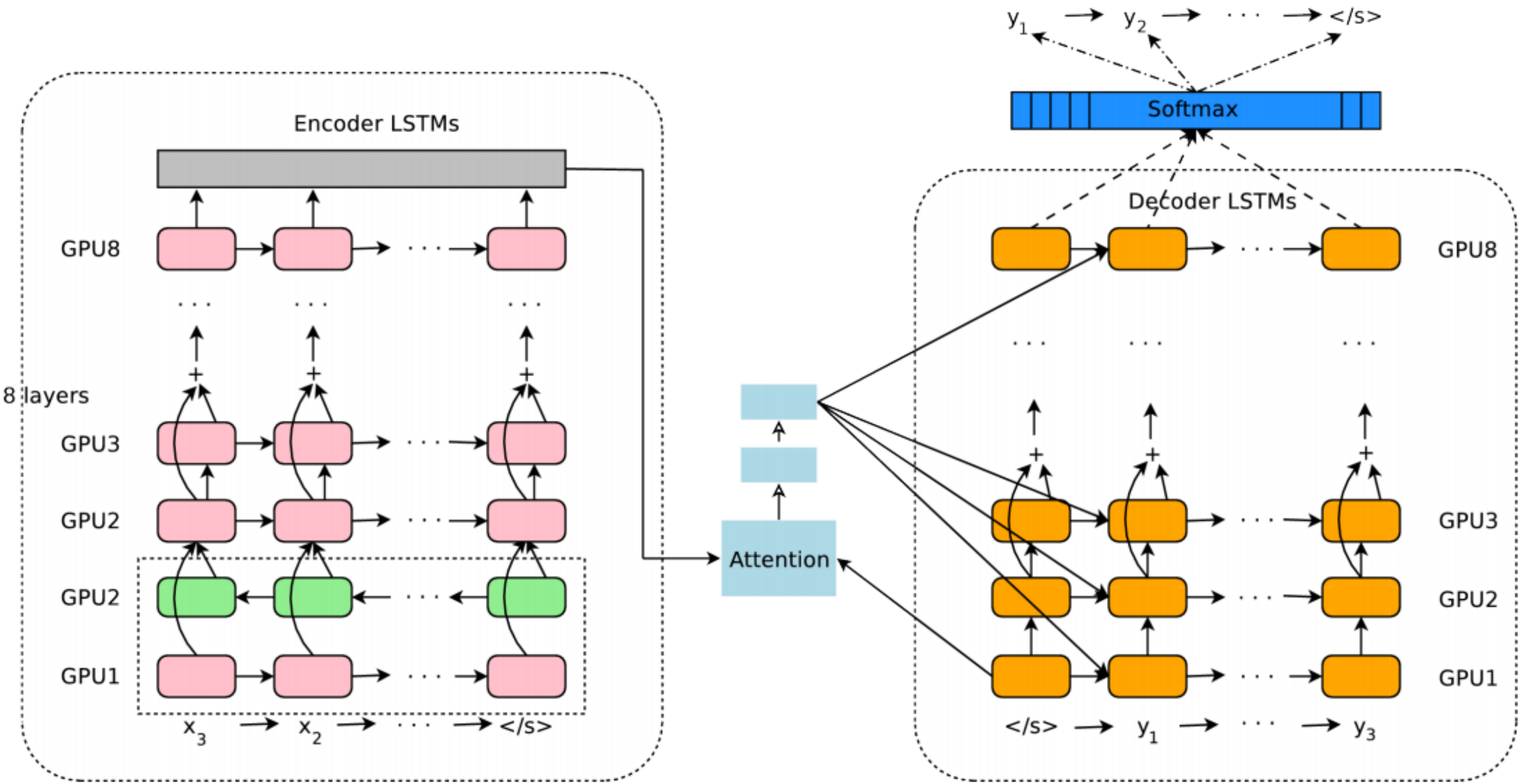
Staff Engineer, Google

Outline

- TensorFlow - ML at Google
- TensorFlow - High-Level API
- TensorFlow - New Features

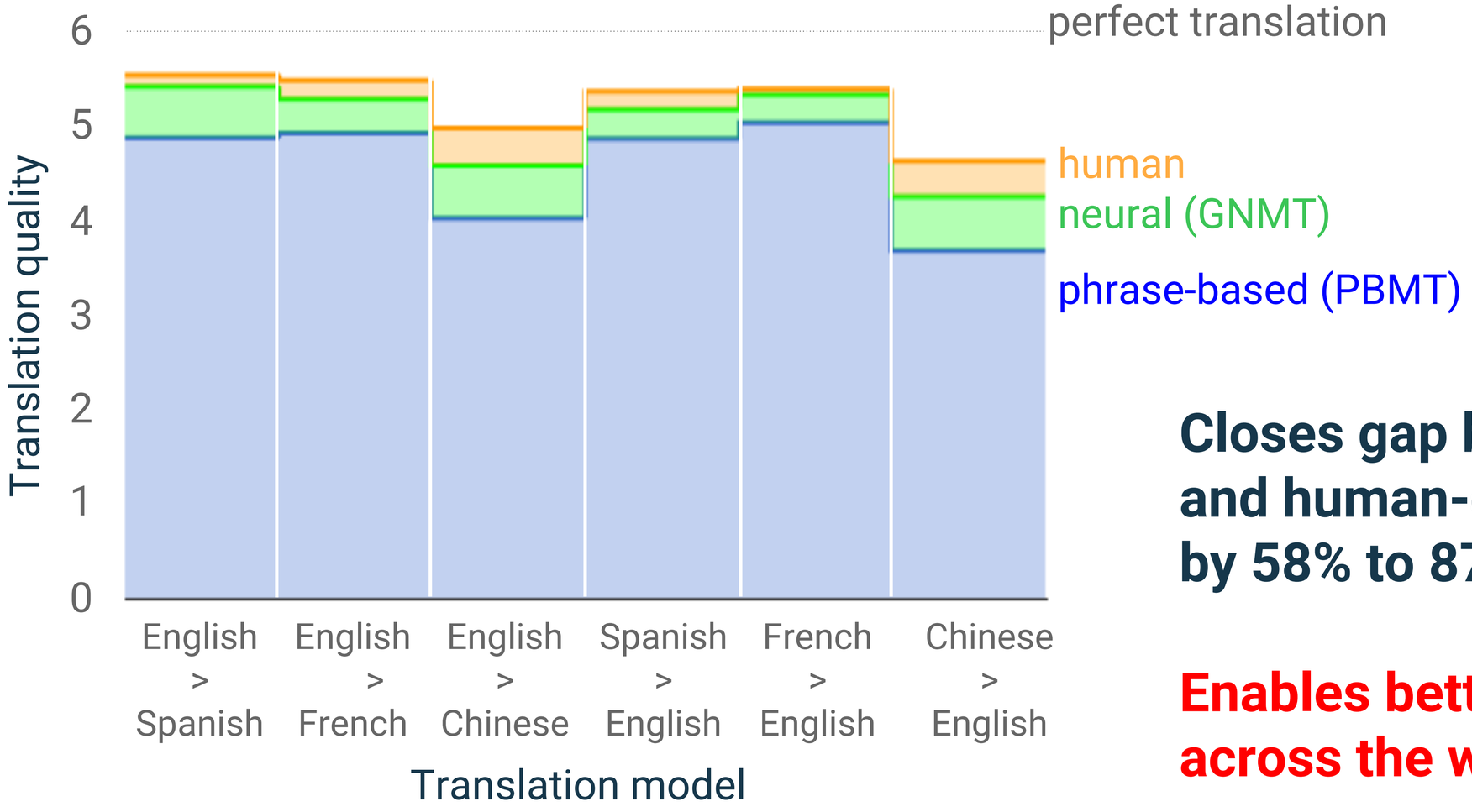
ML at Google

Google Translate


Encoder LSTMs
GPU8
8 layers
GPU3
GPU2
GPU2
GPU1
x3
x2
</s>
Attention
y1
y2
</s>
Softmax
Decoder LSTMs
GPU8
GPU3
GPU2
GPU1
</s>
y1
y3

