Serving an Object Detector via TF Serving

Date: May 23, 2020

Author: Thanaphon Chavengsaksongkram

Email: contact@thanaphon.dev

Introduction

Serving a machine learning prediction can be done by simply running a model against a batch of collected data on a scheduled task. However, in any larger IT operations, it is often a requirement to serve the predictions on an on-demand basis to various parts of the IT infrastructure via a common protocol such as RESTful API.

Furthermore, machine learning models degrade over time (model rotting) and they require continuous training and deployment to production. This problem introduces many engineering challenges such as version controls, deployment, backing out, availability, scalability, etc.

Tensorflow Serving (TF Serving) is a solution designed to tackle many of the existing engineering tasks. It is an efficient model server written in C++ is capable of handling high load and deal with many other production-related tasks. TF Serving can also be deployed to private infrastructure and managed services such as Google Cloud AI Platform with extra benefits such as built-in monitoring.

Running TF Serving

There are many ways to install and run TF Serving: using a Docker image, using a system's package manager, or installing directly from the source. It is recommended by the TensorFlow team to use the TF Serving docker image, as it is one of the fastest ways to get your model to production. Docker images are generally platform-agnostic and can be deployed to various infrastructure. TF Serving Docker images also support easy configuration such as one with a GPU backend and one without.

There are two strategies to leverage TF Serving docker images.

1. Use the base TF Serving docker image as a generic model server and configure it to serve a specific model.
2. Create a new docker image with a model baked into it using TF Serving as a base image.

The first approach requires less maintenance overhead and it is suitable for most applications. The second approach can reduce deployment configuration, which can be useful for deploying a single model to many different platforms.

TF Serving's Model Format

TF Serving server expects a SavedModel, which represents a version of a model generated by tf.saved_model.save() function. It is stored as a directory containing a computation graph and its associated data. SavedModel also provides a CLI tool called saved_model_cli that can be used to inspect the model or make a test prediction.

Tensorflow 1.x saves a model into a frozen graph format. This is not compatible with TF serving as it expects a SavedModel format. Fortunately, a SavedModel is just a wrapper of a frozen graph with additional information such as signatures. Converting a frozen graph to a SavedModel is a pretty straightforward task.

Objective.

1. Deploy the model (locally) using Tensorflow Serving. A little tip: Tensorflow Serving might not be able to use the model in its current frozen graph format. Maybe you have to save it in a different format first!
2. Create a Tensorflow Serving docker image
3. Run the docker image and change `image_example.py` to use the external Tensorflow Serving model.
4. (optional) The code and application structure isn't very neat. Feel free to redesign the application structure and code to create a nicer, more usable client (package)

