Model Deployments with TensorFlow Serving

Snehan Kekre
Coursera

coursera rhyme

Most models don’t get deployed.


https://dineshnirmal.wordpress.com
The Five Pillars of Fluid ML
Featured
Leave a comment
A few months ago, I was talking with the CTO of a major bank about machine learning. At one point he shook his head ruefully and said, "Dinesh, it only took me 3 weeks to develop a model. It's been 11 months, and we still haven't deployed it."

Example of inefficient model deployment

```python
import json
from flask import Flask
from keras.models import load_model
from utils import preprocess

model = load_model('model.h5')
app = Flask(__name__)

@app.route('/classify', methods=['POST'])
def classify():
    review = request.form["review"]
    preprocessed_review = preprocess(review)
    prediction = model.predict_classes([preprocessed_review])[0]
    return json.dumps({"score": int(prediction)})
```

Why Flask is insufficient

- Lack of consistent APIs
- Lack of consistent payloads
- Lack of model versioning
- Lack of mini-batching support
- Inefficient for large models

```
import json
from flask import Flask
from keras.models import load_model
from utils import preprocess

model = load_model('model.h5')
app = Flask(__name__)

@app.route('/classify', methods=['POST'])
def classify():
    review = request.form["review"]
    preprocessed_review = preprocess(review)
    prediction = model.predict_classes([preprocessed_review])[0]
    return json.dumps({"score": int(prediction)})
```

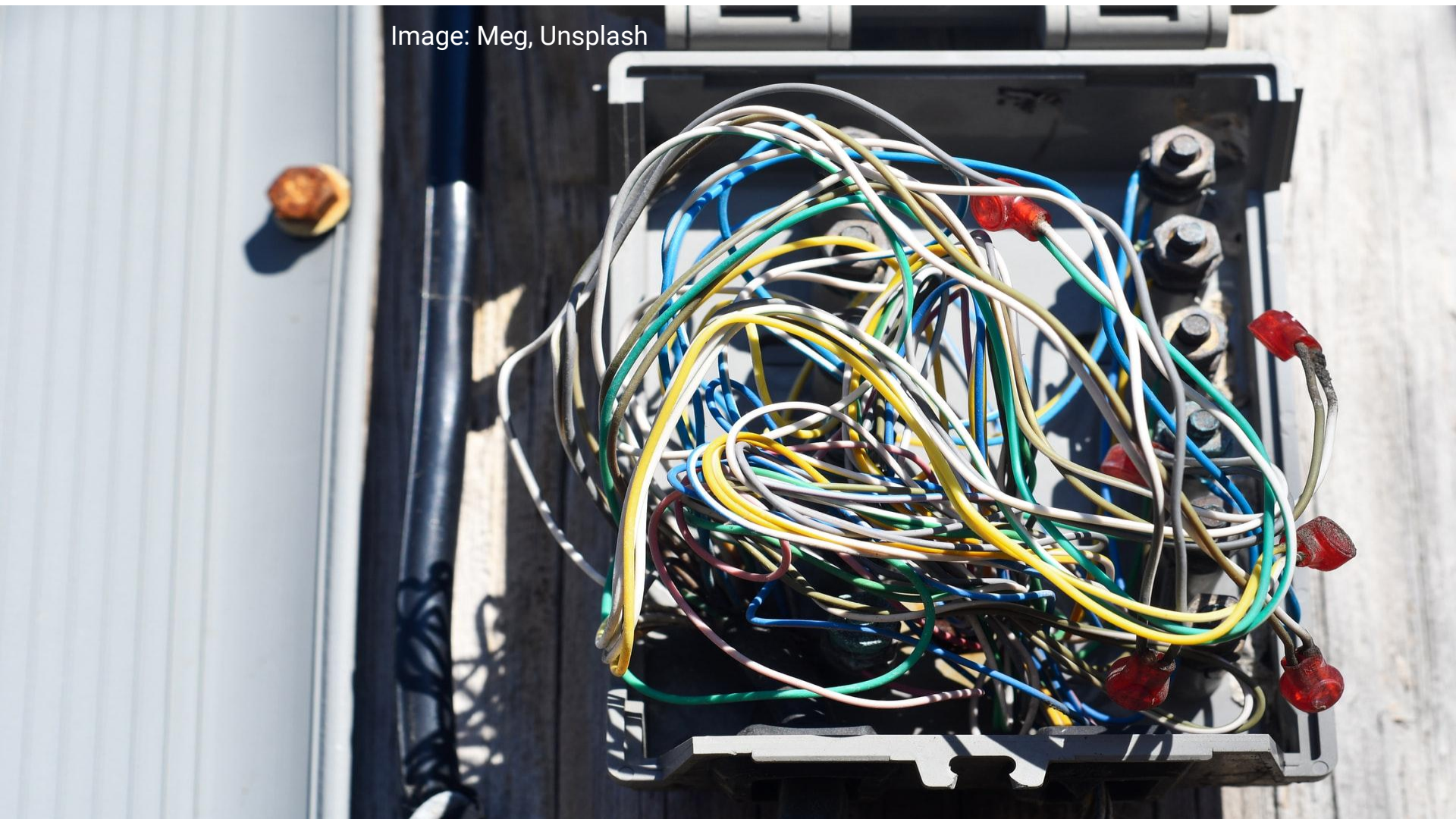
Image: Meg, Unsplash

TensorFlow Serving

TensorFlow Serving

Production ready Model Serving

- Part of the TensorFlow Extended (TFX) Ecosystem
- Used internally at Google
- Highly scalable model serving solution
- Works well for large models up to 2GB

Export your model

- Exported model as protobuf (saved_model.pb)
- Variables and checkpoints

```
import tensorflow as tf

tf.saved_model.save(
        model,
        export_dir="/tmp/saved_model",
        signatures=None
)
```

Export your model

- Consistent export format
- Uses Protobuf format
- Assets contains additional files like vocabularies

```
$ tree saved_models/
saved_models/
        162788939
                assets
                        saved_model.json
                saved_model.pb
                variables
                        checkpoint
                        variables.data-0000-of-0001
                        variables.index
```

coursera rhyme

TensorFlow Serving

- Docker images are available for CPU and GPU hardware
- REST and gRPC endpoints

```
$ docker pull tensorflow/serving

$ docker run -p 8500:8500 \
                -p 8501:8501 \
              --mount type=bind,\
                source=saved_models/,\
                target=/models/my_model \
                -e MODEL_NAME=my_model
                -t tensorflow/serving
```