Wu et al, 2016

Neural Network Translation

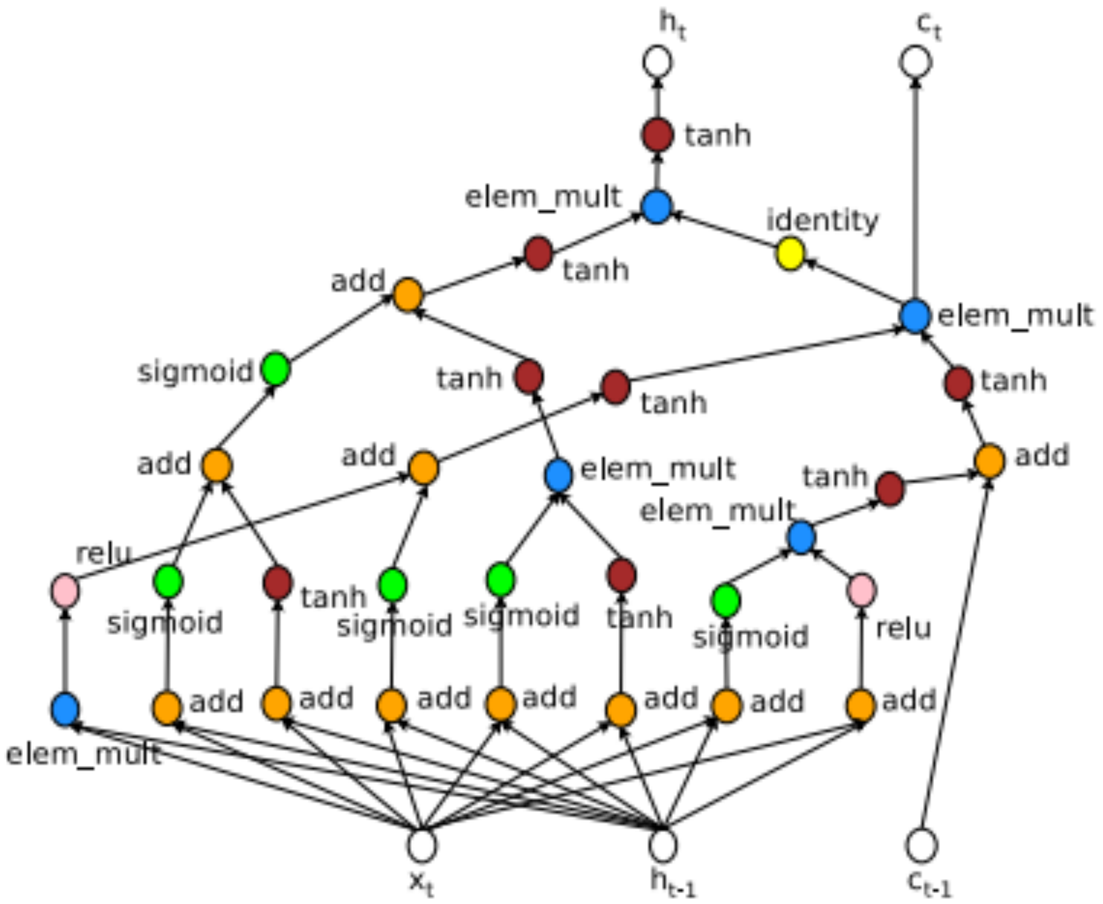


Closes gap between old system and human-quality translation by 58% to 87%

Enables better communication across the world

research.googleblog.com/2016/09/a-neural-network-for-machine.html

Learn to Learn



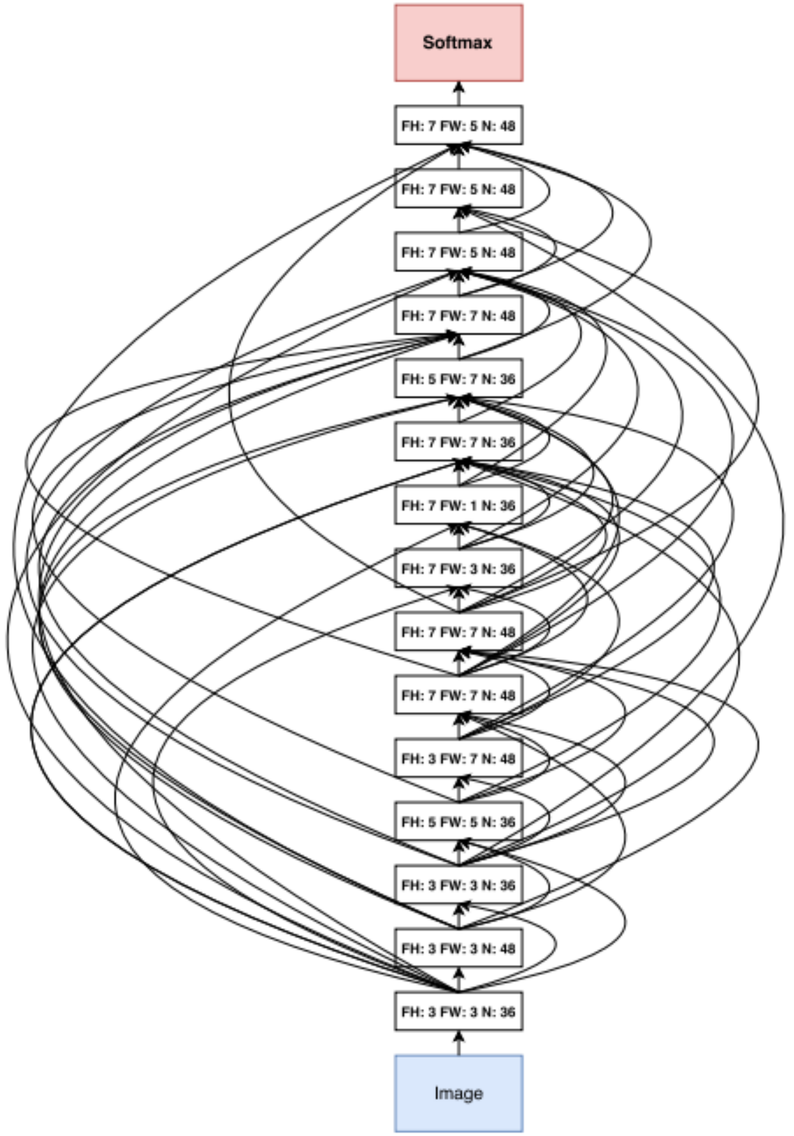


Figure 7: Convolutional architecture discovered by our method, when the search space does not have strides or pooling layers. FH is filter height, FW is filter width and N is number of filters.

Zoph B. & Le, Q. https://arxiv.org/pdf/1611.01578.pdf


Cruise
SPEED LIMIT 35

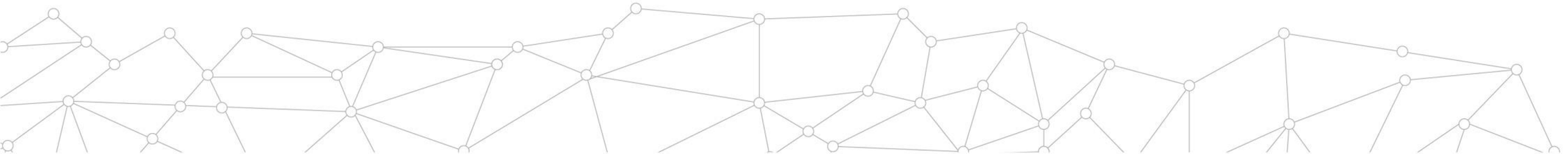
High-Level API

> The first thing to realize about TensorFlow is that it’s a low-level library, meaning you’ll be multiplying matrices and vectors.
>
> Zygmunt Z. on fastml.com

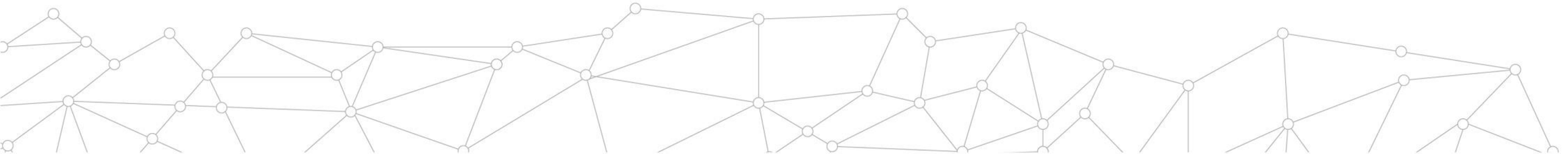
+ Flexible

+ Flexible

+ Extensible



+ Flexible
+ Extensible
+ Maintainable



+ Flexible

+ Extensible

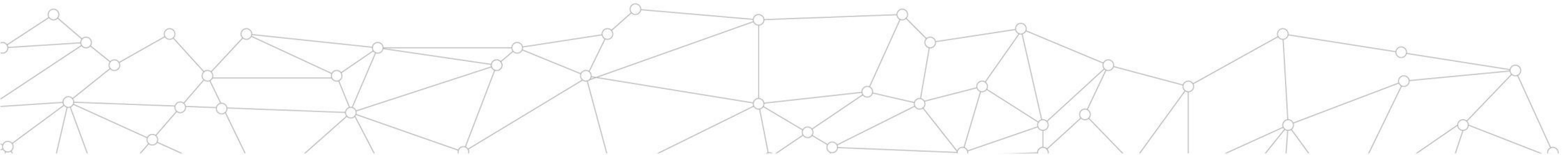
+ Maintainable

✗ Higher-level primitives

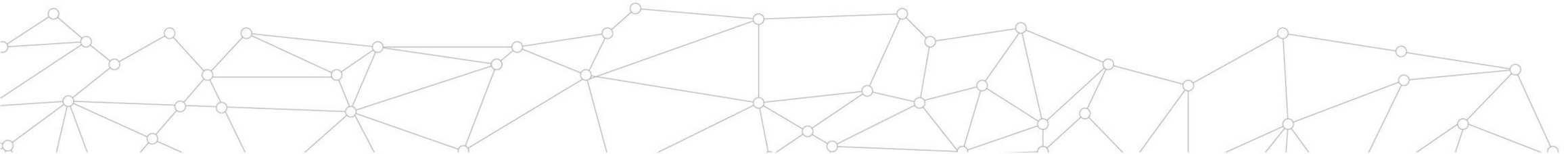
+ Flexible
+ Extensible
+ Maintainable
✗ Higher-level primitives
✗ Out-of-the-box algorithms

Build a TensorFlow high-level API

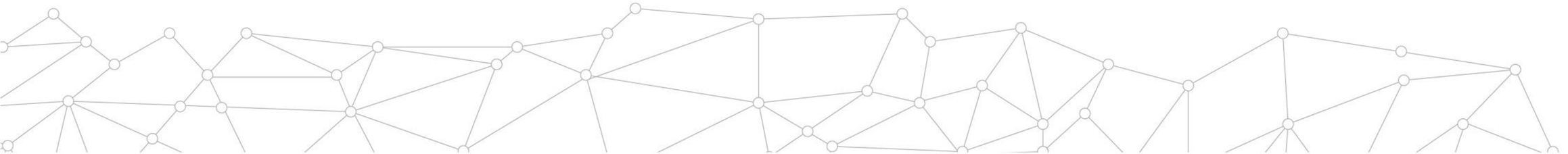
+ Fast iteration



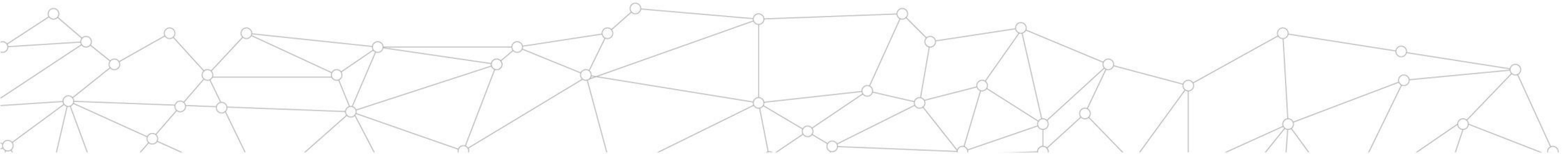
+ Fast iteration
+ Encode best practices