0. Prerequisites

0.1 Install Prerequisites

In [1]:
```
!pip install -r requirements.txt
```

```
Collecting opencv-python==4.1.0.25
  Downloading opencv_python-4.1.0.25-cp37-cp37m-macosx_10_7_x86_64.m
acosx_10_9_intel.macosx_10_9_x86_64.macosx_10_10_intel.macosx_10_10_
x86_64.whl (52.1 MB)
     |████████████████████████████████| 52.1 MB 2.3 MB/s eta 0:00:01
Collecting numpy==1.15.1
  Downloading numpy-1.15.1-cp37-cp37m-macosx_10_6_intel.macosx_10_9_
intel.macosx_10_9_x86_64.macosx_10_10_intel.macosx_10_10_x86_64.whl
(24.5 MB)
     |████████████████████████████████| 24.5 MB 2.8 MB/s eta 0:00:01
1
Collecting tensorflow==1.14.0
```

```
  Downloading tensorflow-1.14.0-cp37-cp37m-macosx_10_11_x86_64.whl (
105.8 MB)
     |████████████████████████████████| 105.8 MB 448 kB/s eta 0:00:0
11
Requirement already satisfied: gast>=0.2.0 in /Users/thanaphonchaven
gsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages
(from tensorflow==1.14.0->-r requirements.txt (line 3)) (0.2.2)
Collecting tensorboard<1.15.0,>=1.14.0
  Downloading tensorboard-1.14.0-py3-none-any.whl (3.1 MB)
     |████████████████████████████████| 3.1 MB 1.6 MB/s eta 0:00:01
Requirement already satisfied: keras-applications>=1.0.6 in /Users/t
hanaphonchavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/
site-packages (from tensorflow==1.14.0->-r requirements.txt (line 3)
) (1.0.8)
Collecting tensorflow-estimator<1.15.0rc0,>=1.14.0rc0
  Downloading tensorflow_estimator-1.14.0-py2.py3-none-any.whl (488
kB)
     |████████████████████████████████| 488 kB 3.9 MB/s eta 0:00:01
Requirement already satisfied: absl-py>=0.7.0 in /Users/thanaphoncha
vengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packag
es (from tensorflow==1.14.0->-r requirements.txt (line 3)) (0.9.0)
Requirement already satisfied: six>=1.10.0 in /Users/thanaphonchaven
gsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages
(from tensorflow==1.14.0->-r requirements.txt (line 3)) (1.15.0)
Requirement already satisfied: wrapt>=1.11.1 in /Users/thanaphonchav
engsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-package
s (from tensorflow==1.14.0->-r requirements.txt (line 3)) (1.12.1)
Requirement already satisfied: wheel>=0.26 in /Users/thanaphonchaven
gsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages
(from tensorflow==1.14.0->-r requirements.txt (line 3)) (0.34.2)
Requirement already satisfied: astor>=0.6.0 in /Users/thanaphonchave
ngsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages
(from tensorflow==1.14.0->-r requirements.txt (line 3)) (0.8.1)
Requirement already satisfied: termcolor>=1.1.0 in /Users/thanaphonc
havengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-pack
ages (from tensorflow==1.14.0->-r requirements.txt (line 3)) (1.1.0)
Requirement already satisfied: grpcio>=1.8.6 in /Users/thanaphonchav
engsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-package
s (from tensorflow==1.14.0->-r requirements.txt (line 3)) (1.29.0)
Requirement already satisfied: keras-preprocessing>=1.0.5 in /Users/
thanaphonchavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7
/site-packages (from tensorflow==1.14.0->-r requirements.txt (line 3
)) (1.1.2)
Requirement already satisfied: google-pasta>=0.1.6 in /Users/thanaph
onchavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-p
ackages (from tensorflow==1.14.0->-r requirements.txt (line 3)) (0.2
.0)
Requirement already satisfied: protobuf>=3.6.1 in /Users/thanaphonch
avengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packa
ges (from tensorflow==1.14.0->-r requirements.txt (line 3)) (3.12.1)
```

```
Requirement already satisfied: werkzeug>=0.11.15 in /Users/thanaphon
chavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-pac
kages (from tensorboard<1.15.0,>=1.14.0->tensorflow==1.14.0->-r requ
irements.txt (line 3)) (1.0.1)
Requirement already satisfied: markdown>=2.6.8 in /Users/thanaphonch
avengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packa
ges (from tensorboard<1.15.0,>=1.14.0->tensorflow==1.14.0->-r requir
ements.txt (line 3)) (3.2.2)
Requirement already satisfied: setuptools>=41.0.0 in /Users/thanapho
nchavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-pa
ckages (from tensorboard<1.15.0,>=1.14.0->tensorflow==1.14.0->-r req
uirements.txt (line 3)) (46.4.0.post20200518)
Requirement already satisfied: h5py in /Users/thanaphonchavengsakson
gkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages (from k
eras-applications>=1.0.6->tensorflow==1.14.0->-r requirements.txt (l
ine 3)) (2.10.0)
Requirement already satisfied: importlib-metadata; python_version <
"3.8" in /Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorfl
ow/lib/python3.7/site-packages (from markdown>=2.6.8->tensorboard<1.
15.0,>=1.14.0->tensorflow==1.14.0->-r requirements.txt (line 3)) (1.
6.0)
Requirement already satisfied: zipp>=0.5 in /Users/thanaphonchavengs
aksongkram/miniconda3/envs/tensorflow/lib/python3.7/site-packages (f
rom importlib-metadata; python_version < "3.8"->markdown>=2.6.8->ten
sorboard<1.15.0,>=1.14.0->tensorflow==1.14.0->-r requirements.txt (l
ine 3)) (3.1.0)
ERROR: tensorflow-serving-api 2.1.0 has requirement tensorflow~=2.1.
0, but you'll have tensorflow 1.14.0 which is incompatible.
Installing collected packages: numpy, opencv-python, tensorboard, te
nsorflow-estimator, tensorflow
  Attempting uninstall: numpy
    Found existing installation: numpy 1.18.4
    Uninstalling numpy-1.18.4:
      Successfully uninstalled numpy-1.18.4
  Attempting uninstall: tensorboard
    Found existing installation: tensorboard 2.1.1
    Uninstalling tensorboard-2.1.1:
      Successfully uninstalled tensorboard-2.1.1
  Attempting uninstall: tensorflow-estimator
    Found existing installation: tensorflow-estimator 2.1.0
    Uninstalling tensorflow-estimator-2.1.0:
      Successfully uninstalled tensorflow-estimator-2.1.0
  Attempting uninstall: tensorflow
    Found existing installation: tensorflow 2.1.0
    Uninstalling tensorflow-2.1.0:
      Successfully uninstalled tensorflow-2.1.0
Successfully installed numpy-1.18.4 opencv-python-4.1.0.25 tensorboa
rd-1.14.0 tensorflow-1.14.0 tensorflow-estimator-1.14.0
```

