Université de Montréal

IFT3150
Projet en informatique

FINAL REPORT
OPEN-SET OBJECT DETECTION USING CLIP MODEL

Van Nam Vu
van.nam.vu@umontreal.ca

Report to
Professor Liam Paull
Professor Frédéric Dupuis

April 26, 2023

REMERCIEMENT

I would like to express my special thanks to my supervisor Liam Paull and Continual-object-detection team (Mostafa ElAraby, Atharva Chandak and Mahtab Sandhu) for all their supports for me on this project □

ABSTRACT

This project aims to capture the object which is not a part of the class in training set from a dataset for autonomous driving. Due to the quality of the unlabeled image from the lost-and-found dataset, it is a challenge for us to evaluate the performance of our pipeline. Inspired by other papers for “Open-Set object Detection” and the recent advance in Contrastive Language-Image Pretraining (CLIP), we design a pipeline using this multi-modal to detect “unknown objects” in the image based on the context that we want. CLIP is known as a very efficient model for image classification and object detection. With a lost-and-found dataset, our pipeline can detect it, but the result may not be as we expected. In the future, we need to work more on the image preprocessing and parameter fine-tuning to improve our pipeline. In addition, we want to link this idea for continual object detection which is an important subject for Data Science and AI.

INTRODUCTION

In the age of data science and AI, robotics and autonomous vehicles are popular in both research and application. Every car brand has installed their autonomous driving system into their cars. This is the future of the world. Nevertheless, all state-of-the-art (SOTA) models for Object Detection were trained with a huge dataset with limited class label. It poses a question whether the model can capture an object not included in its dataset. For example, if the radar of an autonomous vehicle sees an elephant on the street, what will happen? Due to the nature of unknown object anticipation, we name this project “Open-Set Object Detection”.

In our project, we want to use the feature of “Zero-shot object detection” of CLIP and test it with images from lost-and-found dataset. For computer resources, we use google Colab (with 40gb RAM, 16gb GPU) to run and test our pipeline.

LITERATURE REVIEW

Since 2014, numerous state-of-the art deep learning models for Object Detection have been produced. We have classic deep learning models such as CNN to their extended models like RCNN, Fast-RCNN. These deep learning models were trained from many big, labeled datasets using HPC (High Performance Computer) to fine-tune parameters and optimize the weights. Then, starting 2016, “one-stage” object detection algorithms like YOLO, which stands for “You Only Look Once” emerged. Deep learning models like CNN or RCNN have 2 stages:

(1) find objects, propose region.
(2) classify every single object from stage 1 and estimate its size with a bounding box.

On the other hand, deep learning models like YOLO have only one stage in which they try to detect and predict bounding boxes without the region proposal step. This method consumes less time and can be used in real-time applications. Another state-of-the-art deep learning model is Vision Transformer (ViT). This model uses self-attention mechanisms to process images (like Transformer in Natural Language Processing).

We also look into many papers related to Open-set Object Detection. One of their ideas is to detect unknowns (any sub-classes) in a known super-class[1]. Another main idea is to use Faster-RCNN architecture to detect objects and “The unknown-aware RPN labels the proposals that have high scores but do not overlap with any ground-truth bounding box as the potential unknown objects”[2]. After the detection stage, Category Discovery is implemented to cluster unknown objects to novel categories. Additionally, Few-Shot Open-Set

[1] https://arxiv.org/pdf/2207.09775.pdf

[2] https://openaccess.thecvf.com/content/CVPR2022W/L3D-IVU/papers/Zheng_Towards_Open-Set_Object_Detection_and_Discovery_CVPRW_2022_paper.pdf

Object Detection model can detect the data-abundant known objects, the few-shot known objects, and the unknown objects based on an unbalanced training data[3] by using Fast-RCNN architecture and optimizing a parameter called Unknown Decoupling Learner (UDL).