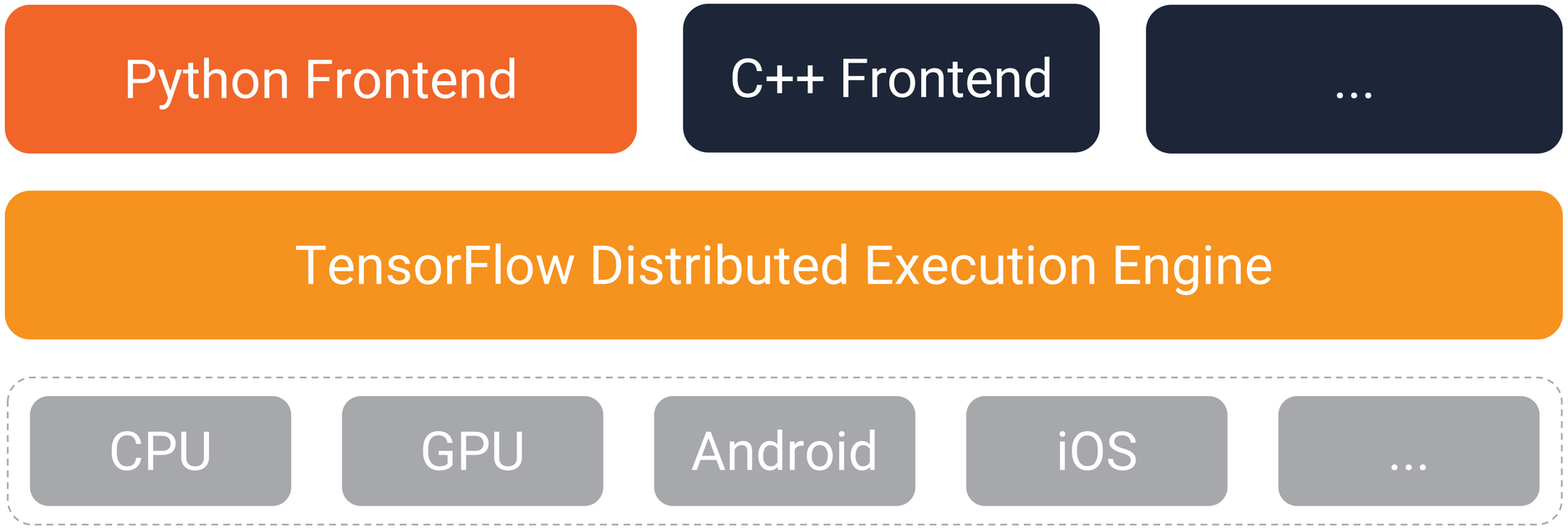


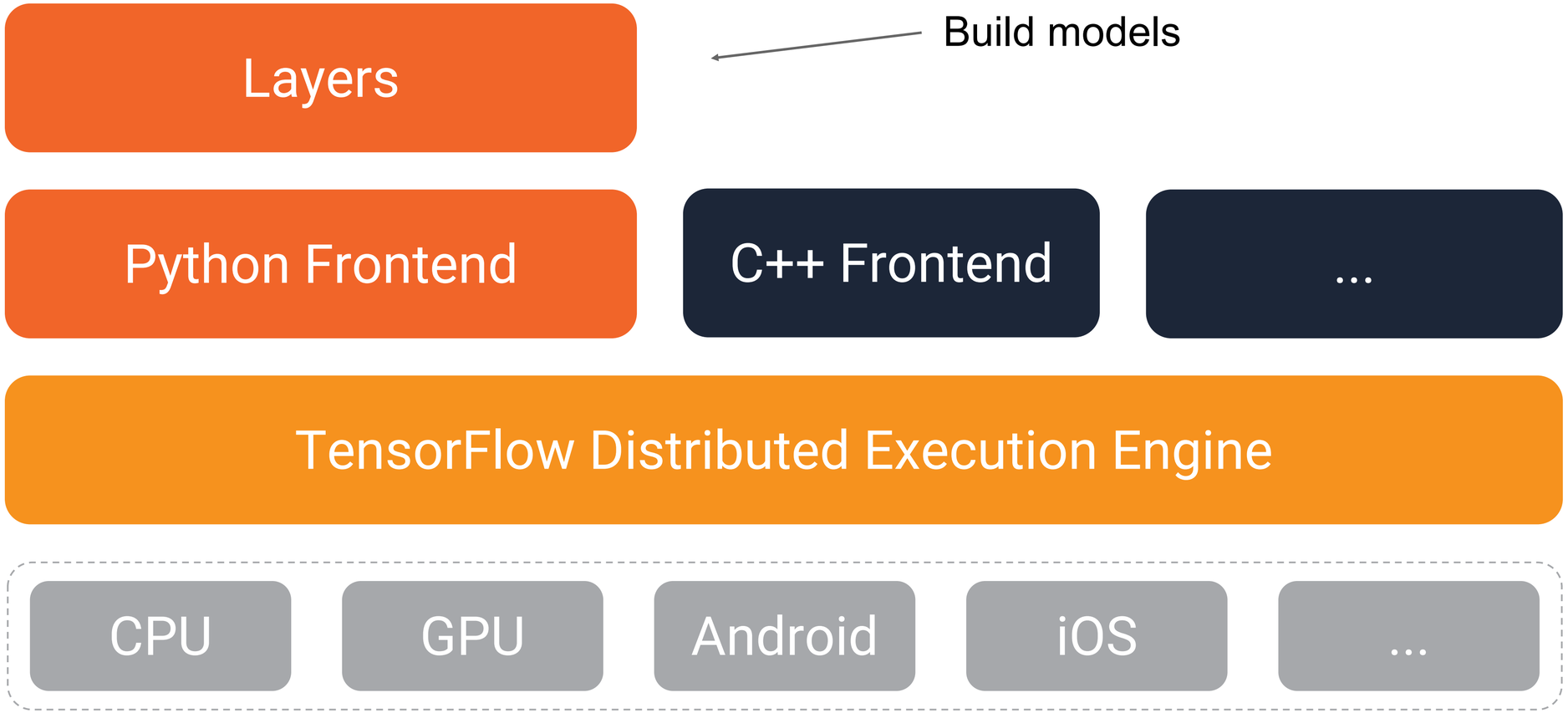
+ Fast iteration
+ Encode best practices
+ Build-in scalability

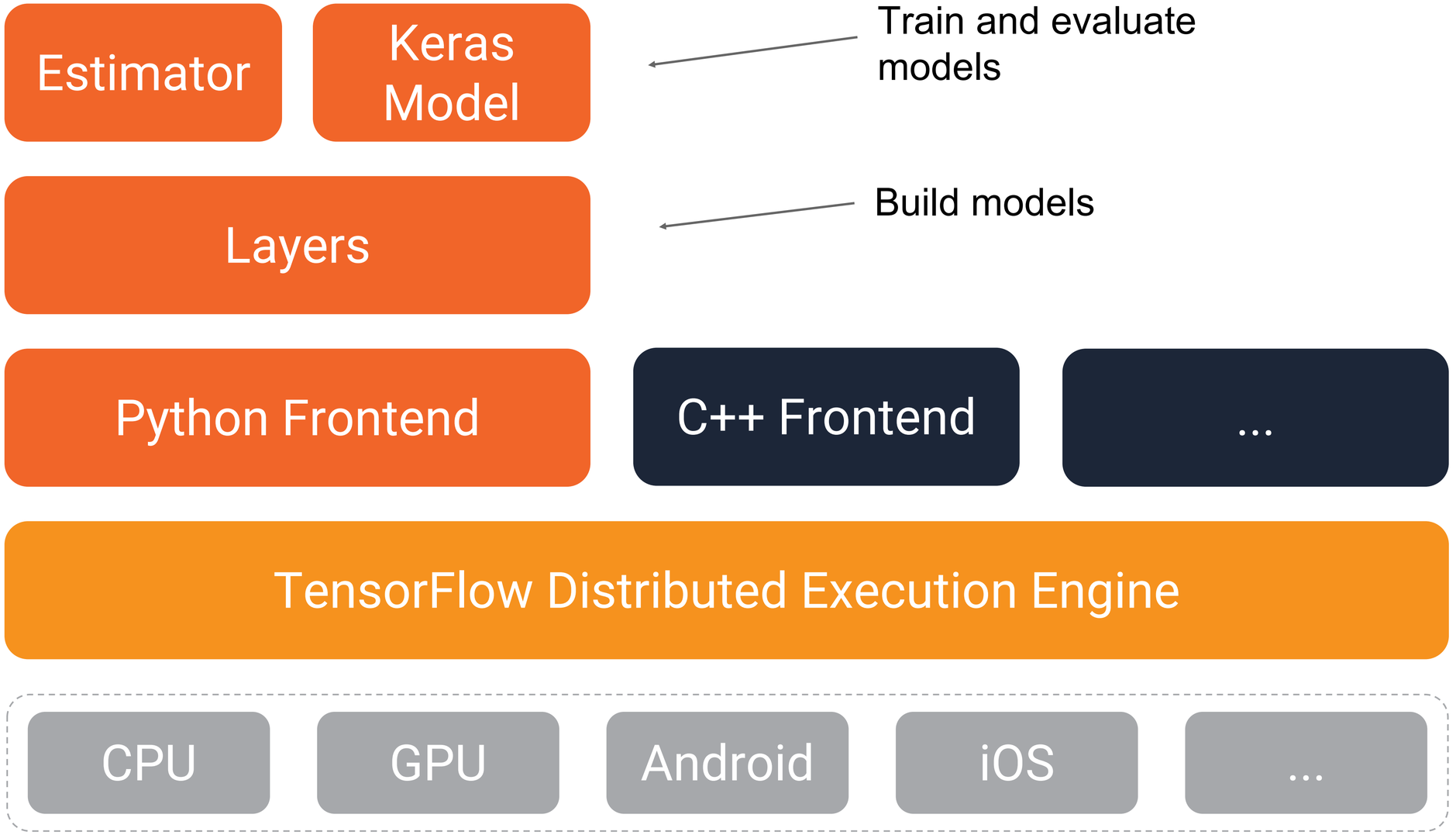


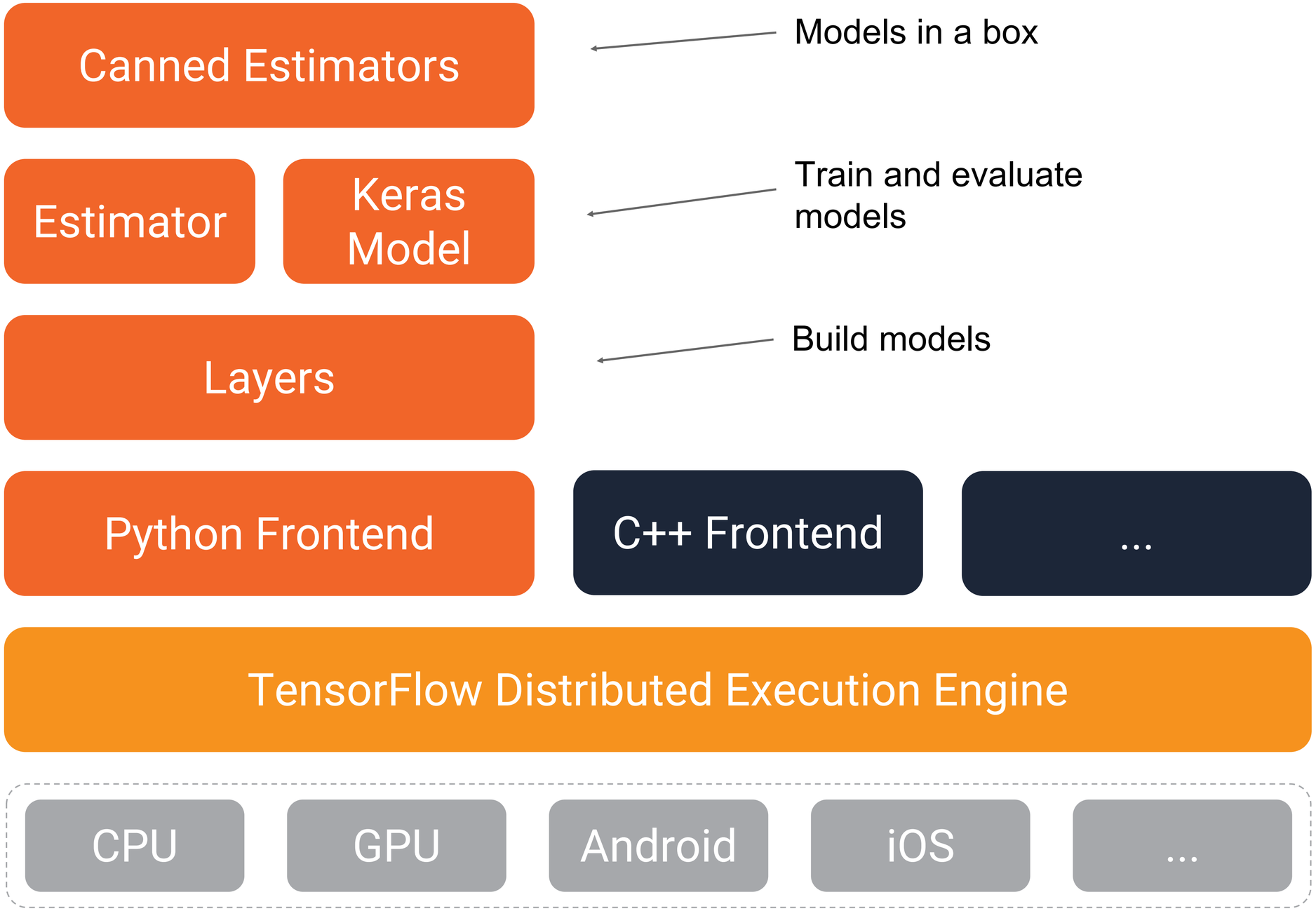
+ Fast iteration
+ Encode best practices
+ Build-in scalability
+ Simple deployment