0.2 Install Missing Packages

In [3]:
```
!pip install Image
```

```
Processing /Users/thanaphonchavengsaksongkram/Library/Caches/pip/whe
els/09/21/3d/d9a06fda40387586027b9963b9558d6b655e0cde968737308f/imag
e-1.5.31-py2.py3-none-any.whl
Collecting django
  Using cached Django-3.0.6-py3-none-any.whl (7.5 MB)
Collecting pillow
  Using cached Pillow-7.1.2-cp37-cp37m-macosx_10_10_x86_64.whl (2.2
MB)
Requirement already satisfied: six in /Users/thanaphonchavengsaksong
kram/miniconda3/envs/tensorflow/lib/python3.7/site-packages (from Im
age) (1.15.0)
Collecting pytz
  Using cached pytz-2020.1-py2.py3-none-any.whl (510 kB)
Collecting asgiref~=3.2
  Using cached asgiref-3.2.7-py2.py3-none-any.whl (19 kB)
Collecting sqlparse>=0.2.2
  Using cached sqlparse-0.3.1-py2.py3-none-any.whl (40 kB)
ERROR: Error checking for conflicts.
Traceback (most recent call last):
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 3021, in _dep_map
    return self.__dep_map
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 2815, in __getattr__
    raise AttributeError(attr)
AttributeError: _DistInfoDistribution__dep_map

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 3012, in _parsed_pkg_info
    return self._pkg_info
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 2815, in __getattr__
    raise AttributeError(attr)
AttributeError: _pkg_info

During handling of the above exception, another exception occurred:
```

```
Traceback (most recent call last):
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_internal/commands/install.py", li
ne 517, in _warn_about_conflicts
    package_set, _dep_info = check_install_conflicts(to_install)
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_internal/operations/check.py", li
ne 114, in check_install_conflicts
    package_set, _ = create_package_set_from_installed()
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_internal/operations/check.py", li
ne 53, in create_package_set_from_installed
    package_set[name] = PackageDetails(dist.version, dist.requires()
)
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 2736, in requires
    dm = self._dep_map
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 3023, in _dep_map
    self.__dep_map = self._compute_dependencies()
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 3032, in _compute_dependencies
    for req in self._parsed_pkg_info.get_all('Requires-Dist') or []:
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 3014, in _parsed_pkg_info
    metadata = self.get_metadata(self.PKG_INFO)
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 1420, in get_metadata
    value = self._get(path)
  File "/Users/thanaphonchavengsaksongkram/miniconda3/envs/tensorflo
w/lib/python3.7/site-packages/pip/_vendor/pkg_resources/__init__.py"
, line 1616, in _get
    with open(path, 'rb') as stream:
FileNotFoundError: [Errno 2] No such file or directory: '/Users/than
aphonchavengsaksongkram/miniconda3/envs/tensorflow/lib/python3.7/sit
e-packages/numpy-1.18.4.dist-info/METADATA'
Installing collected packages: pytz, asgiref, sqlparse, django, pill
ow, Image
Successfully installed Image-1.5.31 asgiref-3.2.7 django-3.0.6 pillo
w-7.1.2 pytz-2020.1 sqlparse-0.3.1
```

0.3 Execute The Provided Script

Execute the provided script to test the functionality of the frozen graph.

In [95]:
```
!python3 image_example.py
```

```
WARNING:tensorflow:From image_example.py:92: FastGFile.__init__ (fro
m tensorflow.python.platform.gfile) is deprecated and will be remove
d in a future version.
Instructions for updating:
Use tf.gfile.GFile.
WARNING:tensorflow:From image_example.py:93: The name tf.GraphDef is
deprecated. Please use tf.compat.v1.GraphDef instead.

WARNING:tensorflow:From image_example.py:196: The name tf.Session is
deprecated. Please use tf.compat.v1.Session instead.

2020-05-22 23:51:52.790135: I tensorflow/core/platform/cpu_feature_g
uard.cc:142] Your CPU supports instructions that this TensorFlow bin
ary was not compiled to use: AVX2 FMA
```

0.4 Imports

All the required imports go here.

In [41]:
```
import tensorflow as tf
from tensorflow.python.platform import gfile
from tensorflow.python.saved_model import signature_constants
from tensorflow.python.saved_model import tag_constants
import os
```

1. Exploring the Frozen Graph

In order to convert the frozen graph to a SavedModel, more information about the graph is required.

1.1 Review the Frozen Graph on Tensorboard

Inspect the graph via Tensorboard. This can give critical information about the input and the prediction layers.

In [34]:
```
%load_ext tensorboard
with tf.Session() as sess:
    model_filename ='detector_frozen.pb'
    with gfile.FastGFile(model_filename, 'rb') as f:
        graph_def = tf.GraphDef()
        graph_def.ParseFromString(f.read())
        g_in = tf.import_graph_def(graph_def)
LOGDIR='tensorboard'
train_writer = tf.summary.FileWriter(LOGDIR)
train_writer.add_graph(sess.graph)
%tensorboard --logdir tensorboard
```

```
The tensorboard extension is already loaded. To reload it, use:
  %reload_ext tensorboard
WARNING:tensorflow:From <ipython-input-34-6ae794422132>:4: FastGFile
.__init__ (from tensorflow.python.platform.gfile) is deprecated and
will be removed in a future version.
Instructions for updating:
Use tf.gfile.GFile.
```

1.2 Input and Output Layers

According to the Tensorboard, the model contains one input layer and 3 output layers.

1.2.1 Input Layer:

- input/input_data

1.2.2 Output Layers

- pred_mbbox/concat_2
- pred_sbbox/concat_2
- pred_lbbox/concat_2

1.3 Converting Frozen Graph to SavedModel

In this step, a frozen graph model is exported into a SavedModel named sertis-detector with a version number of 1. A default predict-signature definition will be used. The input and output tensors are specified based on the information previously obtained.

