# Machine Learning Model Deployment

Contents

[Machine Learning Model Deployment 1](#_Toc112566853)

[Overview: 1](#_Toc112566854)

[Architecture: 1](#_Toc112566855)

[Fast API: 2](#_Toc112566856)

[Ngrok: 2](#_Toc112566857)

[Model: 2](#_Toc112566858)

[Web pages: 2](#_Toc112566859)

[Steps to deploy model: 2](#_Toc112566860)

[References 3](#_Toc112566861)

# Overview:

Machine learning model can be deployed in webserver for production purpose or wide client access. Following document describe about architecture and steps in deployment. Model can deployed Uvicorn server or google colab for experimental purpose.

# Architecture:

Server

Browser

Ngrok

<<gateway>>

FastAPI

<<webapi>>

Machine Learning Model

Web pages

<<static>>

### Fast API:

FastAPI is a modern, fast (high-performance), web framework for building APIs with Python 3.6+ based on standard Python.

### Ngrok:

Ngrok is gateway and load balancer, it will open secure web connections with browser and revere proxy to FASTAPI web server. it also provide public url to access website.

### Model:

ML model is trained and tested model and it can be saved in different formats. save\_model

Provide option to saving model.

### Web pages:

These are static html file, scripts and css.

# Steps to deploy model:

* Install ngork

!pip install flask**-**ngrok

!pip install flask**==**0.12.2

* Install DL/ML libarries

!pip install tensorflow

!pip install keras

!pip install pillow

!pip install numpy

!pip install tqdm

!pip install keras**.**utils

* Load google drive

from google.colab import drive

drive**.**mount('/content/drive')

* Import important API

**from** keras.models **import** Model, load\_model

**from** keras.applications.xception **import** Xception, preprocess\_input

* Update nrok token

!**/**ngrok authtoken 2DI4bkOAQIhGOsc75CHmWQFEVuq\_5pFDKAMNaBtXXXXX

* Develop API

# return webpage

@app**.**route("/")

**def** home():

**return** render\_template('index.html')

# load model and prediction

@app**.**route('/getprediction', methods**=**['POST'])

**def** getprediction():

tokenizer **=** load(open('/content/drive/MyDrive/website/model/tokenizer.p',"rb"))

model **=** load\_model('/content/drive/MyDrive/website/model/model\_9.h5')

xception\_model **=** Xception(include\_top**=False**, pooling**=**"avg")

photo **=** extract\_features(img\_path, xception\_model)

img **=** Image**.**open(img\_path)

description **=** generate\_desc(model, tokenizer, photo, max\_length)

description **=** description**.**replace("start ", "")

description **=** description**.**replace(" end", "")

* Public URL

\* Running on <http://95e0-34-148-31-45.ngrok.io>

\* Traffic stats available on http://127.0.0.1:4040

# References

<https://keras.io/api/models/model_saving_apis/https://keras.io/api/models/model_saving_apis/>

<https://fastapi.tiangolo.com/>

<https://ngrok.com/>