Video Object Detection with LMMs for Visual Q/A

Start Date: 10/1/2024

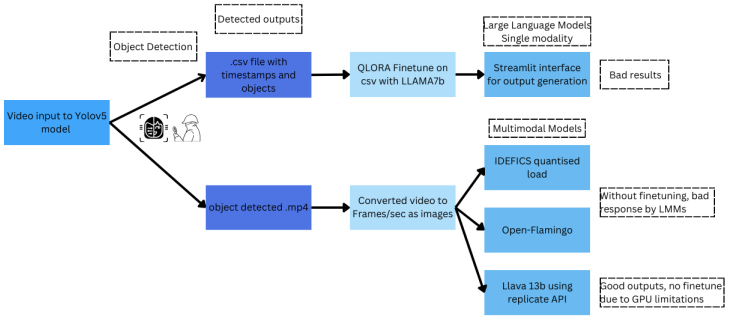
Submission Date: 9/2/2024

by
Agam Pandey
Civil Engineering
22113009
agam_p@ce.iitr.ac.in

Refer to the GitHub Repo for documentation and source code
https://github.com/AGAMPANDEYY/Yolov_LLM.git

Indian Institute of Technology Roorkee

1 Introduction



Computer Vision and Natural Language Processing using LMMs

The project worked by me is based on Object detection using YOLO technique and Visual Language Querying using the Llava 13b model.The link to proposal PDF attached here:
https://drive.google.com/file/d/1GRPIvfd5LSjEIfrp22TOsCJUH0YTHQqT/view?usp=sharing
In the fast-changing world, the problem of analyzing complex real-time data is getting tough. With the advancement of Generative AI and Computer vision, the problem of monitoring real-time footage and querying needed questions from a model can be beneficial.

Inventory management can be simplified
Suppose, a manager would want to know how many packages of company A were kept at a place.
At what time was it picked up?
How many workers were handling it?
What type of vehicle stored it?
Traffic Management. In an area, CCTVs can give us videos from which our model detects objects and converts and stores them in the form of metadata, LLM model can be used to query information regarding the number of cars passing through the area in a time interval.
Total number of people/cars?
Traffic violations can also be listed (if we work on speed detection) An advanced CNN supported by the LLM model that are multimodal models can help identify objects from CCTVs and generate responses to questions by managers!

Proposal Implementation:
Creating a model using Computer Vision to detect objects in a Video and Large Language The project implementation was divided into 3 phases:

- Video Object Detection
- Integrated LMM for VQA
- Data Collection for Fine-tuning

Figure 1: Major packages and Libraries used

2 Model and Datasets used

The process of creating an end-to-end project required me to explore new concepts and understand the working of codes, during this phase I utilised different libraries and models. The tech stack is mentioned:

- Yolov5 for Object Detection Car, Pedestrian,Traffic Signs
- Large Language Model- Llama
- Multimodal Visual Language Models - Open Flamingo,Llava
- Streamlit for web UI and Flask for REST APIs

The concepts learnt during the steps were :

- CNNs working
- YOLO and related segmentation concepts
- NLP-related concepts
- LSTMs,Transfomers, and Multimodal models
- Visual Encoders/Language encoders and Architecture of LMMs
- Hosting Streamlit web application

Dataset

The dataset which i initially worked on was video frames of BDD100K images which I decided to train my Yolo model on. But due to limited GPU RAM and disk space, the training was done on a subset image, label of BDD100K which I used through Roboflow
https://universe.roboflow.com/pedro-azevedo-3c9ol/bdd100k-3zgda/browse.
The dataset that I first thought of using with Llava and Llama was a traffic related Q/A dataset which i couldn't find for fine tuning, so I worked on few shot inferences to get desired output.

3 Model Pipeline Discussion

The model end to end pipeline is discussed below.

1. Cloning GitHub repo of Yolov5 model, trying Panoptic segmentation on custom data
2. Training Yolov5 on a custom dataset called with Roboflow API.
3. Downloading the trained model checkpoints
4. Object detection on custom traffic dashcam video
5. Saved .mp4 file broken down to frames/sec into images with detected objects