In [36]:
```
export_dir = './serving/sertis-detector/1/'
graph_pb = 'detector_frozen.pb'

builder = tf.saved_model.builder.SavedModelBuilder(export_dir)

with tf.gfile.GFile(graph_pb, "rb") as f:
    graph_def = tf.GraphDef()
    graph_def.ParseFromString(f.read())

sigs = {}

with tf.Session(graph=tf.Graph()) as sess:
    # name="" is important to ensure we don't get spurious prefixing
    tf.import_graph_def(graph_def, name="")
    g = tf.get_default_graph()
   # print([n.name for n in tf.get_default_graph().as_graph_def().node
])
    inp = g.get_tensor_by_name("input/input_data:0")
    pred_mbbox = g.get_tensor_by_name("pred_mbbox/concat_2:0")
    pred_sbbox = g.get_tensor_by_name("pred_sbbox/concat_2:0")
    pred_lbbox = g.get_tensor_by_name("pred_lbbox/concat_2:0")
    sigs[signature_constants.DEFAULT_SERVING_SIGNATURE_DEF_KEY] = \
        tf.saved_model.signature_def_utils.predict_signature_def(
            {"in": inp}, {"out_mbbox": pred_mbbox, "out_sbbox": pred_s
bbox, "out_lbbox": pred_lbbox })

    builder.add_meta_graph_and_variables(sess,
                                         [tag_constants.SERVING],
                                         signature_def_map=sigs)

    builder.save()
```

```
INFO:tensorflow:No assets to save.
INFO:tensorflow:No assets to write.
INFO:tensorflow:SavedModel written to: ./serving/sertis-detector/1/s
aved_model.pb
```

1.4 Inspect the SavedModel

Using saved_model_cli, inspect the exported SavedModel for sanity check.

In [38]:
```
!saved_model_cli show --dir serving/sertis-detector/1 --tag_set serve
--signature_def serving_default
```

```
The given SavedModel SignatureDef contains the following input(s):
  inputs['in'] tensor_info:
      dtype: DT_FLOAT
      shape: unknown_rank
      name: input/input_data:0
The given SavedModel SignatureDef contains the following output(s):
  outputs['out_lbbox'] tensor_info:
      dtype: DT_FLOAT
      shape: (-1, -1, -1, 3, 85)
      name: pred_lbbox/concat_2:0
  outputs['out_mbbox'] tensor_info:
      dtype: DT_FLOAT
      shape: (-1, -1, -1, 3, 85)
      name: pred_mbbox/concat_2:0
  outputs['out_sbbox'] tensor_info:
      dtype: DT_FLOAT
      shape: (-1, -1, -1, 3, 85)
      name: pred_sbbox/concat_2:0
Method name is: tensorflow/serving/predict
```

2. Using TF Serving to Serve the Model

Before creating a docker image for this model, test if the model can be served by running it on the base TF serving image.

2.1 Download Tensorflow Serving Image

In [2]:
```
!docker pull tensorflow/serving
```

```
Using default tag: latest
latest: Pulling from tensorflow/serving

a4a261c9: Pulling fs layer
20cdee96: Pulling fs layer
60e1d0de: Pulling fs layer
7668deea: Pulling fs layer
b5699598: Pulling fs layer
8f5dbe31: Pulling fs layer
011e11a2: Pulling fs layer
Digest: sha256:ea44bf657f8cff7b07df12361749ea94628185352836bb0806534
5f5c8284bae
Status: Downloaded newer image for tensorflow/serving:latest
docker.io/tensorflow/serving:latest
```

2.1.1 Check the Downloaded Image

In [39]:
```
!docker images
```

```
REPOSITORY          TAG                 IMAGE ID            CREATED
SIZE
jwt-api-test        1.0                 f690ec151b72        5 weeks
ago         1.04GB
tensorflow/serving  latest              7c20ddd72597        4 month
s ago       251MB
python              stretch             b9d77e48a75c        8 month
s ago       940MB
```

2.2 Set Environment Variables

Unfortunately, Jupyter's notebook does not persist variables set by the shell. So Python's os package will be used instead to set the variables.

In [42]:
```
os.environ["MODEL_PATH"] = os.path.join(os.path.sep, os.getcwd(), "ser
ving", "sertis-detector")
```

In [43]:
```
!echo $MODEL_PATH
```

```
/Users/thanaphonchavengsaksongkram/Projects/ML-Practical/mle-take-ho
me-test/serving/sertis-detector
```

2.3 Start Tensorflow Serving Server

Start the model server and mount the sertis-detector model to the container file system.

In [64]:
```
!docker run -it --rm -p 8500:8500 -p 8501:8501 --name sertis-detector
--detach \
        -v "$MODEL_PATH:/models/sertis-detector" \
        -e MODEL_NAME=sertis-detector tensorflow/serving
```

```
docker: Error response from daemon: Conflict. The container name "/s
ertis-detector" is already in use by container "3f6cc2102b4fe8a0c4d9
6f4ed2415560a4f08f7cc18f113af283dfbac5839430". You have to remove (o
r rename) that container to be able to reuse that name.
See 'docker run --help'.
```

2.3.1 Explaination for Each Argument

For reference, a short explaination of every parameters is listed here.

--detach

run the image in the background

--name

name the container so we can stop or restart it later.

-v "$MODEL_PATH:/models/sertis-detector"

Mount the host file system that contains the model to the container file system at the specified path.

-e MODEL_NAME=my_mnist_model

Sets the container's MODEL_NAME environment variable, so TF Serving knows which model to serve. By default, it will look for models in the /models directory, and it will automatically serve the latest version it finds.

--rm

Deletes the container when you stop it (no need to clutter your machine with interrupted containers). However, it does not delete the image.

