Zero-Shot Object Detection using Grounding DINO

Mentor: Richa Thakur

Abstract:
This project explores Zero-Shot Object Detection using the Grounding DINO framework. Unlike traditional detectors trained on fixed datasets, Grounding DINO leverages vision-language alignment to detect unseen objects based on textual prompts. In this work, the model is applied to the Open Images V6 Food dataset containing 54 food classes, which are not part of the COCO dataset. The results demonstrate qualitative success in detecting new objects with text prompts, highlighting the potential of open-set object detection.

1. Introduction

Object detection is essential in computer vision for identifying and localizing objects. Conventional models are limited to predefined classes, requiring extensive labeled data. Zero-Shot Object Detection (ZSD) solves this by enabling detection of unseen objects using only textual prompts. Grounding DINO, a state-of-the-art open-set detector, aligns image features with text embeddings through cross-modal attention. This project implements Grounding DINO under the guidance of Richa Thakur, evaluating it on a food dataset outside the COCO domain.

2. Dataset and Tools

- **Dataset:** Open Images V6 – Foods dataset from Roboflow (54 food-related classes).
- **Pre-trained Model:** Grounding DINO (official IDEA Research repo).
- **Weights:** Official pre-trained weights.
- **Environment:** Python, PyTorch, PIL, OpenCV, NumPy.

3. Methodology

1. Cloned the Grounding DINO repository and loaded pre-trained weights.
2. Performed inference on single and multiple images using textual prompts.
3. Visualized detections with bounding boxes (red = predicted, green = ground truth).
4. Saved predictions in COCO format and evaluated using COCO metrics.

4. Results

Grounding DINO successfully detected food objects using zero-shot prompts. Although evaluation scores are low due to the domain shift (COCO → Food), qualitative results demonstrate its zero-shot capability.

Metric	Value (Food Dataset, Zero-Shot)	Notes
mAP@0.5:0.95	0.4 – 0.5 %	Mean Average Precision across IoU thresholds (low due to domain shift)
Precision (IoU=0.5)	~1 – 2 %	Correct detections / total predictions
Recall (IoU=0.5)	~2 – 3 %	Detected objects / total ground truths
Small Objects	Very Low	Model struggles with small food items
Medium Objects	Moderate	Better detection compared to small
Large Objects	Higher	Larger food items are easier to detect

Since Grounding DINO was trained on the COCO dataset but tested on unseen Food dataset classes (54 categories), the evaluation scores are expectedly low. This demonstrates its zero-shot capability rather than high accuracy.

5. Challenges & Limitations

- Domain shift from COCO to Food dataset reduced accuracy.
- Small objects were harder to detect.
- Limited by prompt design and lack of fine-tuning.

6. Future Enhancements

- Improve prompt engineering strategies.
- Fine-tune Grounding DINO on food-specific data.
- Integrate with SAM (Segment Anything Model) for segmentation.

7. Acknowledgment

I sincerely thank my mentor, Ms. Richa Thakur, for her constant guidance, patience, and support throughout this project. Her inputs were invaluable in completing this work successfully.

8. References

[1] IDEA Research, Grounding DINO GitHub Repository:
https://github.com/IDEA-Research/GroundingDINO
[2] Roboflow, Open Images V6 Food Dataset:
https://universe.roboflow.com/foodai/open-images-dataset-v6-foods