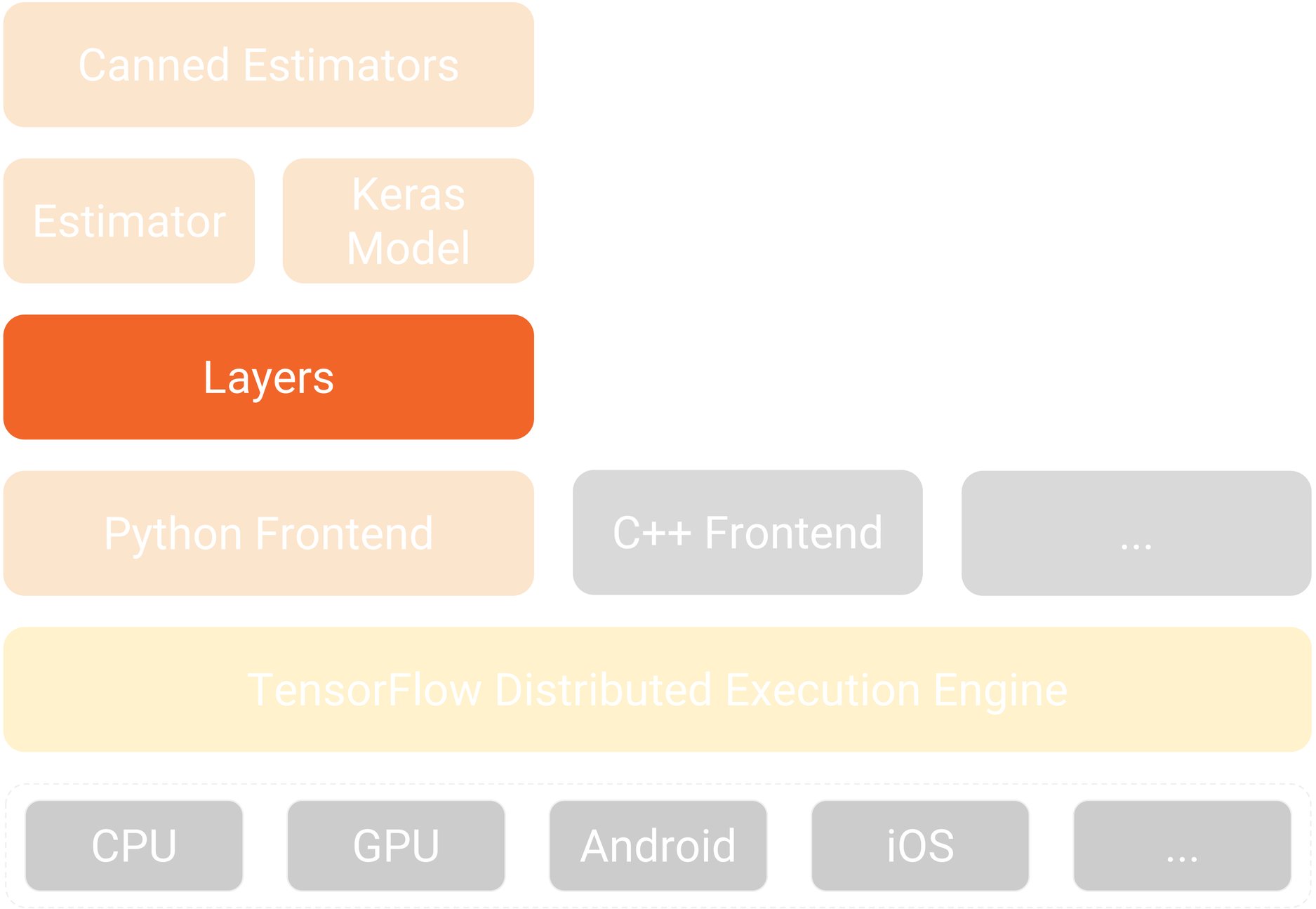



Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...


Layers
Build models
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...


Estimator
Keras Model
Train and evaluate models
Layers
Build models
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...


Canned Estimators
Models in a box
Estimator
Keras Model
Train and evaluate models
Layers
Build models
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...

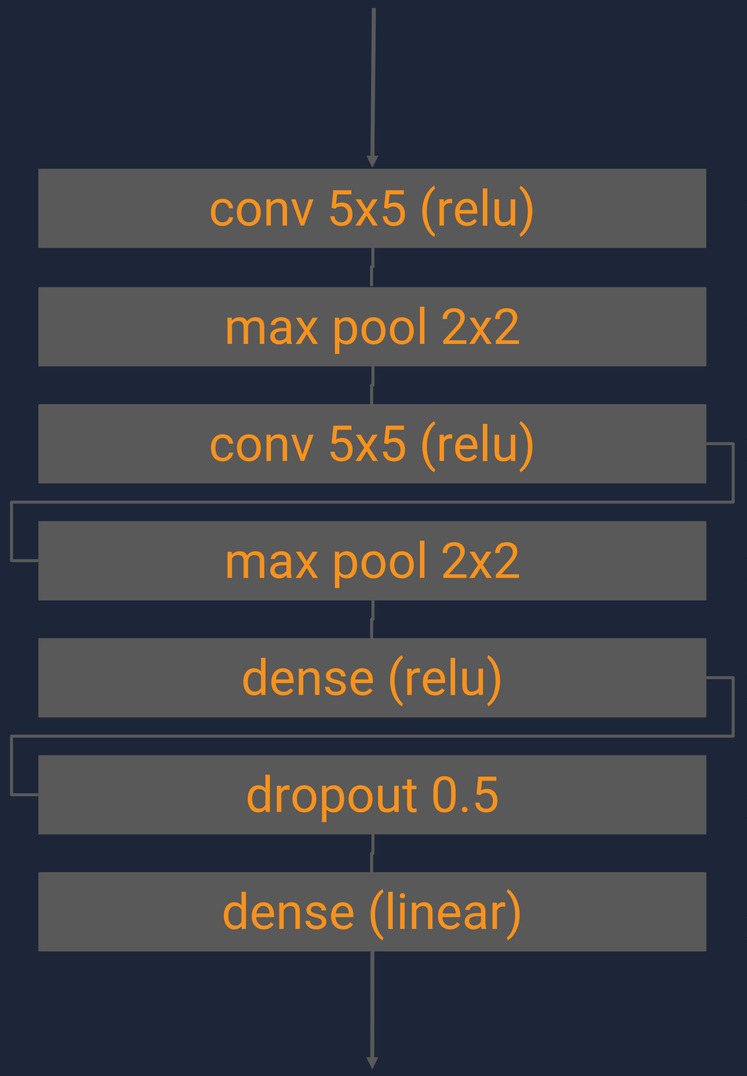

Canned Estimators
Estimator
Keras Model
Layers
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...


conv 5x5 (relu)
max pool 2x2
conv 5x5 (relu)
max pool 2x2
dense (relu)
dropout 0.5
dense (linear)

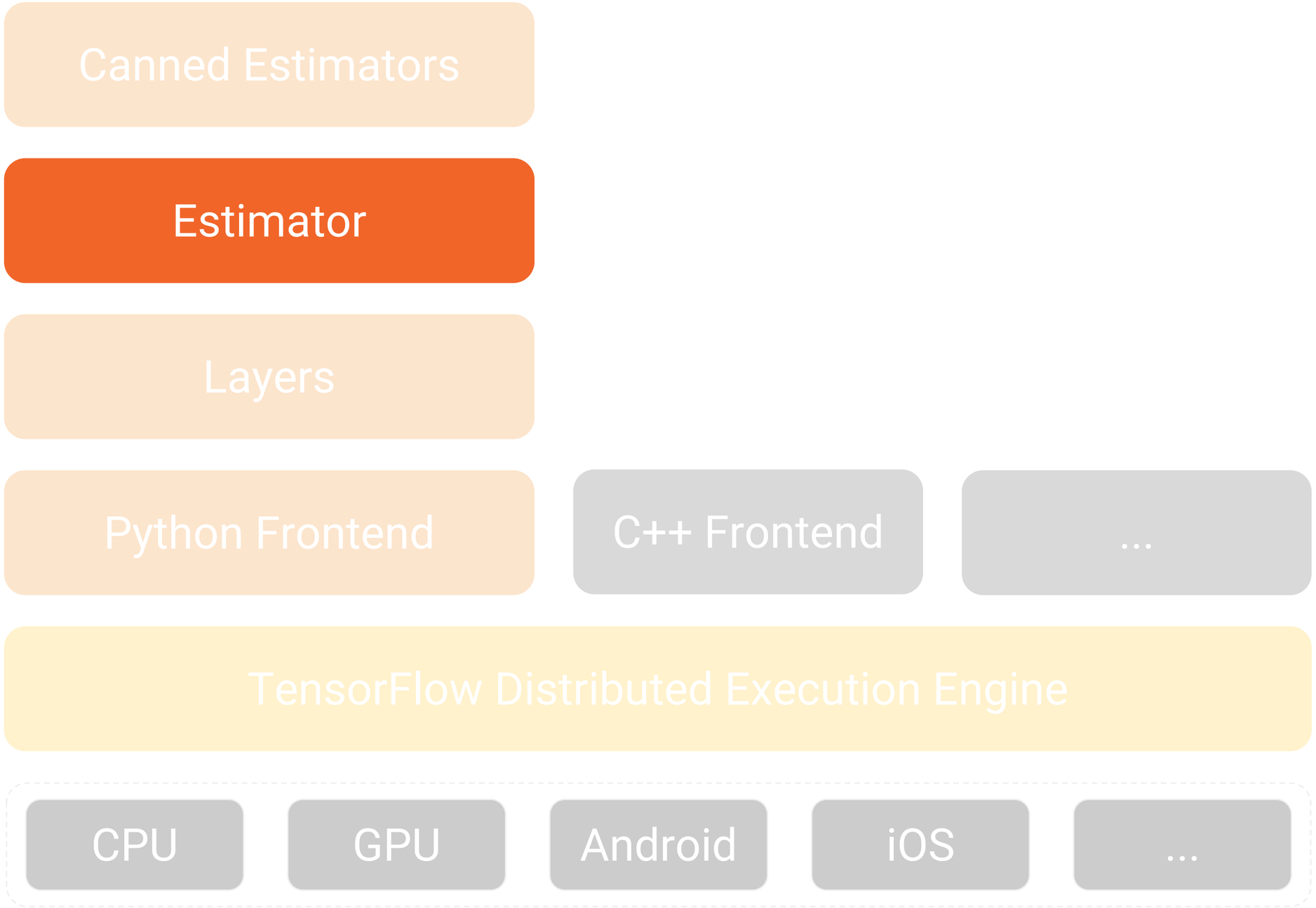

```
x = tf.layers.conv2d(x, kernel_size=[5,5], ...)
```