-p 8500:8500

Makes the Docker engine forward the host's TCP port 8500 to the container's TCP port 8500. By default, TF Serving uses this port to serve the gRPC API.

-p 8501:8501

Forwards the host's TCP port 8501 to the container's TCP port 8501. By default, TF Serving uses this port to serve the REST API.

2.3.2 Check if the Container is running.

In [45]:
```
!docker ps --all
```

```
CONTAINER ID        IMAGE                  COMMAND                  CR
EATED              STATUS                    PORTS                                    
NAMES
06dfb17894b5        tensorflow/serving    "/usr/bin/tf_serving…"    2
hours ago          Up 2 hours                0.0.0.0:8500-8501->8500-8
501/tcp   sertis-object-detector
cbf5a53282d3        jwt-api-test:1.0      "gunicorn -b :8080 m…"    5
weeks ago          Exited (0) 5 weeks ago                             
my-app
```

2.4 Test the Prediction Service

Due to the required preprocessing steps, using CURL may not be appropriate. Instead, the API will be tested by running a Python script predict_via_rest_api.py which preprocess the image and create an HTTP post request to the predict endpoint.

In [46]:
```
!python3 predict_via_rest_api.py
```

2.5 Clean-up

Stop and remove the running container.

In [49]:
```
!docker stop sertis-detector && docker rm sertis-detector
!docker ps --all
```

```
sertis-object-detector
Error: No such container: sertis-object-detector
CONTAINER ID        IMAGE                  COMMAND                  CR
EATED              STATUS                    PORTS                NAMES
50a327d2f825        tensorflow/serving    "/usr/bin/tf_serving…"    28
seconds ago        Up 27 seconds             8500-8501/tcp       servin
g_base
cbf5a53282d3        jwt-api-test:1.0      "gunicorn -b :8080 m…"    5
weeks ago          Exited (0) 5 weeks ago                         my-ap
p
```

3. Create a docker image to serve the model

The model has been tested on a TF Serving server. This task will simply copy the model and bake it into a new docker image for distribution.

3.1 Start Tensorflow Serving Server

In [47]:
```
!docker run -d --name serving_base tensorflow/serving
!docker ps -a
```

```
50a327d2f82589895703cd774122bf30df754908bb6412b43de0b1254acb02d8
CONTAINER ID        IMAGE                  COMMAND                  CR
EATED              STATUS                    PORTS                                              
NAMES
50a327d2f825        tensorflow/serving    "/usr/bin/tf_serving…"    1
second ago         Up Less than a second     8500-8501/tcp                                      
serving_base
06dfb17894b5        tensorflow/serving    "/usr/bin/tf_serving…"    2
hours ago          Up 2 hours                0.0.0.0:8500-8501->8500-8
501/tcp   sertis-object-detector
cbf5a53282d3        jwt-api-test:1.0      "gunicorn -b :8080 m…"    5
weeks ago          Exited (0) 5 weeks ago                                                       
my-app
```

3.2 Copy the model from the local filesystem into the container file system

In [51]:
```
!docker cp serving/sertis-detector serving_base:/models/sertis-detecto
r
```

3.3 Create a docker image with the new change applied

In [53]:
```
!docker commit --change "ENV MODEL_NAME sertis-detector" serving_base
tf-sertis-detector
```

```
sha256:b958cf060a5592c166fab0e00b5e4f9f6a201f79a92ad3e8aa91abfa18dbb
c60
```

3.4 Stop TF Serving container.

In [81]:
```
!docker kill serving_base
!docker stop serving_base && docker rm serving_base
```

```
Error response from daemon: Cannot kill container: serving_base: No
such container: serving_base
Error response from daemon: No such container: serving_base
```

In [82]:
```
!docker ps -a
```

```
CONTAINER ID        IMAGE                COMMAND                  CRE
ATED             STATUS                     PORTS
NAMES
3f6cc2102b4f        b958cf060a55         "/usr/bin/tf_serving…"   22
minutes ago      Up 22 minutes              0.0.0.0:8500-8501->8500-85
01/tcp   sertis-detector
cbf5a53282d3        jwt-api-test:1.0     "gunicorn -b :8080 m…"   5 w
eeks ago         Exited (0) 5 weeks ago
my-app
```

3.5 Check if the image is created

In [77]:
```
!docker images
```

```
REPOSITORY                                TAG                 IMAGE ID
CREATED             SIZE
as12production/sertis-object-detector     1.0                 b958cf06
0a55        24 minutes ago      499MB
tf-sertis-detector                        latest              b958cf06
0a55        24 minutes ago      499MB
jwt-api-test                              1.0                 f690ec15
1b72        5 weeks ago         1.04GB
tensorflow/serving                        latest              7c20ddd7
2597        4 months ago        251MB
python                                    stretch             b9d77e48
a75c        8 months ago        940MB
centurylink/dockerfile-from-image         latest              970eaf37
5dfd        4 years ago         19.2MB
```

3.6 Test the Prediction Service

Start a docker container using the new image. Then run the predict_via_rest_api.py script to test its functionality.

In [62]:
```
!docker run -it --rm -p 8500:8500 -p 8501:8501 --name sertis-detector
--detach \
        -e MODEL_NAME=sertis-detector tf-sertis-detector
```

```
3f6cc2102b4fe8a0c4d96f4ed2415560a4f08f7cc18f113af283dfbac5839430
```

In [65]:
```
!python3 predict_via_rest_api.py
```

3.7 Clean up

In [83]:
```
!docker stop sertis-detector && docker rm sertis-detector
```

```
sertis-detector
Error: No such container: sertis-detector
```

4. Deploy to Docker hub

Deploy the newly created image to a docker registry (docker hub).

