ML model deployment using Docker container

Preeti AI · Follow
2 min read · Aug 24, 2023

Deploying a machine learning model on Docker is a common practice to ensure that your model runs consistently across different environments. Docker containers package your application and its dependencies into a single unit that can be easily deployed and run on various platforms. Here's a step-by-step guide on how to deploy a machine learning model on Docker:

1. Set Up Your Environment:

Before you begin, make sure you have the following installed on your machine:

- *Docker: You can download and install Docker from the official website (https://www.docker.com/get-started).*

2. Create a Directory for Your Project:

Create a directory for your project and navigate to it using your terminal or command prompt.

mkdir my_ml_model
cd my_ml_model

3. Prepare Your Model Code:

Put your machine learning model code and any necessary files (trained model weights, data preprocessing scripts, etc.) into this directory. Ensure that all dependencies are properly listed in a requirements.txt or similar file.

4. Create a Dockerfile:

Create a file called `Dockerfile` *(without any file extension) in your project directory. This file contains instructions on how to build your Docker image. Here's a simple example of a Dockerfile for a Python-based ML model:*

```
# Use an official Python runtime as a parent image
FROM python:3.8-slim

# Set the working directory to /app
WORKDIR /app

# Copy the current directory contents into the container at /app
COPY . /app

# Install any needed packages specified in requirements.txt
RUN pip install — trusted-host pypi.python.org -r requirements.txt

# Make port 80 available to the world outside this container
EXPOSE 80

# Define environment variable
ENV NAME World
```

 # Run app.py when the container launches
CMD [“python”, “your_model_script.py”]

Replace `your_model_script.py` with the actual name of your model deployment script.

5. Build the Docker Image:

Open your terminal or command prompt, navigate to the project directory, and run the following command to build your Docker image:

docker build -t my-ml-model .

This command tells Docker to build an image named `my-ml-model` using the current directory (denoted by `.`) as the build context.

6. Run a Docker Container:

Once the image is built successfully, you can run a Docker container based on this image:

docker run -p 4000:80 my-ml-model

This command maps port 4000 on your host machine to port 80 in the Docker container. Adjust the port numbers as needed.

7. Access Your Model API:

Your model should now be accessible via HTTP at `http://localhost:4000` (or the port you specified in the `docker run` command). You can send requests to this URL to make predictions or perform inference using your model.

Docker ML AI Deployment Python

Preeti AI

Instruction-tuning Llama-2–7B for News Classification

The purpose of this notebook is to provide a comprehensive, step-by-step tutorial for fine-tuning any LLM (Large Language Model).

15 min read · Oct 22, 2023

56

Preeti AI

Medical Chatbot using GPT-3

We'll build a simple medical chatbot using the GPT-3 language model from OpenAI. Please note that you'll need access to the OpenAI GPT-3...

2 min read · Jul 28, 2023

51 1

Preeti AI

ResNet-Computer Vision

Use a pre-trained image classification model (e.g., ResNet50) for a specific problem (e.g., classifying cats and dogs) with Python code...

2 min read · Jul 31, 2023

2

Preeti AI

LLM Project for beginners

This is a basic example on how to use OpenAI's GPT-3 (which is one of their language models) in a Python code snippet. Please note that...

1 min read · Jul 28, 2023

See all from Preeti AI

Recommended from Medium

Moez Ali

Easily deploy machine learning models from the comfort of your Notebook

Streamline your Machine Learning model deployment with Modelbit.com. Seamlessly move from Jupyter notebooks to production-ready REST APIs...