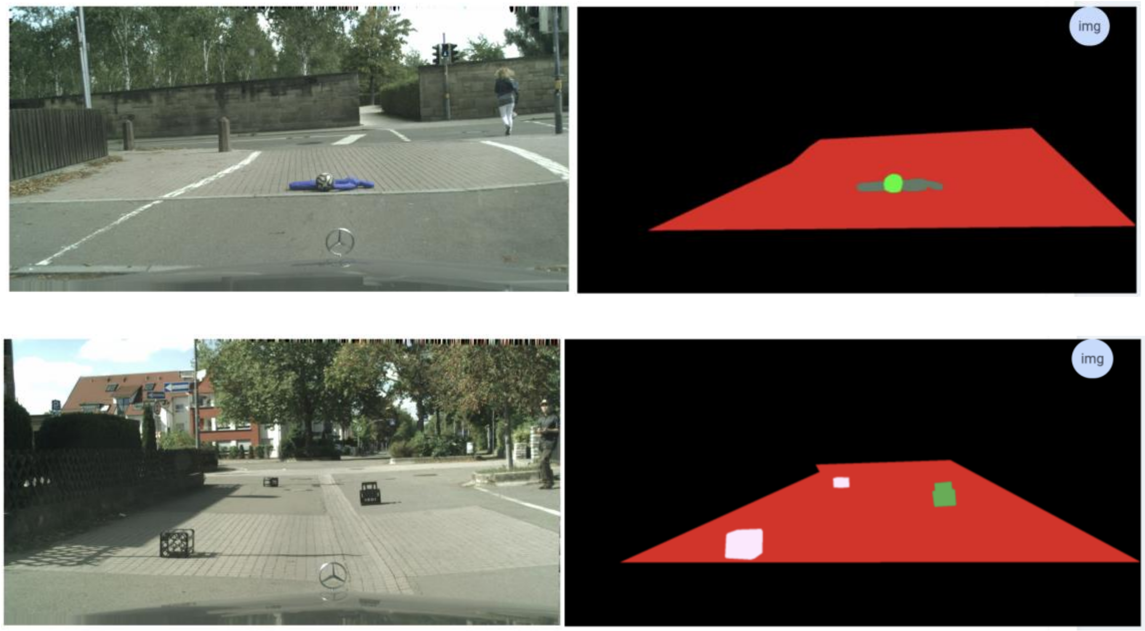
Moreover, we also reviewed an article named “Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition”. The authors “proposed to bound the approximate posterior by fitting regions of high density on the basis of correctly classified data points”[4] so unseen unknown objects can be detected from trained tasks.

GOALS

- Being able to load and preprocess images for autonomous driving dataset.
- Understanding the concept of SOTA model for Object Detection.
- Creating a pipeline to detect objects not from dataset (based on context of autonomous driving).
- Testing pipeline and have initial results to discuss.
- Clustering out-of-distribution objects and in-of-distribution objects.

METHODS

1. **Dataset and image processing**
   a. **Dataset:** Lost-and-found dataset
      i. **Description:** Image with small obstacles on the road (often caused by lost cargo)
      ii. **Size:** Train Set - 1036 image / Test Set – 1203 image, with image size = (1024,2048,3)
      iii. **Details:** Dataset give raw image and segmented image
      iv. **Sample:**