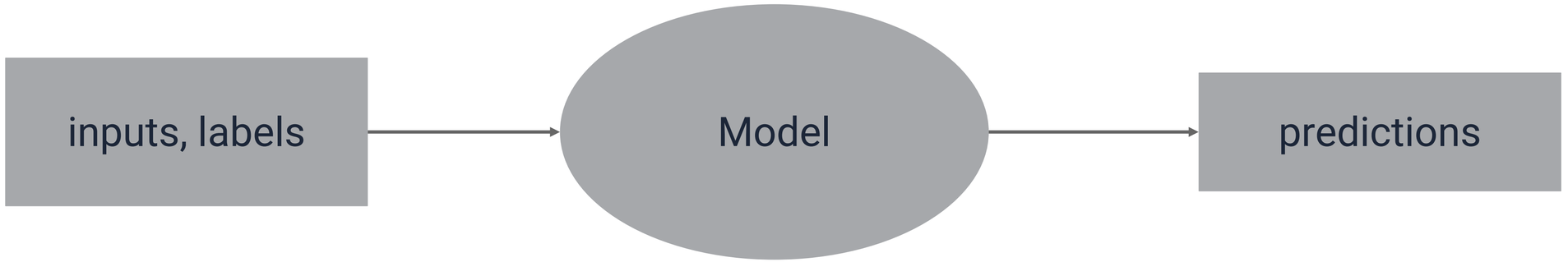

```
x = tf.layers.conv2d(x, kernel_size=[5,5], ...)
```

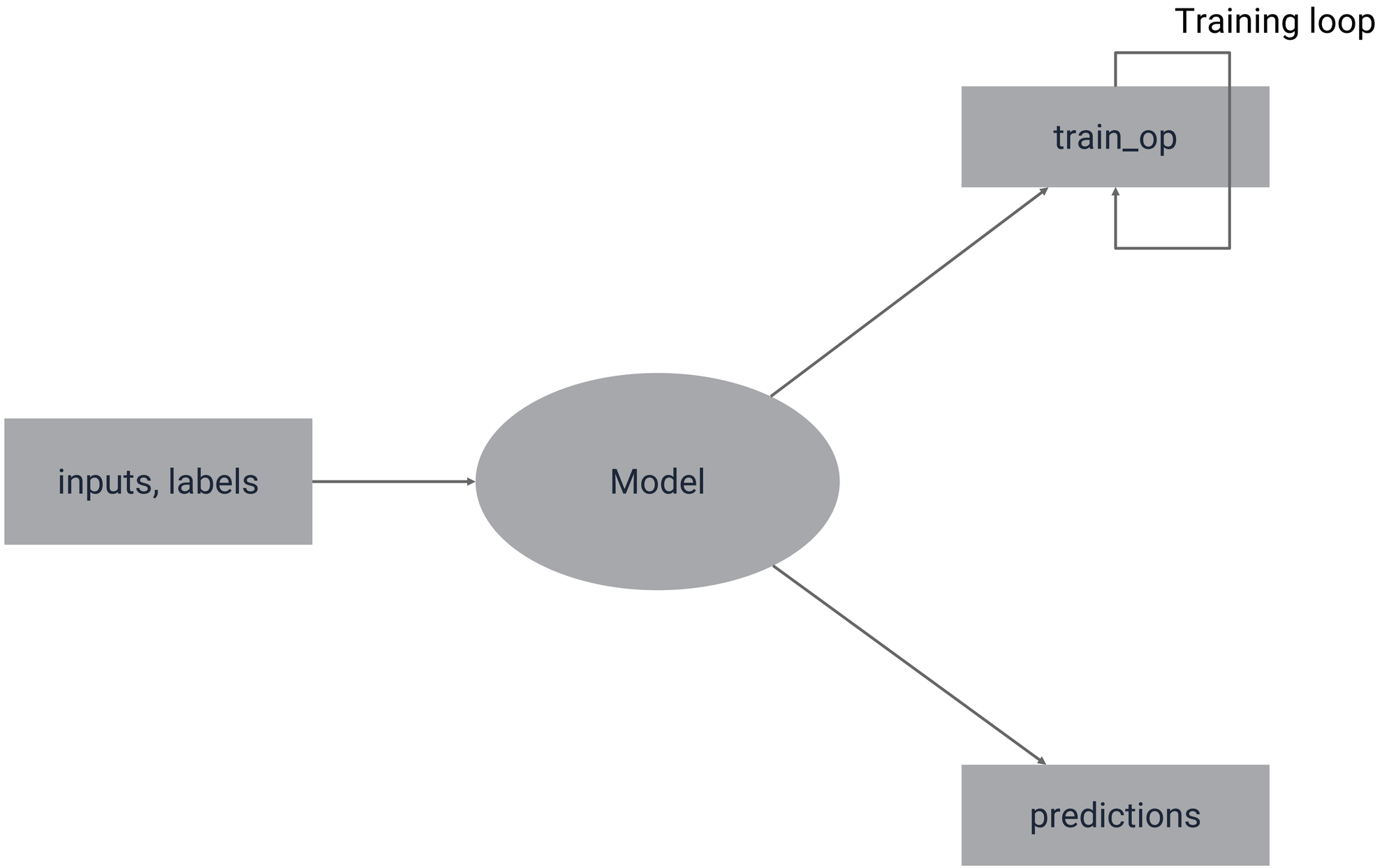
```
x = tf.layers.max_pooling2d(x, kernel_size=[2,2], ...)
```

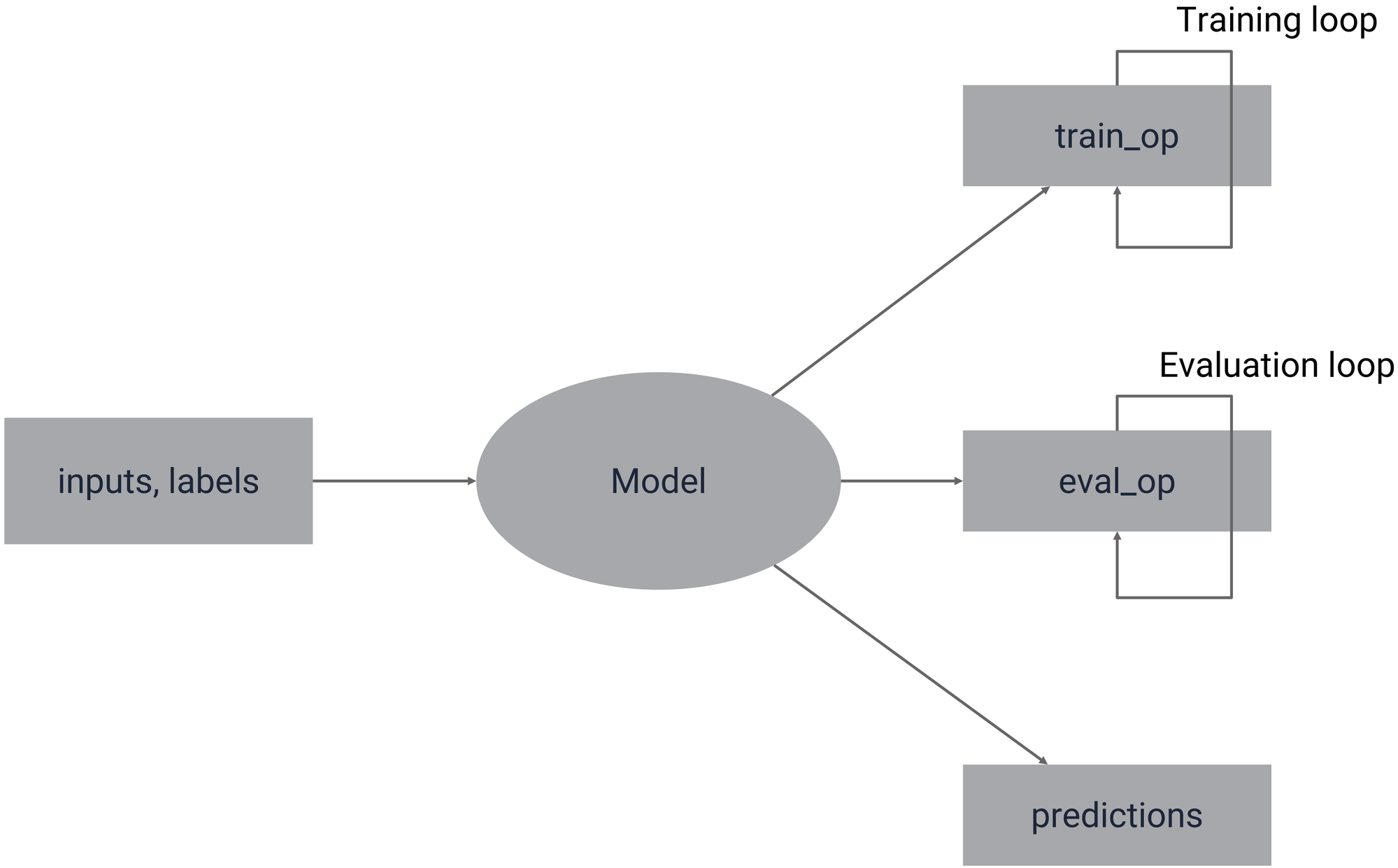


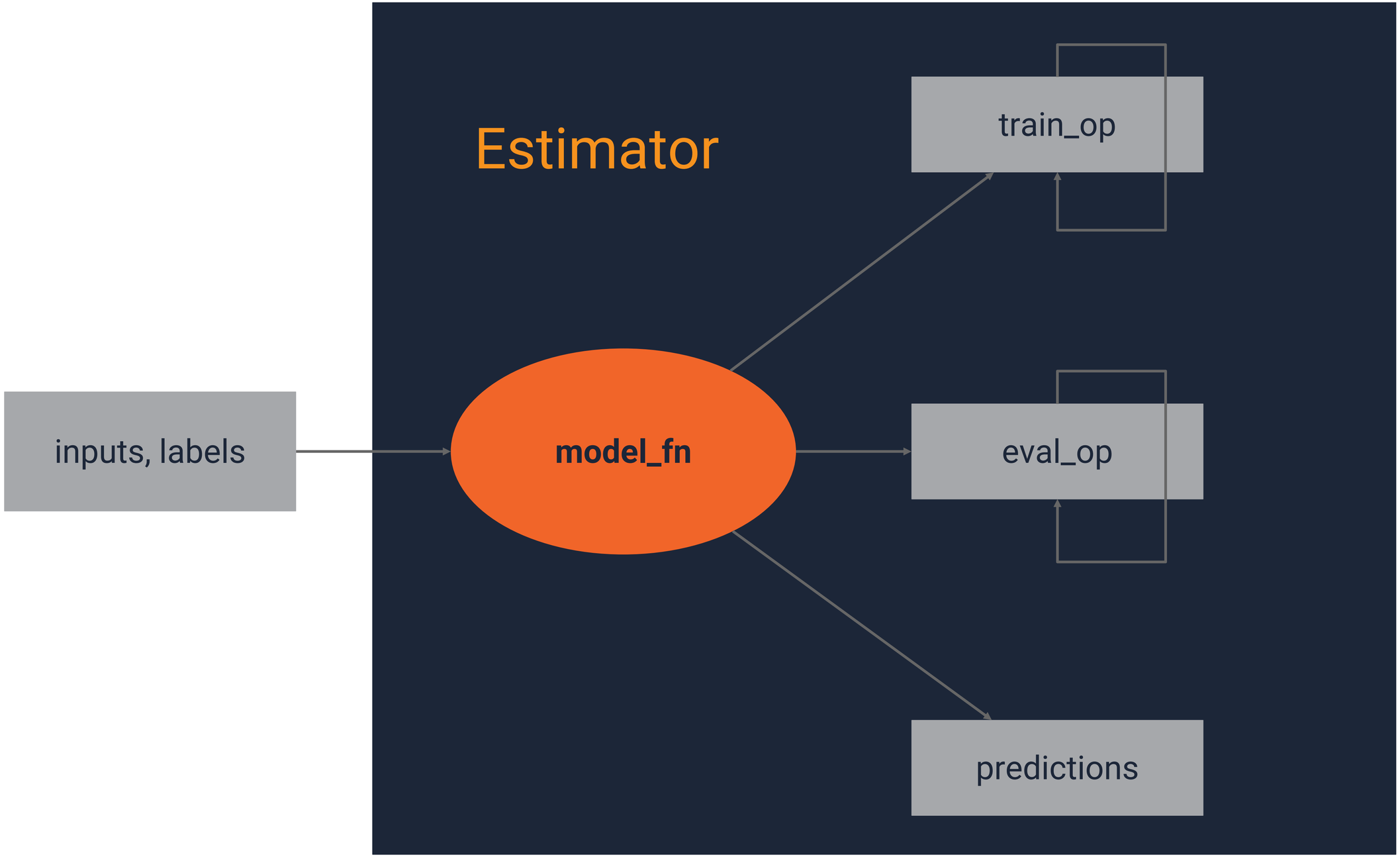

```
x = tf.layers.conv2d(x, kernel_size=[5,5], ...)
x = tf.layers.max_pooling2d(x, kernel_size=[2,2], ...)
x = tf.layers.conv2d(x, kernel_size=[5,5], ...)
x = tf.layers.max_pooling2d(x, kernel_size=[2,2], ...)
x = tf.layers.dense(x, activation_fn=tf.nn.relu)
x = tf.layers.dropout(x, 0.5)
x = tf.layers.dense(x)
```

