instructions.md

Deploy Models with TensorFlow Serving and Flask

If you want to run the code from the hands on project Deploy Models with TensorFlow Serving and Flask on your local machine, please follow the instructions given in this file.

What's Included

Following folders and files are included in this zip file:

1. `pets` - TensorFlow SavedModel Directory
2. `static` - Empty directory which will be used for storing images by the flask app
3. `templates` - HTML templates are here
4. `app.py` - Flask app
5. `instructions.pdf` - This file

Environment

You will require Python3 installed. I used python 3.7 and TensorFlow 2.1.0, and I'd recommend you do the same. It is recommended that you create a new virtual environment to avoid issues with existing installations.

Install the python packages required:

```
$ pip3 install tensorflow==2.1.0 flask flask-bootstrap requests
```

Docker Instance

Launch the docker instance which will serve the TensorFlow SavedModel (in the **pets** folder):

```
$ sudo docker run -p PORT_NUMBER:8501 --name=pets -v "YOUR_SAVED_MODEL_PATH:/models/pets/1" -e MODEL_NAME=pets tensorflow/serving
```

In the project, we used 8502 for the `PORT_NUMBER`, and `YOUR_SAVED_MODEL_PATH` needs to be the *absolute* path of the **pets** folder in your local machine. So, if you extracted the downloaded zip file in, say, `/home/example/`, and want to use 8502 for the server port, the above command will become:

```
$ sudo docker run -p 8502:8501 --name=pets -v "/home/example/pets/:/models/pets/1" -e MODEL_NAME=pets tensorflow/serving
```

Please note if you use any other port, you will have to change the `MODEL_URI` in the `app.py` file accordingly.

Flask App

Once the docker instance is running, you can launch the flask app:

```
$ python3 app.py
```

And that's it! The default port for flask is 5000, so you can access the app by going to `localhost:5000` in your browser.