4.1 Tag the image

In [72]:
```
!docker tag b958cf060a55 as12production/sertis-object-detector:1.0
```

4.2 Docker Hub - Login

In [75]:
```
!docker login --username as12production
```

```
Password:
```

4.3 Push the image to Docker Hub

In [76]:
```
!docker push as12production/sertis-object-detector:1.0
```

```
The push refers to repository [docker.io/as12production/sertis-objec
t-detector]

4b169550: Preparing
d98b810c: Preparing
55bd8fcf: Preparing
61ac0e5e: Preparing
3374c0b5: Preparing
fb8f161b: Preparing
43ea46a8: Preparing
fcc4a1a8: Preparing
4b169550: Pushed   248.3MB/248.3MBserving 1.0: digest: sha256:4b6e58
f60a825e34a39b968610e9af3deec30f0acdd9c4493f9820a892784ec0 size: 220
2
```

https://hub.docker.com/r/as12production/sertis-object-detector
(https://hub.docker.com/r/as12production/sertis-object-detector)

4.4 Test the Prediction Service

Using the image acquired from Docker Hub, the prediction service will be tested using the same script as previously.

4.4.1 Remove any old images

In [84]:
```
!docker rmi tf-sertis-detector
!docker rmi as12production/sertis-object-detector
!docker rmi tensorflow/serving
!docker images
```

```
Error: No such image: tf-sertis-detector
Untagged: as12production/sertis-object-detector:1.0
Untagged: as12production/sertis-object-detector@sha256:4b6e58f60a825
e34a39b968610e9af3deec30f0acdd9c4493f9820a892784ec0
Deleted: sha256:b958cf060a5592c166fab0e00b5e4f9f6a201f79a92ad3e8aa91
abfa18dbbc60
Deleted: sha256:3559d4427da077b38cddcf889c1a4c9b385f7f6d58a7daf2d68d
b863955e7ee9
Deleted: sha256:7c20ddd72597be37ca64e0393fdc219b8906b8709becacf51f74
6c9f812a8121
Deleted: sha256:a4cc2c00fdca74c89dec852801b1824cb5fd22e90ac97be2e843
57ea3145f95b
Deleted: sha256:6decd594d39b31482b1d147650d855358a26fc600dc06a67ca81
228bc7feef6c
Deleted: sha256:6e949cb9cd885c847557035e725919b879b201f37df07e2b19fa
e80a088058a3
Deleted: sha256:2d95a023d1fa3fc0caabcc97ee5dcdb7e75dd79e24567431f8e3
4047ae660ee7
Deleted: sha256:7c52cdc1e32d67e3d5d9f83c95ebe18a58857e68bb6985b0381e
bdcec73ff303
Deleted: sha256:a3c2e83788e20188bb7d720f36ebeef2f111c7b939f1b19aa1b4
756791beece0
Deleted: sha256:61199b56f34827cbab596c63fd6e0ac0c448faa7e026e3309948
18190852d479
Deleted: sha256:2dc9f76fb25b31e0ae9d36adce713364c682ba0d2fa70756486e
5cedfaf40012
Error: No such image: tensorflow/serving
REPOSITORY          TAG                 IMAGE ID            CREATED
SIZE
jwt-api-test        1.0                 f690ec151b72        5 weeks
ago         1.04GB
python              stretch             b9d77e48a75c        8 months
ago         940MB
```

4.4.2 Pull the image from Docker Hub

In [86]:
```
!docker pull as12production/sertis-object-detector:1.0
```

```
1.0: Pulling from as12production/sertis-object-detector

a4a261c9: Pulling fs layer
20cdee96: Pulling fs layer
60e1d0de: Pulling fs layer
7668deea: Pulling fs layer
b5699598: Pulling fs layer
8f5dbe31: Pulling fs layer
011e11a2: Pulling fs layer
075f0126: Pulling fs layer
Digest: sha256:4b6e58f60a825e34a39b968610e9af3deec30f0acdd9c4493f982
0a892784ec0[9A
Status: Downloaded newer image for as12production/sertis-object-dete
ctor:1.0
docker.io/as12production/sertis-object-detector:1.0
```

4.4.3 List all images

In [87]:
```
!docker images
```

```
REPOSITORY                                TAG                 IMAGE ID
CREATED             SIZE
as12production/sertis-object-detector   1.0                 b958cf06
0a55        29 minutes ago      499MB
jwt-api-test                              1.0                 f690ec15
1b72        5 weeks ago         1.04GB
python                                    stretch             b9d77e48
a75c        8 months ago        940MB
```

4.4.4 Start a Container with the new image

In [89]:
```
!docker run -it --rm -p 8500:8500 -p 8501:8501 --name sertis-detector
--detach \
        -e MODEL_NAME=sertis-detector as12production/sertis-object-det
ector:1.0
```

```
ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522
```

