Neural Networks and Deep Learning CS6366

Team: Crimson Noir

1. Jithendra Katta -G34424060
2. Nikhil Reddy Kodumuru -G27606239

Project Proposal: InterACT: Utilizing Large Language Models for Interactive Object detection.

- **Problem Statement or Task:** Our project aims to generate an application that enables interactive object detection through natural language queries. Specifically, the application will respond to questions or perform tasks related to object detection and localization in videos. For example, when a user asks questions like "Where is the table?" and the application would respond by detecting and bounding the table in the environment.

- **Previous Work and Approach:** Previous work in computer vision and natural language processing has explored these concepts separately as chatgpt for Interactions and YOLO kind of models for object detection. We are interested in using the pre-trained YOLO World model which is trained under a publicly available dataset named as COCO (Common Objects in Context) for training and validation purposes.We want to fine tune it using some real time video data and Integrating it with an LLM which can be trained under some large datasets like cornell movie dialogs corpus dataset and fine tuned under own textual dataset. However, our approach differs by integrating these two domains and allowing for real-time object detection in a live environment.

- **Data Sourcing:** We plan to source video data from real time for fine tuning the model. Additionally, we collect some textual data for our large language model which can be used to give responses to the YOLO model.

- **Use of Neural Networks:** Our project heavily relies on neural networks, particularly convolutional neural networks (CNNs) for object detection tasks and transformer-based models for natural language understanding. We will leverage pre-trained models like YOLO (You Only Look Once) for object detection and fine-tune them using transfer learning techniques. In the same way, we would like to use transformer models such as BERT (Bidirectional Encoder Representations from Transformers) for language modeling tasks.

- **Division of Responsibilities:** The team's responsibilities can be divided as follows:

    ➢ Computer Vision Developer: Assigned to perform tasks related to using pre-trained models and fine-tuning them using environment data, optimizing it for real-time performance, and handling video processing tasks.

- ➢ NLP Developer: Assigned to perform tasks related to training and fine-tuning language models, and developing the interface for interacting with the application.

- **Measures of Success:** The success of our project will be evaluated based on several factors, including:

  - ➔ Accuracy of object detection in a real-time environment.
  - ➔ Effectiveness of Large language model and its response generation.
  - ➔ Integration of both models and effectiveness of the final hybrid model.
  - ➔ Scalability and adaptability to various domains or use cases.

- **Risks to Success:** Some potential risks to the success of our project include:

  - ➢ Difficulty in integrating object detection and language understanding components and Most importantly, the computation is challenging.
  - ➢ Challenges in optimizing the application for real-time performance, especially on resource-constrained devices.
  - ➢ Data limitations or biases in the training data affecting the model's generalization capabilities.
  - ➢ Technical issues or unforeseen complexities during implementation that may delay the project timeline.