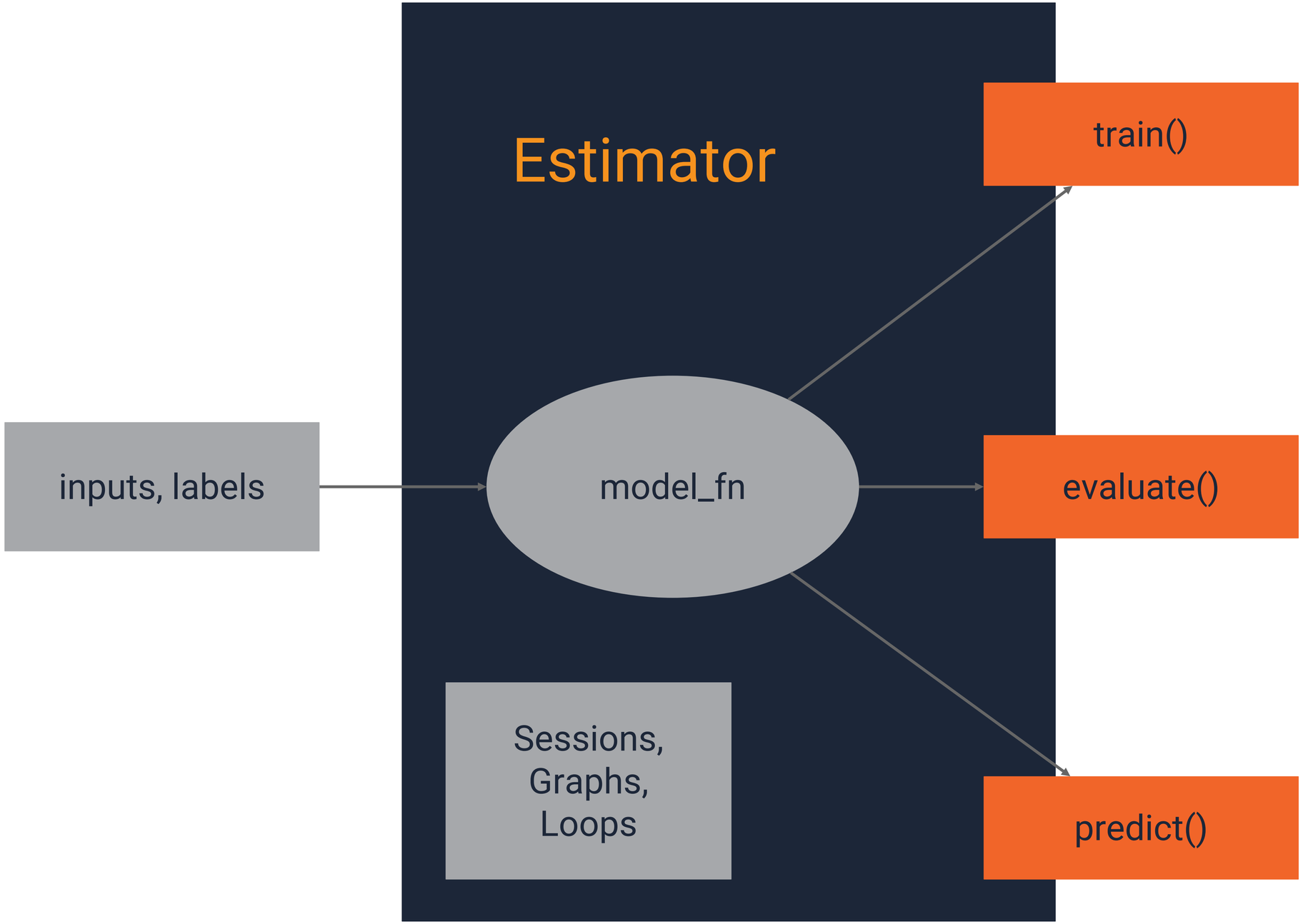

Canned Estimators
Estimator
Layers
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...

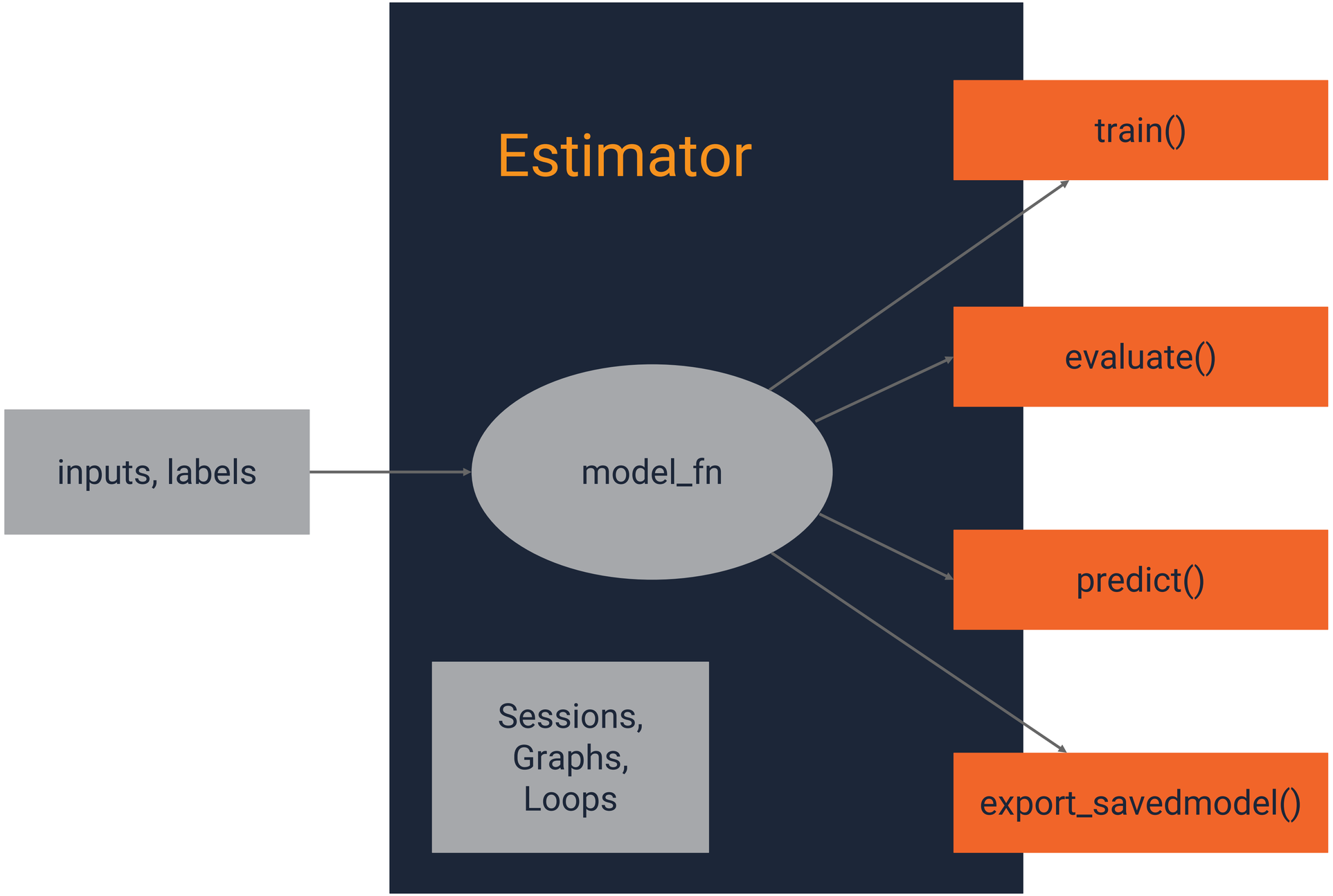

inputs, labels
Model
predictions


Training loop
train_op
inputs, labels
Model
predictions

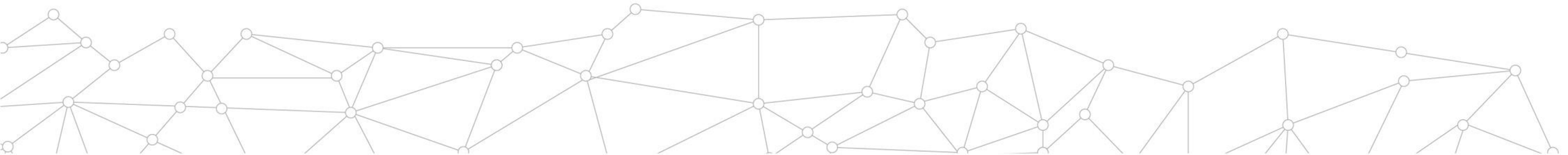

Training loop
train_op
Evaluation loop
inputs, labels
Model
eval_op
predictions


Estimator
train_op
inputs, labels
model_fn
eval_op
predictions

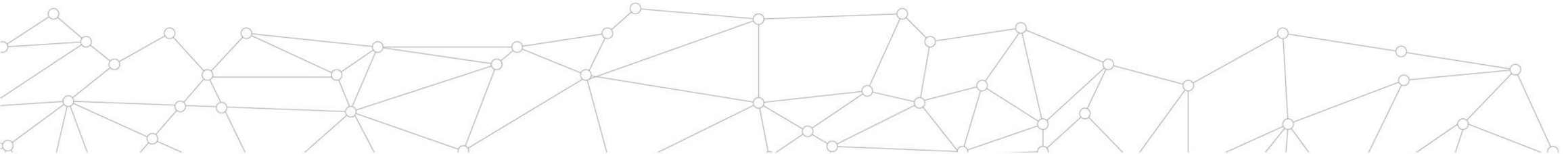

Estimator
train()
inputs, labels
model_fn
evaluate()
Sessions,
Graphs,
Loops
predict()


Estimator
train()
evaluate()
inputs, labels
model_fn
predict()
Sessions,
Graphs,
Loops
export_savedmodel()