4.4.5 Inspect the container

In [91]:
```
!docker inspect ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522
```


```
[
    {
        "Id": "ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522",
        "Created": "2020-05-22T16:35:13.9663642Z",
        "Path": "/usr/bin/tf_serving_entrypoint.sh",
        "Args": [],
        "State": {
            "Status": "running",
            "Running": true,
            "Paused": false,
            "Restarting": false,
            "OOMKilled": false,
            "Dead": false,
            "Pid": 6785,
            "ExitCode": 0,
            "Error": "",
            "StartedAt": "2020-05-22T16:35:14.2965373Z",
            "FinishedAt": "0001-01-01T00:00:00Z"
        },
        "Image": "sha256:b958cf060a5592c166fab0e00b5e4f9f6a201f79a92ad3e8aa91abfa18dbbc60",
        "ResolvConfPath": "/var/lib/docker/containers/ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522/resolv.conf",
        "HostnamePath": "/var/lib/docker/containers/ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522/hostname",
        "HostsPath": "/var/lib/docker/containers/ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522/hosts",
        "LogPath": "/var/lib/docker/containers/ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522/ea7018f0e1ff409a97437020fe0e5c903b6cb254ae77ebcd81618d28939dc522-json.log",
        "Name": "/sertis-detector",
        "RestartCount": 0,
        "Driver": "overlay2",
        "Platform": "linux",
        "MountLabel": "",
        "ProcessLabel": "",
        "AppArmorProfile": "",
        "ExecIDs": null,
        "HostConfig": {
            "Binds": null,
            "ContainerIDFile": "",
            "LogConfig": {
                "Type": "json-file",
                "Config": {}
            },
```

```
"NetworkMode": "default",
"PortBindings": {
    "8500/tcp": [
        {
            "HostIp": "",
            "HostPort": "8500"
        }
    ],
    "8501/tcp": [
        {
            "HostIp": "",
            "HostPort": "8501"
        }
    ]
},
"RestartPolicy": {
    "Name": "no",
    "MaximumRetryCount": 0
},
"AutoRemove": true,
"VolumeDriver": "",
"VolumesFrom": null,
"CapAdd": null,
"CapDrop": null,
"Capabilities": null,
"Dns": [],
"DnsOptions": [],
"DnsSearch": [],
"ExtraHosts": null,
"GroupAdd": null,
"IpcMode": "private",
"Cgroup": "",
"Links": null,
"OomScoreAdj": 0,
"PidMode": "",
"Privileged": false,
"PublishAllPorts": false,
"ReadonlyRootfs": false,
"SecurityOpt": null,
"UTSMode": "",
"UsernsMode": "",
"ShmSize": 67108864,
"Runtime": "runc",
"ConsoleSize": [
    0,
    0
],
"Isolation": "",
"CpuShares": 0,
"Memory": 0,
```

```
            "NanoCpus": 0,
            "CgroupParent": "",
            "BlkioWeight": 0,
            "BlkioWeightDevice": [],
            "BlkioDeviceReadBps": null,
            "BlkioDeviceWriteBps": null,
            "BlkioDeviceReadIOps": null,
            "BlkioDeviceWriteIOps": null,
            "CpuPeriod": 0,
            "CpuQuota": 0,
            "CpuRealtimePeriod": 0,
            "CpuRealtimeRuntime": 0,
            "CpusetCpus": "",
            "CpusetMems": "",
            "Devices": [],
            "DeviceCgroupRules": null,
            "DeviceRequests": null,
            "KernelMemory": 0,
            "KernelMemoryTCP": 0,
            "MemoryReservation": 0,
            "MemorySwap": 0,
            "MemorySwappiness": null,
            "OomKillDisable": false,
            "PidsLimit": null,
            "Ulimits": null,
            "CpuCount": 0,
            "CpuPercent": 0,
            "IOMaximumIOps": 0,
            "IOMaximumBandwidth": 0,
            "MaskedPaths": [
                "/proc/asound",
                "/proc/acpi",
                "/proc/kcore",
                "/proc/keys",
                "/proc/latency_stats",
                "/proc/timer_list",
                "/proc/timer_stats",
                "/proc/sched_debug",
                "/proc/scsi",
                "/sys/firmware"
            ],
            "ReadonlyPaths": [
                "/proc/bus",
                "/proc/fs",
                "/proc/irq",
                "/proc/sys",
                "/proc/sysrq-trigger"
            ]
        },
        "GraphDriver": {
```

```
            "Data": {
                "LowerDir": "/var/lib/docker/overlay2/919ade47a8758c
b34572da967e4c80ade7044a224d17607a9d913f2b405202e1-init/diff:/var/li
b/docker/overlay2/ea660bf6f1d74d7007ca87c7c76aa5f70255288c6a9b8d8755
a74831e4c31624/diff:/var/lib/docker/overlay2/c02465433b84490bc0a3e1a
826d85ae4dfd1821f0430a32317ad1ac4537c62e9/diff:/var/lib/docker/overl
ay2/a7cd4efddb304376745f1b567bc03e685c24c3bc4903e77bcae7ee7d8d30b37c
/diff:/var/lib/docker/overlay2/dd39f3977a94c0c7b522bb4a2ca4156146a32
a682187aa37609226abb873f90a/diff:/var/lib/docker/overlay2/73e3f3dd9b
9adab3269336e5d3a8ac7dc149166a55964078caffc21e044ccc2c/diff:/var/lib
/docker/overlay2/ba5ae2f61a135b74d5b0ff70aea02588b050134b325b51af7cf
5f3395029073a/diff:/var/lib/docker/overlay2/6d918807d973566ccc6bb17b
