COSE474-2024F: Final Project Proposal
"CLIP-based Real-Time Feedback System for Image-Text Understanding"

Taehee Jeong

1. Introduction

In autonomous driving systems, it is crucial to make immediate situational assessments and take appropriate actions. This project aims to implement a real-time feedback system using CLIP (Contrastive Language–Image Pretraining) to assess the relationship between images and text, providing action commands for vehicles in autonomous driving environments. The system will classify various traffic scenarios and generate commands such as "move left," "stop," or "move right," contributing to safer autonomous driving decisions.

2. Problem definition & chanllenges

The problem addressed in this project is how to enable an autonomous vehicle to make real-time decisions based on visual data. The system will define 10 specific scenarios, such as "move left," "stop," or "move right," and use CLIP to classify these scenarios accurately. The main challenges include ensuring real-time performance while maintaining accuracy in image-based command generation, as well as handling the complexity of diverse road conditions and potential data limitations.

3. Related Works

Recent research in autonomous driving action prediction has typically involved complex setups combining multiple sensors such as video and LiDAR. However, this project simplifies the approach by focusing on image-based decision-making using CLIP and leveraging zero-shot learning. CLIP's ability to associate images with text commands without extensive retraining makes it suitable for this task.

Zero-shot learning allows a model to classify new categories it hasn't explicitly trained on. Studies have shown CLIP's effectiveness in image-text retrieval, object classification, and captioning, which are relevant to real-time decision-making in autonomous driving. This project will build on these works by applying CLIP to traffic scenarios.

4. Datasets

This project will use the A2D2 (Audi Autonomous Driving Dataset), specifically the Front Center Camera images. The dataset consists of high-resolution images captured in various traffic scenarios. These images will be used to train and evaluate the system's ability to classify different commands based on road conditions.

5. State-of-the-art methods and baselines

The project will leverage CLIP's zero-shot learning approach to evaluate image-text associations for action classification. The performance of this simpler image-based method will be compared against more complex state-of-the-art systems that typically combine video and LiDAR data for object detection and decision-making in autonomous vehicles.

6. Schedule

- **10/10 - 10/17**: Submit project proposal and refine research plan
- **10/18 - 10/20**: Prepare and preprocess the A2D2 dataset, particularly the Front Center Camera images
- **10/21 - 11/01**: Implement the CLIP model and run initial tests with preliminary datasets
- **11/02 - 11/15**: Complete command classification system for 10 traffic scenarios and run evaluation tests
- **11/16 - 12/01**: Final system testing, performance evaluation, and optimization
- **12/02 - 12/10**: Final report writing, code documentation, and submission