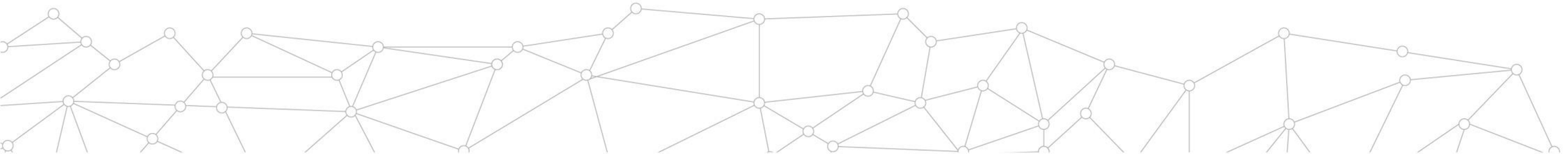
+ Encodes best practices

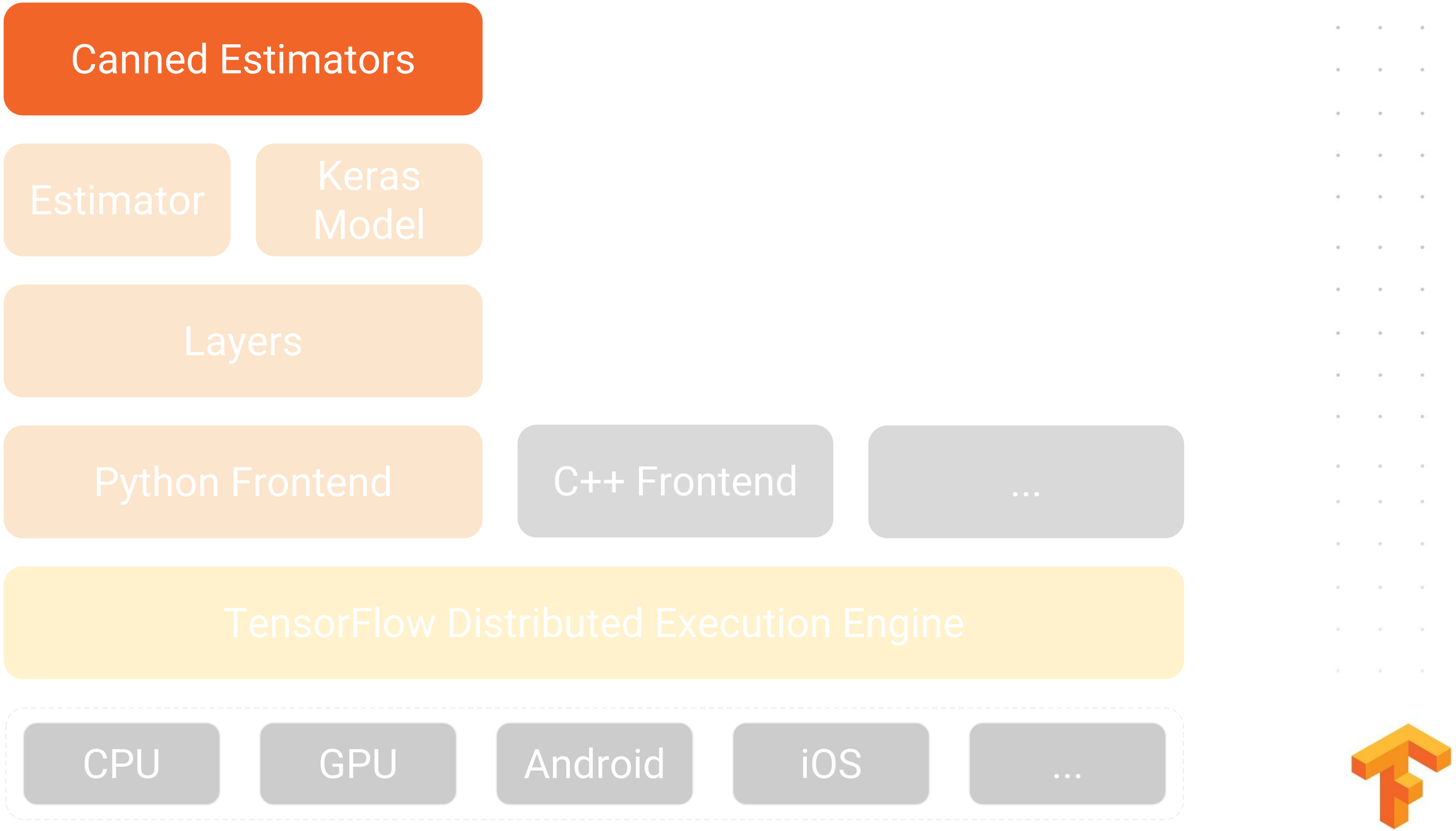


+ Encodes best practices
+ Deploy with TensorFlow Serving



+ Encodes best practices
+ Deploy with TensorFlow Serving
+ Distributed/Scalable by design




Canned Estimators
Estimator
Keras Model
Layers
Python Frontend
C++ Frontend
...
TensorFlow Distributed Execution Engine
CPU
GPU
Android
iOS
...

```
area = real_valued_column(“square_foot”),
rooms = real_valued_column(“num_rooms”),
zip_code = sparse_column_with_integerized_feature(“zip_code”, 100000)


regressor = LinearRegressor(feature_columns=[area, rooms, zip_code], ...)



regressor.train(train_input_fn)

regressor.evaluate(eval_input_fn)
```

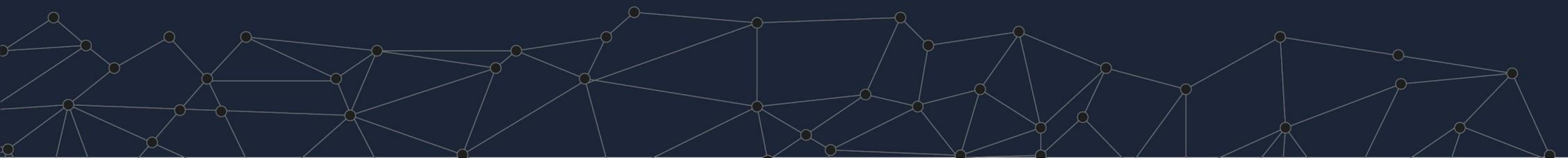
```
area = real_valued_column("square_foot"),
rooms = real_valued_column("num_rooms"),
zip_code = sparse_column_with_integerized_feature("zip_code", 100000)


regressor = LinearRegressor(feature_columns=[area, rooms, zip_code], ...)



regressor.train(train_input_fn)

regressor.evaluate(eval_input_fn)
```

```
area = real_valued_column(“square_foot”),
rooms = real_valued_column(“num_rooms”),
zip_code = sparse_column_with_integerized_feature(“zip_code”, 100000)


regressor = DNNRegressor(
    feature_columns=[area, rooms, embedding_column(zip_code, 8)],
    hidden_units=[1024, 512, 256])


regressor.train(train_input_fn)

regressor.evaluate(eval_input_fn)
```

```
area = real_valued_column(“square_foot”),
rooms = real_valued_column(“num_rooms”),
zip_code = sparse_column_with_integerized_feature(“zip_code”, 100000)


regressor = DNNRegressor(
    feature_columns=[area, rooms, embedding_column(zip_code, 8)],
    hidden_units=[1024, 512, 256])


regressor.train(train_input_fn)

regressor.evaluate(eval_input_fn)
```

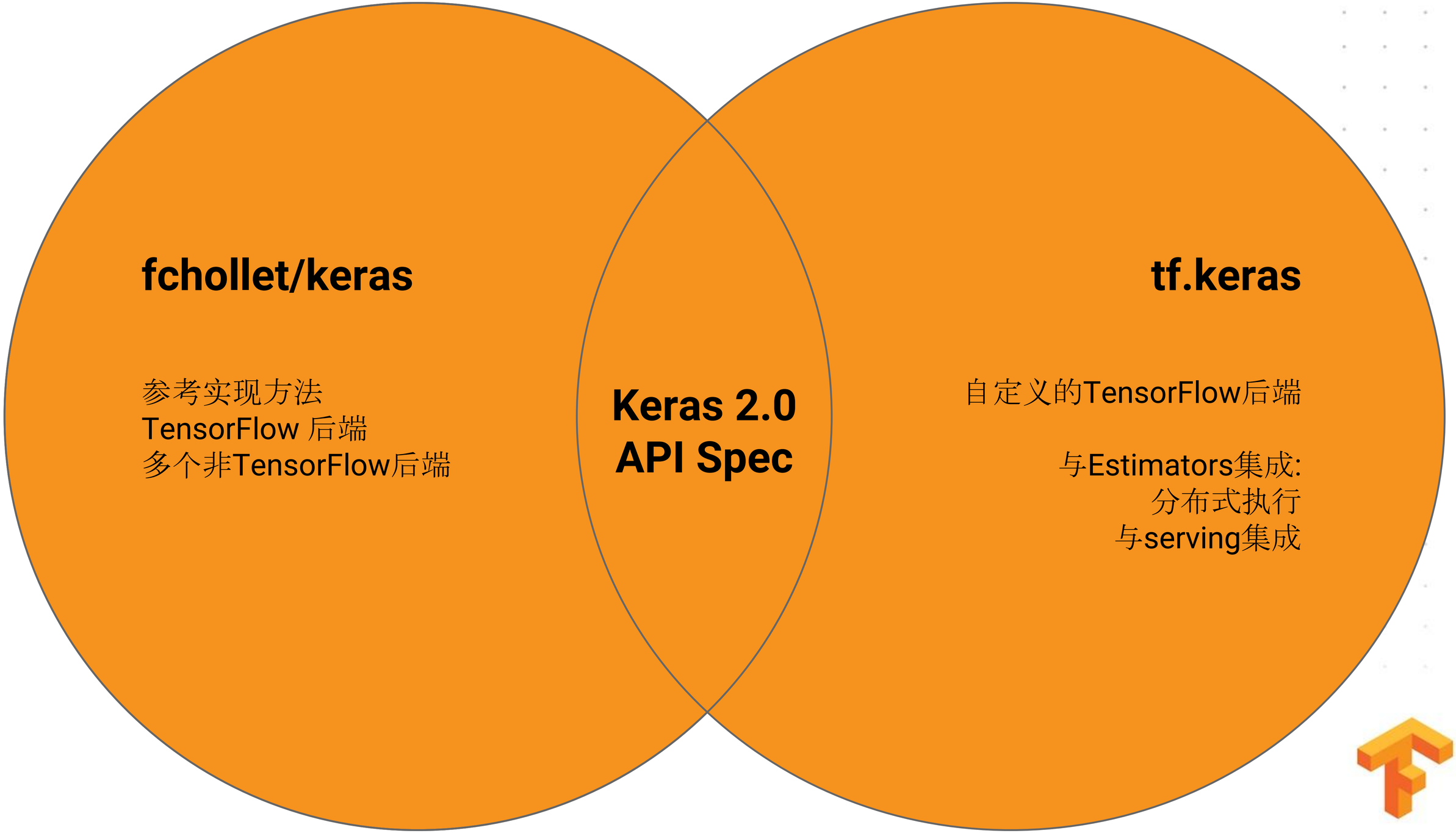
Roadmap

	1.0	1.1	1.2
core	Layers	Layers Estimator	Layers Estimator Canned Estimators
contrib	Estimator Canned Estimators	Canned Estimators	More Canned Estimators



New Features

TensorFlow & Keras


fchollet/keras
参考实现方法
TensorFlow 后端
多个非TensorFlow后端
Keras 2.0
API Spec
tf.keras
自定义的TensorFlow后端
与Estimators集成:
分布式执行
与serving集成

Check out...