a0f88f24f6126f004c53e4f2cc69bde1e84852b4/diff:/var/lib/docker/overla
y2/029dc0958bf8527dfecf2e5545129a308d7a27746f9a0de59ccb9c4697940f47/
diff:/var/lib/docker/overlay2/11f9cb0331912e55b56b511c6af1b37aae47d3
6c6e186d46a6ff274af97fb3c4/diff",
                "MergedDir": "/var/lib/docker/overlay2/919ade47a8758
cb34572da967e4c80ade7044a224d17607a9d913f2b405202e1/merged",
                "UpperDir": "/var/lib/docker/overlay2/919ade47a8758c
b34572da967e4c80ade7044a224d17607a9d913f2b405202e1/diff",
                "WorkDir": "/var/lib/docker/overlay2/919ade47a8758cb
34572da967e4c80ade7044a224d17607a9d913f2b405202e1/work"
            },
            "Name": "overlay2"
        },
        "Mounts": [],
        "Config": {
            "Hostname": "ea7018f0e1ff",
            "Domainname": "",
            "User": "",
            "AttachStdin": false,
            "AttachStdout": false,
            "AttachStderr": false,
            "ExposedPorts": {
                "8500/tcp": {},
                "8501/tcp": {}
            },
            "Tty": true,
            "OpenStdin": true,
            "StdinOnce": false,
            "Env": [
                "MODEL_NAME=sertis-detector",
                "PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/
bin:/sbin:/bin",
                "MODEL_BASE_PATH=/models"
            ],
            "Cmd": null,
            "Image": "as12production/sertis-object-detector:1.0",
            "Volumes": null,
            "WorkingDir": "",
```

```
            "Entrypoint": [
                "/usr/bin/tf_serving_entrypoint.sh"
            ],
            "OnBuild": null,
            "Labels": {
                "maintainer": "gvasudevan@google.com",
                "tensorflow_serving_github_branchtag": "2.1.0",
                "tensorflow_serving_github_commit": "d83512c6b5b2b84
33df2fd61bbbfb22e0295b3d3"
            }
        },
        "NetworkSettings": {
            "Bridge": "",
            "SandboxID": "3a8df7040862bda053bb114cd24d8b30f382896aa6
f773985763b760298fe424",
            "HairpinMode": false,
            "LinkLocalIPv6Address": "",
            "LinkLocalIPv6PrefixLen": 0,
            "Ports": {
                "8500/tcp": [
                    {
                        "HostIp": "0.0.0.0",
                        "HostPort": "8500"
                    }
                ],
                "8501/tcp": [
                    {
                        "HostIp": "0.0.0.0",
                        "HostPort": "8501"
                    }
                ]
            },
            "SandboxKey": "/var/run/docker/netns/3a8df7040862",
            "SecondaryIPAddresses": null,
            "SecondaryIPv6Addresses": null,
            "EndpointID": "0e1f4061c0a630b5c6b308ff231afb124afa97945
5ab135154542c420d91a89b",
            "Gateway": "172.17.0.1",
            "GlobalIPv6Address": "",
            "GlobalIPv6PrefixLen": 0,
            "IPAddress": "172.17.0.2",
            "IPPrefixLen": 16,
            "IPv6Gateway": "",
            "MacAddress": "02:42:ac:11:00:02",
            "Networks": {
                "bridge": {
                    "IPAMConfig": null,
                    "Links": null,
                    "Aliases": null,
                    "NetworkID": "7ebccccd906873b483c625a2f33ac8fd38
```

```
c06b03b6d6e69d92eaf8edfbe9401d",
                    "EndpointID": "0e1f4061c0a630b5c6b308ff231afb124
afa979455ab135154542c420d91a89b",
                    "Gateway": "172.17.0.1",
                    "IPAddress": "172.17.0.2",
                    "IPPrefixLen": 16,
                    "IPv6Gateway": "",
                    "GlobalIPv6Address": "",
                    "GlobalIPv6PrefixLen": 0,
                    "MacAddress": "02:42:ac:11:00:02",
                    "DriverOpts": null
                }
            }
        }
    }
]
```

4.6 Test the Prediction Service

In [92]:
```
!python3 predict_via_rest_api.py
```

4.7 Clean-up

In [94]:
```
!docker stop sertis-detector  && docker rm sertis-detector
!docker ps -a
```

```
sertis-detector
Error: No such container: sertis-detector
CONTAINER ID        IMAGE               COMMAND                  CRE
ATED             STATUS                  PORTS               NAMES
cbf5a53282d3        jwt-api-test:1.0    "gunicorn -b :8080 m…"   5 w
eeks ago         Exited (0) 5 weeks ago                      my-app
```

5. Conclusion

In this task, a frozen graph has been converted to a SavedModel and deployed to a Tensorflow Serving server. A new docker image is also created with the model included and is deployed to Docker Hub.

5.1 Future Tasks

In my opinion, the model is not user friendly, because it requires an image preprocessing steps before the input can be fed into the prediction service. I believe that the transformation pipeline should be included as part of the prediction service or perhaps create a wrapper API that performs this data preprocessing. It is a lot more user friendly if the API simply accepts a list of base64 encoded images and return the predictions.

In [ ]: