ZERO-SHOT-OBJECT-DETECTION

Priyanka Singh

1.Introduction

This project demonstrates the application of Grounding DINO, a state-of-the-art open-set object detector, for zero-shot detection. Unlike traditional models that are limited to a predefined set of classes, Grounding DINO can detect objects based on natural language prompts, allowing for flexible and dynamic detection tasks without requiring model retraining.

2.Problem Statement

The primary challenge addressed is the inflexibility of conventional object detectors. This project aims to overcome this limitation by enabling the detection and localization of objects described via text prompts. The core objectives are to build a functional detection pipeline, evaluate its performance quantitatively and qualitatively, and analyze the impact of different prompt formulations.

3.Objectives

Pipeline Implementation: Successfully implemented a zero-shot object detection pipeline using the IDEA-Research/grounding-dino-base model.

Model Evaluation: Evaluated the model's ability to detect unseen object classes using a subset of the COCO 2017 Validation dataset.

Prompt Analysis: Explored how different prompt strategies—minimal, descriptive, and synonym-based—affect detection performance.

Result Visualization: Visualized model outputs with bounding boxes and confidence scores, and generated charts to analyze evaluation metrics.

4.Methodology

Dataset: A subset of 60 diverse images from the COCO 2017 Validation set was used for testing.

Model: The IDEA-Research/grounding-dino-base model from Hugging Face was employed, leveraging its pre-trained language and vision capabilities.

Pipeline:

1. **Milestone 1:** Selected and prepared the test image set and defined baseline prompts.
2. **Milestone 2:** Performed initial zero-shot detection with baseline prompts and saved the results.
3. **Milestone 3:** Performed a comprehensive quantitative evaluation using COCO metrics (mAP) and a qualitative evaluation through visual overlays.
4. **Milestone 4:** Conducted prompt variability experiments using descriptive and synonym-based prompts and analyzed the impact on detection performance.

Tools: PyTorch, Transformers, pycocotools, Matplotlib, Pandas, Kaggle GPU.

5.Result and Observation

The model's performance was analyzed both quantitatively (using COCO metrics) and qualitatively (through visual inspection).

Quantitative Analysis (mAP): The overall performance was modest, reflecting the difficulty of the zero-shot task. The AP scores across all strategies were in the range of 0.10 to 0.15, with the AP@.50 scores being significantly higher, indicating that while the model could find objects, its bounding box localization was less precise at stricter IoU thresholds.

Analysis of Prompt Strategies:

- Minimal vs. Detailed: Descriptive prompts like "a photo of a person." generally yielded slightly better and more consistent results than minimal prompts like "a person." This suggests that providing additional context helps the model better ground the text to the image content.
- Synonyms: The synonym-based prompt strategy often led to decreased performance. The model struggled with uncommon synonyms, occasionally resulting in missed detections or incorrect classifications.

Analysis of Object Size: The model's performance was significantly impacted by object size. As expected, the model performed best on large objects, followed by medium objects, and struggled most with small objects, which often went undetected. This is a common failure mode in object detection.

6.Conclusion And Future work

This project successfully demonstrates the implementation and evaluation of a zero-shot object detection pipeline using Grounding DINO. The key findings are:

- Grounding DINO is effective for detecting a wide range of objects based on natural language, though performance is a challenge for zero-shot tasks.
- Prompt formulation critically influences results; descriptive prompts generally outperform minimal or synonym-based ones.

- The model's performance is highly dependent on object size, with detection of small objects remaining a significant challenge.

Future Work:

- Model Exploration: Investigate alternative open-vocabulary models like YOLO-World or OV-DINO for comparative analysis.
- Expanded Dataset: Test the model on a larger and more diverse dataset (e.g., 200-500 images) to ensure more robust and generalizable results.
- Interactive Application: Develop a lightweight, user-friendly web application to allow real-time zero-shot detection with custom prompts.