tf.keras

tf.layers

tf.estimator

New Programmer's Guide:

www.tensorflow.org/programmers_guide

BlogPost: https://goo.gl/RyLuUw

Eager Execution

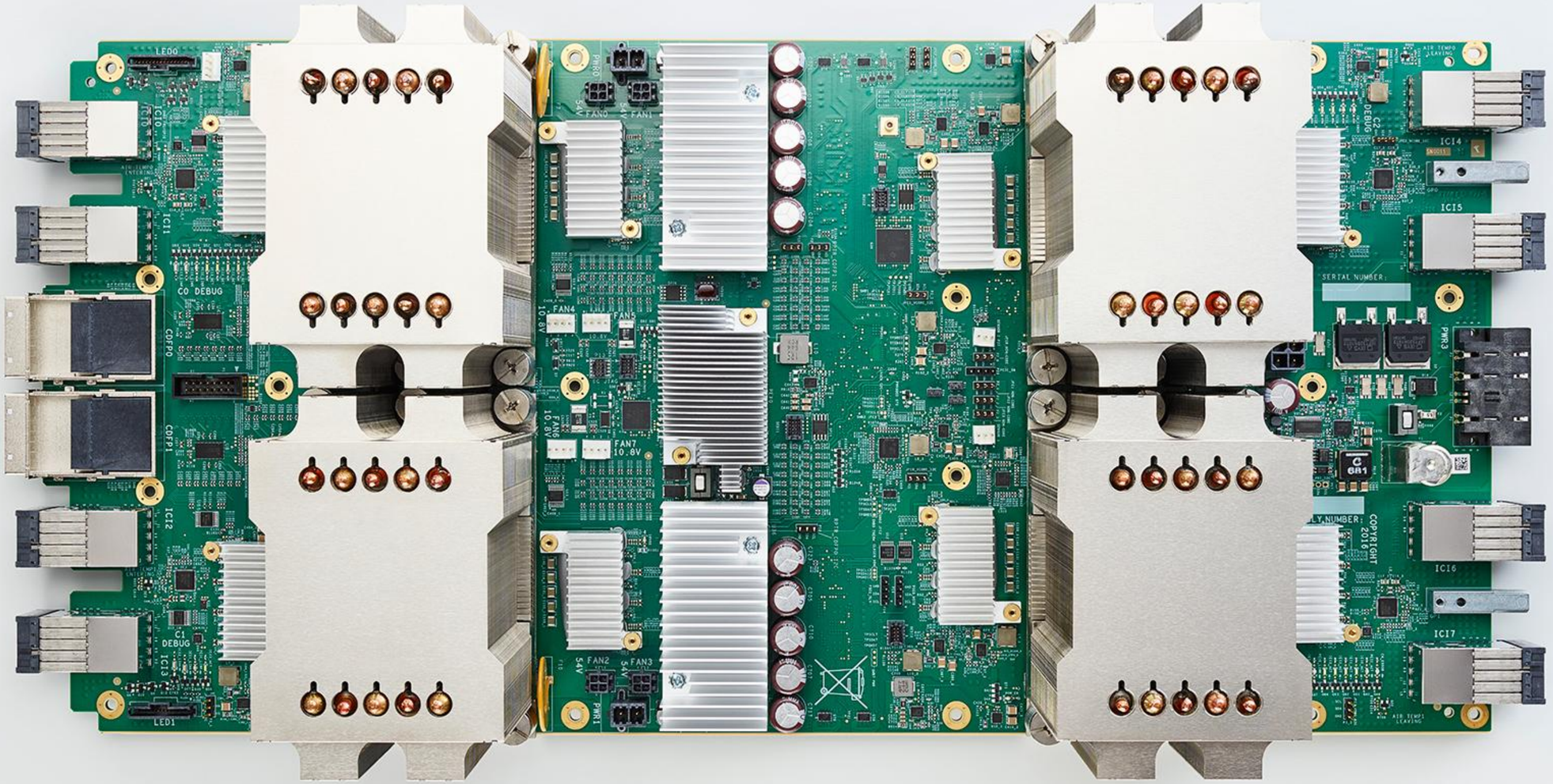
As simple as possible

Simple, Instant error

```
a = tf.constant(6)
while a != 1:
  if a % 2 == 0:
    a = a / 2
  else:
    a = 3 * a + 1
  print(a)
```

```
# Outputs
tf.Tensor(3, dtype=int32)
tf.Tensor(10, dtype=int32)
tf.Tensor(5, dtype=int32)
tf.Tensor(16, dtype=int32)
tf.Tensor(8, dtype=int32)
tf.Tensor(4, dtype=int32)
tf.Tensor(2, dtype=int32)
tf.Tensor(1, dtype=int32)
```

2nd Generation TPU


LED0
IC10
IC11
C0 DEBUG
CDFP0
CDFP1
IC12
C1 DEBUG
IC13
LED1
PWR0
54V
FAN0
FAN1
FAN4
10.8V
FAN5
FAN6
10.8V
FAN7
10.8V
FAN2
54V
FAN3
PWR1
AIR TEMP1 LEAVING
C2 DEBUG
IC14
IC15
SERIAL NUMBER:
PWR3
IC16
IC17
COPYRIGHT 2016

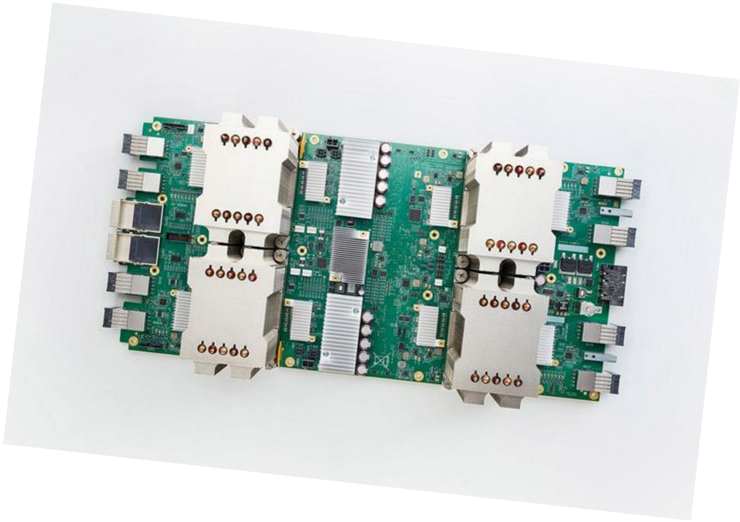
性能预览

ResNet-50 训练在 20 小时内
达到 ~74% 准确度

TensorFlow Lite

Trends

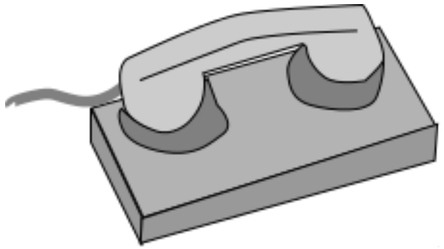
More on-device ML



More ML hardware

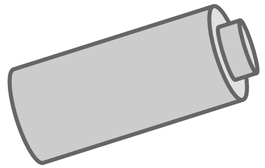
Why on-device?

Offline



Low-bandwidth

Latency



Power

Challenges

Resources

Bandwidth

Memory

Computation

Heterogeneity

GPUs

CPUs

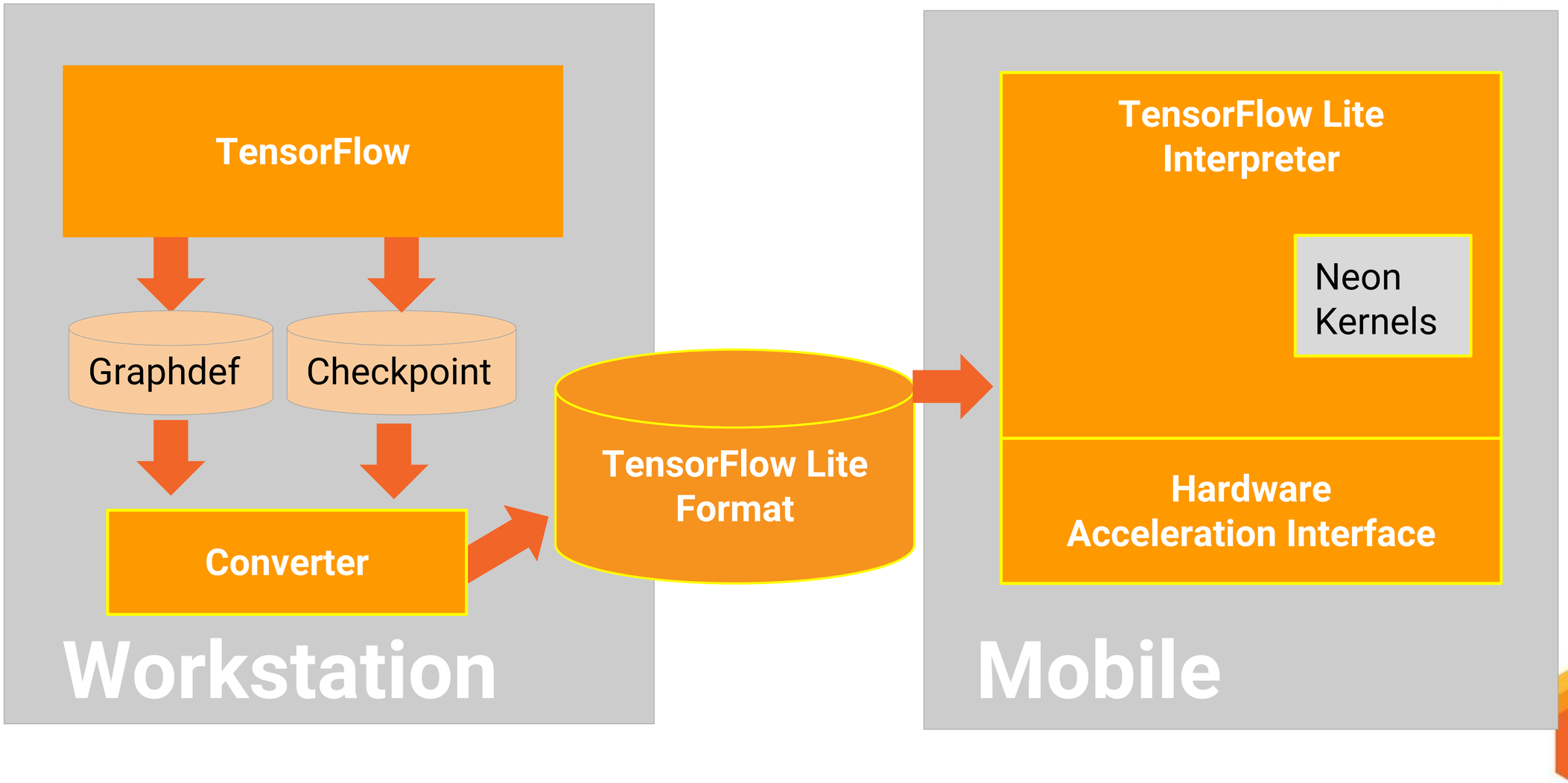
DSPs

...

TensorFlow

TensorFlow works well on **large** devices.
TensorFlow Lite is focused on **small** devices

TensorFlow Lite



Tf.data

tf.data 模块

tf.data.Dataset
Represents input pipeline using functional transformations

tf.data.Iterator
Provides sequential access to elements of a Dataset

open source machine learning

tensorflow.google.cn/