Video Object Detection Integrated LLM Chatbot

Project Domain -- *<u>Computer Vision and Natural Language Processing using LLMs</u>*

Name- Agam Pandey

Enrollment | Branch - 22113009 | Civil Engineering

Mail id- agam_p@ce.iitr.ac.in

Phone No- +91- 9118205853

Problem description:

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.

<u>Inventory management can be simplified</u>. 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?

<u>Traffic Management</u>. 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? Traffic violations can also be listed (if we work on speed detection)

An advanced **CNN supported by the LLM model** 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 Models to query anything related to objects**.

The project implementation will be divided into 3 phases:

- Data Collection/Cleaning
- Object Detection Phase

- LLM Chatbot Phase

I have experience in working with LLMs (Llama2 7b Chat model) and will learn concepts of Computer vision Object detection throughout the project.

Data Collection Phase:

First, the aim will be to arrange a Dataset to train our model on.

For the dataset, I have decided to work on **BDD100K Dataset** which is the largest open driving video dataset on the internet. The model will be trained on this dataset.

Source-https://doc.bdd100k.com/download.html

Data will include 100k videos and images.

Object Detection Phase:

Now, the aim will be to start Object Detection.

What are we aiming to classify?

- Semantic Segmentation
- Instance Segmentation
- Panoptic Segmentation

Among the above 3, the best method to identify *things* and *stuffs* will be using **Panoptic segmentation (**to give semantic labels as well as instance identifier to a pixel**)**

The dataset for Panoptic segmentation has the following classes:

0: unlabeled 1: dynamic 2: ego vehicle 3: ground 4: static 5: parking 6: rail track 7: road 8: sidewalk 9: bridge 10: building
11: fence 12: garage 13: guard rail 14: tunnel 15: wall 16: banner 17: billboard 18: lane divider 19: parking sign 20: pole
21:polegroup 22: street light 23: traffic cone 24: traffic device 25: traffic light 26: traffic sign 27: traffic sign frame 28:Terrain
29: vegetation 30: sky 31: person 32: rider 33: bicycle 34: bus 35: car 36: caravan 37: motorcycle 38: trailer 39: train 40: truck

Out of these 40 classes in the dataset, I will be working on classifying 30things and 10stuffs.

The file is in JSON and mask formats. The segmentation will be stored in RLE (run length encoding) or mask format(Mask format handles pixel segmentation overlap). After classification, I will store the JSON/Mask file into a COCO format file.

The resource to understand this- https://doc.bdd100k.com/format.html

Among models like Faster R-CNN, YOLO, etc for *instantaneous object segmentation*, *multi-object tracking*, and *segmentation tracking* using CNN, I will be using the **Single Shot Detector algorithm**, which aims at finding the perfect bounding box within which an object fits based on the previous training datasets.

The model for Object detection will be SSD MobileNet
ssd_mobilenet_v2_320x320_coco17_tpu-8.config

First, create a TensorFlow environment. The model runs faster in GPU so install CUDA and cudNN.

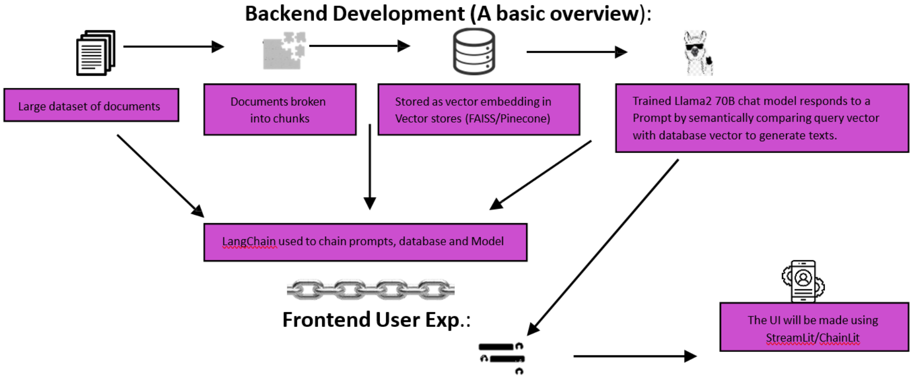
The model will run panoptic segmentation on the training set and store the output in a file.

LLM Chatbot Phase:

Once we get the file with Video object detection trained by SSD algorithm, we would then use it for creating a Q/A bot.

Steps for LLM bot:

- Use of **Llama 2 7b chat quantized model** (4 bit quantization)
- Dataset to train on.
- **PEFT** config (Parameter Efficient Fine Tuning) for fine-tuning the model on custom dataset
- **Langchain**- to chain all conversations, used to give memory to our conversation to the model.
- **Streamlit**- to create a web application interface for the Q/A bot


Backend Development (A basic overview):
Large dataset of documents
Documents broken into chunks
Stored as vector embedding in Vector stores (FAISS/Pinecone)
Trained Llama2 70B chat model responds to a Prompt by semantically comparing query vector with database vector to generate texts.
LangChain used to chain prompts, database and Model
Frontend User Exp.:
The UI will be made using StreamLit/ChainLit

Timeline:

Week 1- Creating a deep and thorough understanding of object detection SSD algorithm and mathematics involved. Understanding the concepts used in Large Language Models and dataset format.

Week 2- Start working on the Video object detection using MobileNet

Week 3-Start training LLM on the output metadata file generated from MobileNet

Week 4-Create a web application for the same using Streamlit.(I don’t have web development knowledge so I’ll use streamlit for easy development)