Figure 2: Object detection Yolov5

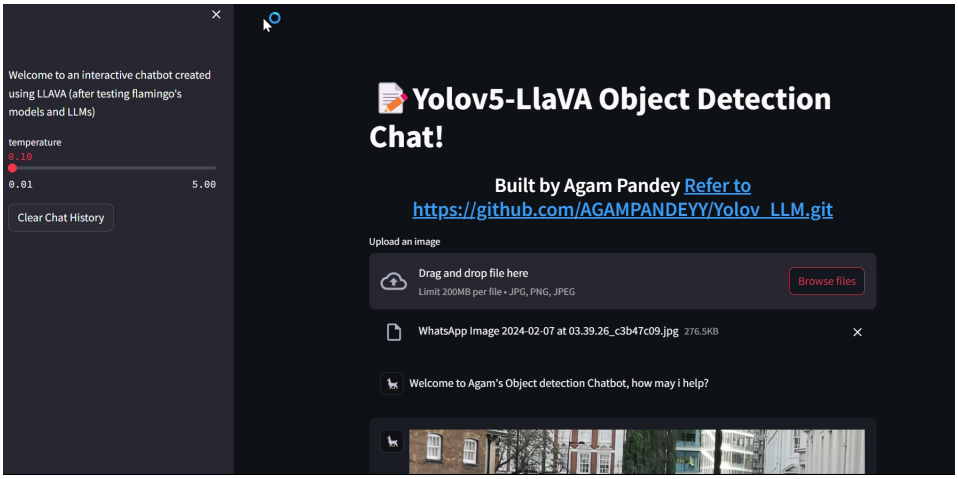


Figure 3: Streamlit web app

6. Llama 13b 4-bit quantized model called and QLORA fine tuning on general Q/A and image captioning dataset, but the notebooks crashed.
7. Llava 13b chat model called using Replicate API (had issues with loading huge model on Google colab)
8. Inferences on image dataset extracted from Yolov5 object detection video
9. Created a web interface application using Streamlit and ngrok to chat with the Llava 13b model

4 Final Results

The final result was a non-trained LLava13b chatbot that could integrate with Yolov5 model for object detection and Visual Question Answering through a web applicationlike-

- The number of cars present?
- The numbers of people?
- Are the cars parked illegally?

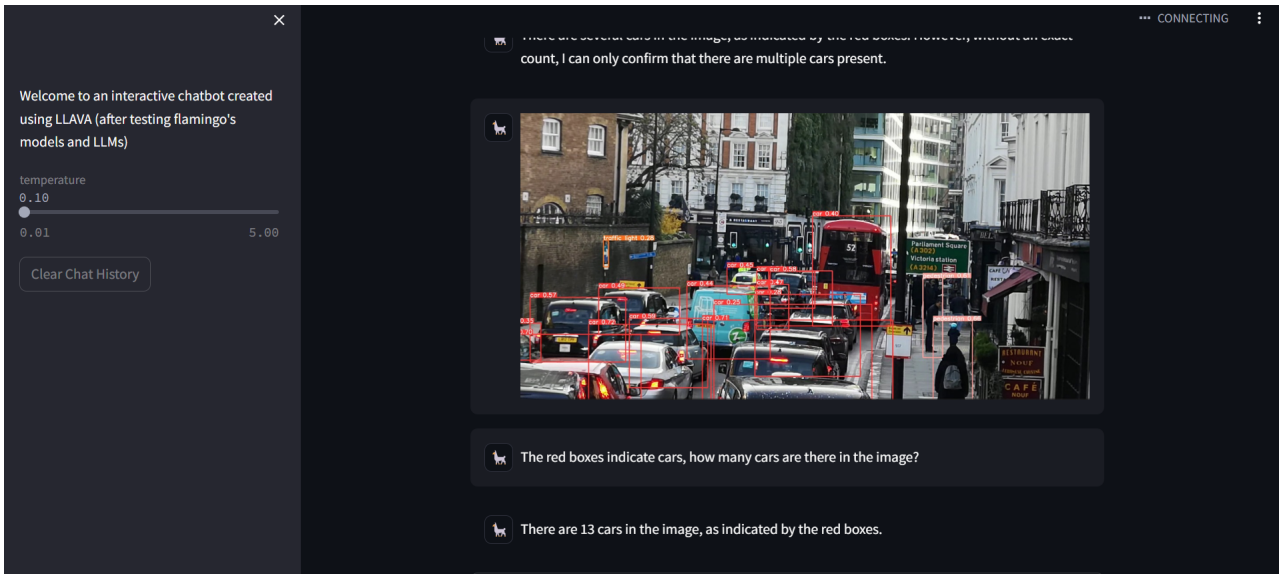


Figure 4: Model shows the count of cars andpedestrian in the image

- Is there any accident?

5 Future scope

Here,I will talk about the future expansion of the project i could work on if given good working environment with peers.
As mentioned in the proposal (link attached above) the expansion could be a mobile/web application for taking either real time video inputs and query any thing related to videos.

- A store owner could use it for inventory management
- A building inspection could be done by drone video capturing and streaming it to the app then VQA
- Can be used in medical field, patients or dermatologists can record videos of skin, ask query related to the diseases,etc

The furture expansion requires:

- Quality dataset for specific inudtry
- Good computation resources such as HPCs
- Development for smooth integration of Object detection and VQA both through a same app.
- My study on Multimodal models was limited to 2 weeks, if we research in this field,we could find more capable LMMs or ways to increase the accuracy.