9 min read · Sep 8, 2023


295

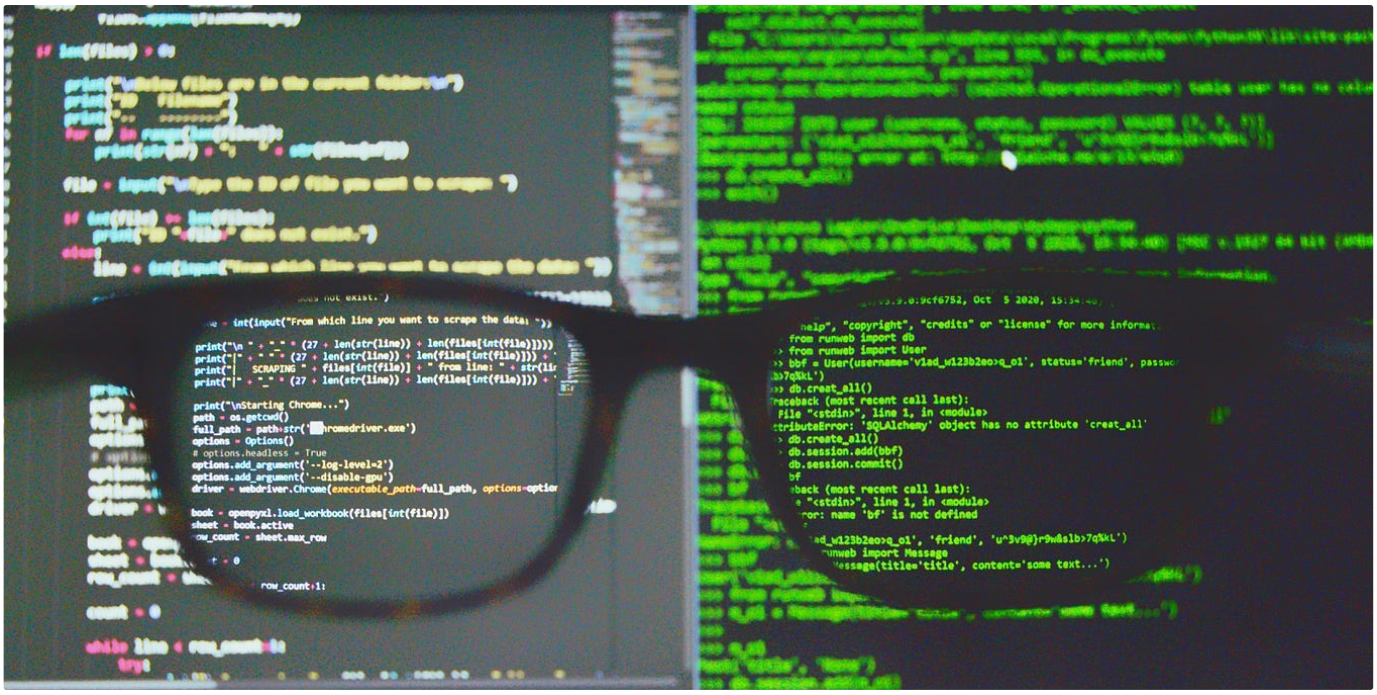
1


Daniel García in LatinXinAI

Step on Step guide to deploy YOLO model using FastAPI

Introduction

8 min read · Sep 9, 2023



Charu Makhijani

Machine Learning Model Deployment as a Web App using Streamlit

Complete Guide to deploy ML model using Streamlit

5 min read · Nov 7, 2023

302 2

Madhav Thaker

(Part 1) Build your own RAG with Mistral-7B and LangChain

LLMs have taken the world by storm and rightfully so. They help you with every day tasks like building a coding project, creating a recipe...

14 min read · Nov 14, 2023

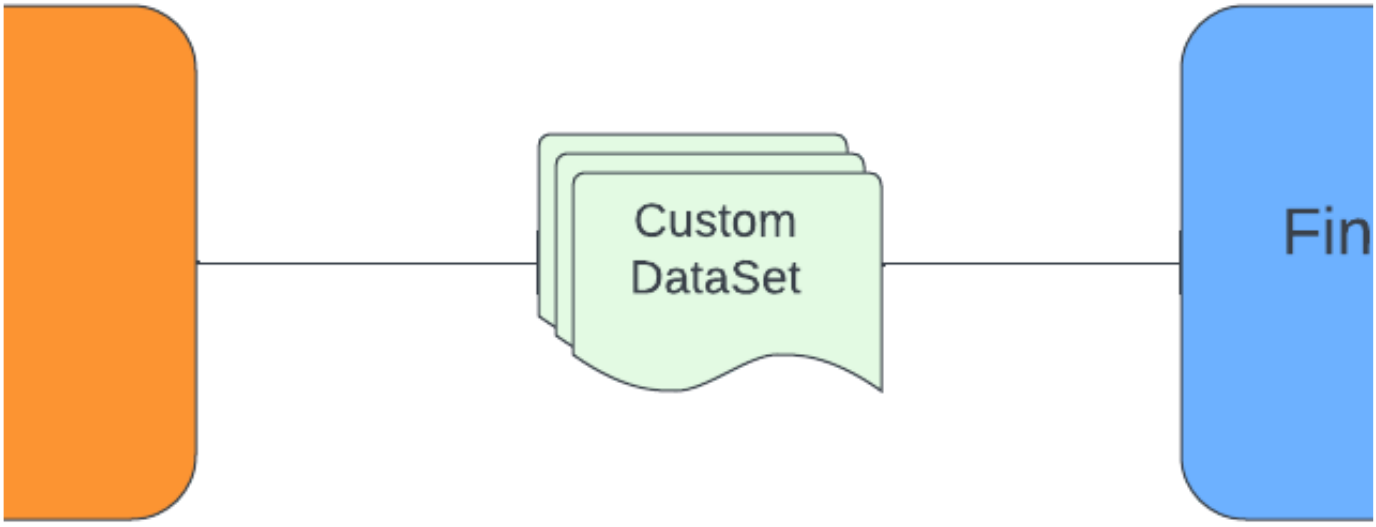
Prasenjit Chowdhury

Deployment of Machine Learning Models: On-Premises and in the Cloud

In the rapidly evolving field of machine learning, creating a powerful model is just the beginning. To make the most out of your machine...

3 min read · Sep 24, 2023

2



Suman Das

Fine Tune Large Language Model (LLM) on a Custom Dataset with QLoRA

The field of natural language processing has been revolutionized by large language models (LLMs), which showcase advanced capabilities and...

15 min read · Jan 25, 2024

790 7

See more recommendations