Enhancing Robotic Vision through Vision-Language Models: A Trash Sorting Framework with Physically Grounded Understanding

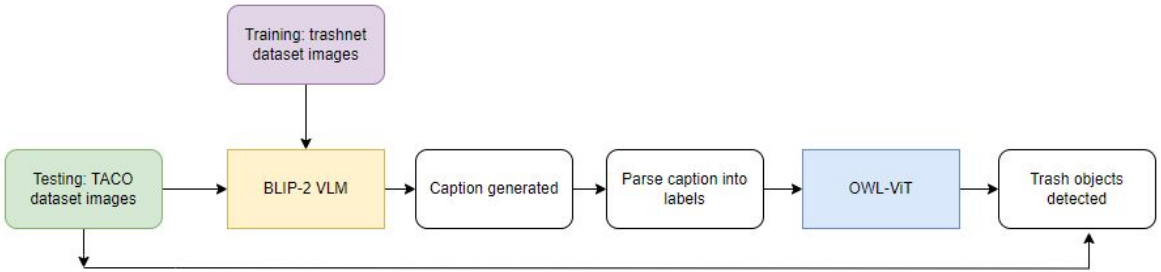
Joe Lisk, Prakrit Sinha, Sam Komonosky

Problem Description

- Trash detection has few datasets with small number of data.
  - TACO has 1500 images, 4784 annotations
- Definition of trash is ambiguous
- Trash classification is highly context-dependent
- “Trash” objects can have a large, diverse number of class labels.
  - Difficult to train on every kind of label
-  

Methodology

- Trained the BLIP-2 VLM on trashnet dataset (~2500 images).
- Original dataset had labels in the form of a numeric id (0-5), in order to generate captions we mapped these label ids into text categories (0:cardboard, 1:glass, 2:metal, 3:paper, 4:plastic, 5:trash).
- Captions were then parsed and fed as labels into a zero-shot (model which can predict on unseen class labels) object detection model: OWL-ViT



Challenges

Problem: Fine-tuning PG-Instruct BLIP on custom datasets proved intractable

Problem: COCO objects had objects where trash/not trash classification was ambiguous

Problem:

- Generated caption words not always usable as a set of class labels
- Generated caption inconsistent for single image
- OWL-ViT detects objects that would be very unlikely to be trash (e.g tables)
- Generated caption is sometimes not that useful
    - Ex. `can can of soda can with green and white background`

Problem:

- Training + Inference on VLM + Inference on OWL-ViT computationally expensive

Solutions

PG-Instruct BLIP issue:

- Used BLIP-2 for fine tuning (precursor to PG-Instruct BLIP)

COCO Dataset issue:

- Fine-tuned BLIP-2 on Trashnet dataset (well defined categories of trash; all objects trash)

Caption-related issues:

- Match words in caption with classes in unseen TACO dataset to use as labels for OWL-ViT
- Potential solution: natural language processing methods for caption refinement

Computational complexity issue:

- Open issue

Baseline method

- Mask RCNN used for TACO dataset, segments compared against ground truth segments.
- Faster RCNN can be used for general object detection. Bounding box compared against ground truth bounding box.
- Metric: Bounding box score using IoU. Class label score using AP. (TACO uses slightly different metric)
- Comparison metric still an open issue with our work

$$Score = \begin{cases} \max_i p_i, & class\_score \\ 1 - p_{N+1}, & litter\_score \\ \frac{\max_i p_i}{p_{N+1}+\epsilon}, & ratio\_score \end{cases}$$

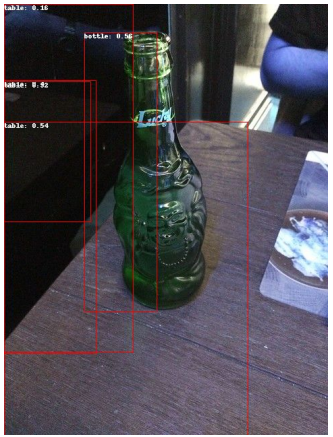
[2]

Results:

- Able to detect trash objects with unseen labels in training data (zero-shot)
- Able to detect cluttered objects
- BLIP-2 image captioning:

caption →

```
cardboard bowl with white background
```



- OWL-ViT zero-shot:

Future Work

This project:

- Create a metric to compare performance of our method with baseline

Future research ideas:

- Increase the number of images in trash datasets for greater scene contexts and richer labels
- Single architecture for caption -> label generation -> zero-shot object detection with less computational cost
- Include scene context in the label for object

References

[1] Dive into VLM article: A Dive into Vision-Language Models (huggingface.co)

[2] Taco Paper: https://arxiv.org/abs/2003.06975

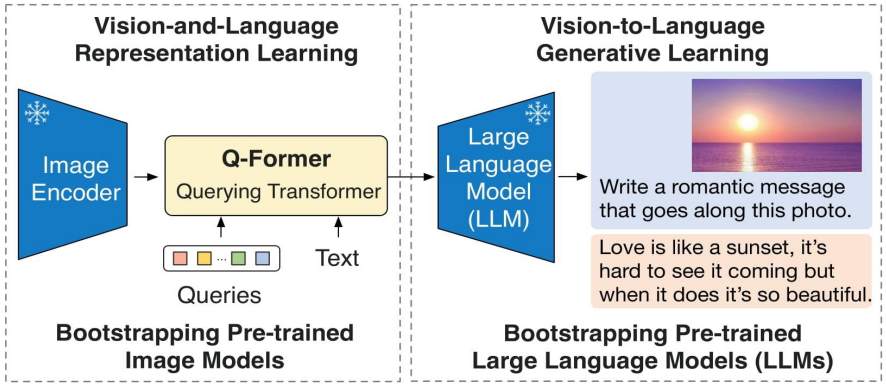
[3] Taco Dataset: http://tacodataset.org/

[4] Trashnet Dataset: https://huggingface.co/datasets/garythung/trashnet

[5] BLIP-2 Paper: https://huggingface.co/docs/transformers/en/model_doc/blip-2

[6] OWL-ViT Paper: https://huggingface.co/docs/transformers/en/model_doc/owlvit

BLIP-2 Architecture [5]



OWL-ViT Architecture [6]