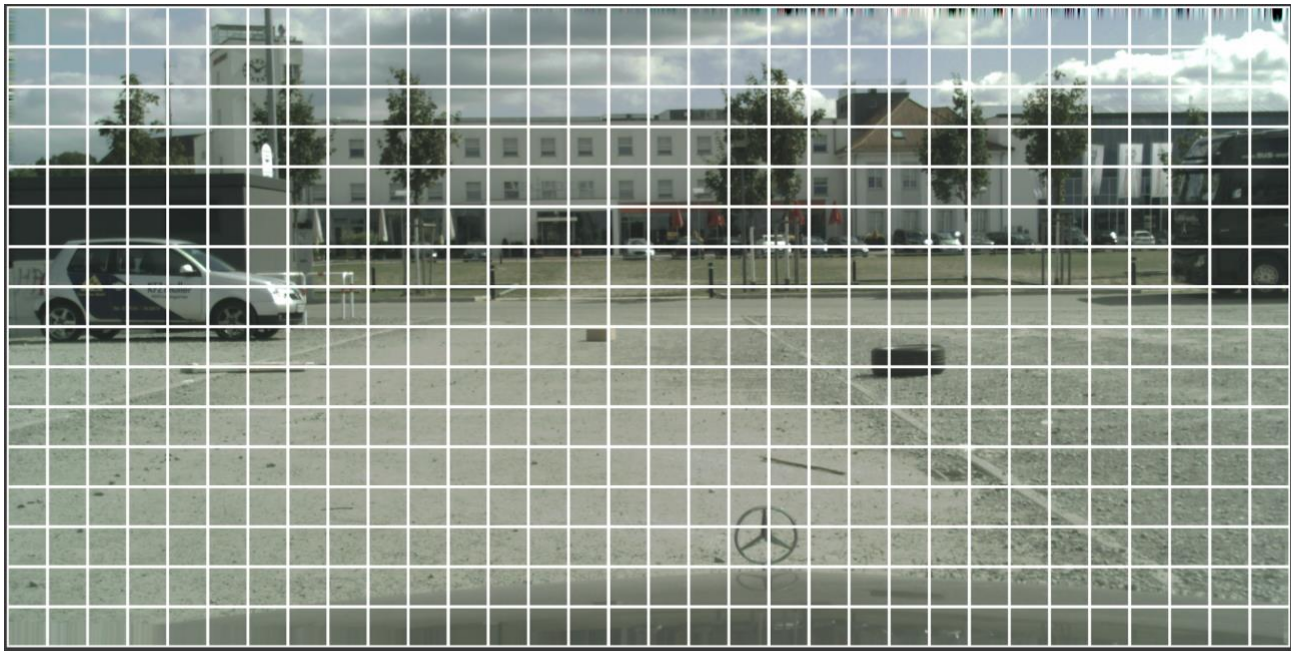
[3] https://arxiv.org/pdf/2210.15996.pdf

[4] https://arxiv.org/pdf/1905.12019v5.pdf

b. Processing: Since the image has shape of (1024,2048, 3), then we split it to patches with size of 64x64

Image.shape = (3,1024,2048)

Patches.shape = torch.Size([1,16,32,3,64,64])



2. CLIP Zero-shot object detection

a. CLIP model:

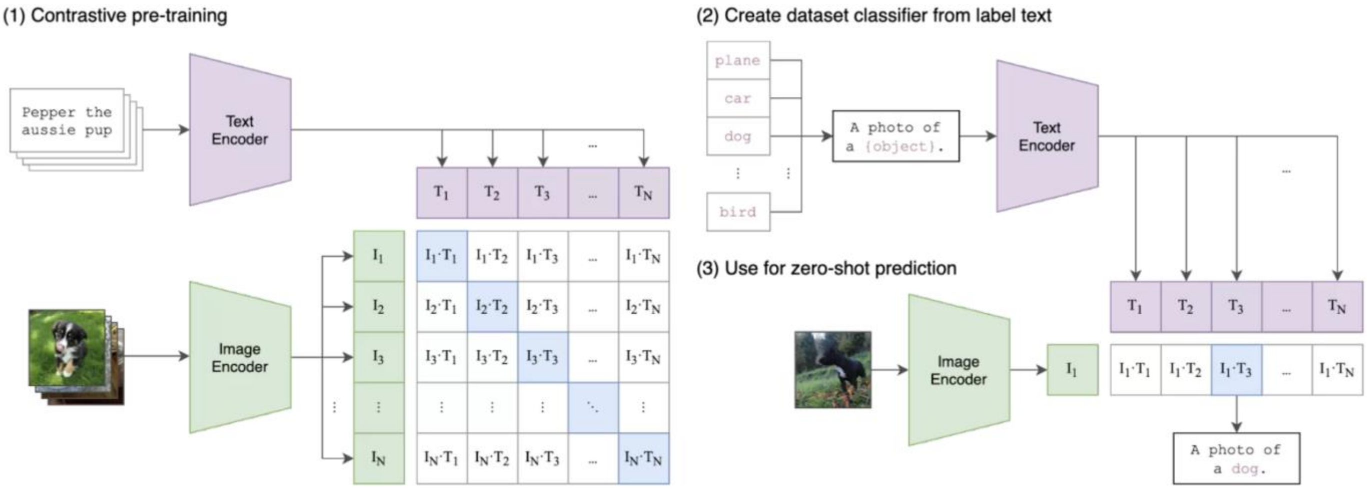
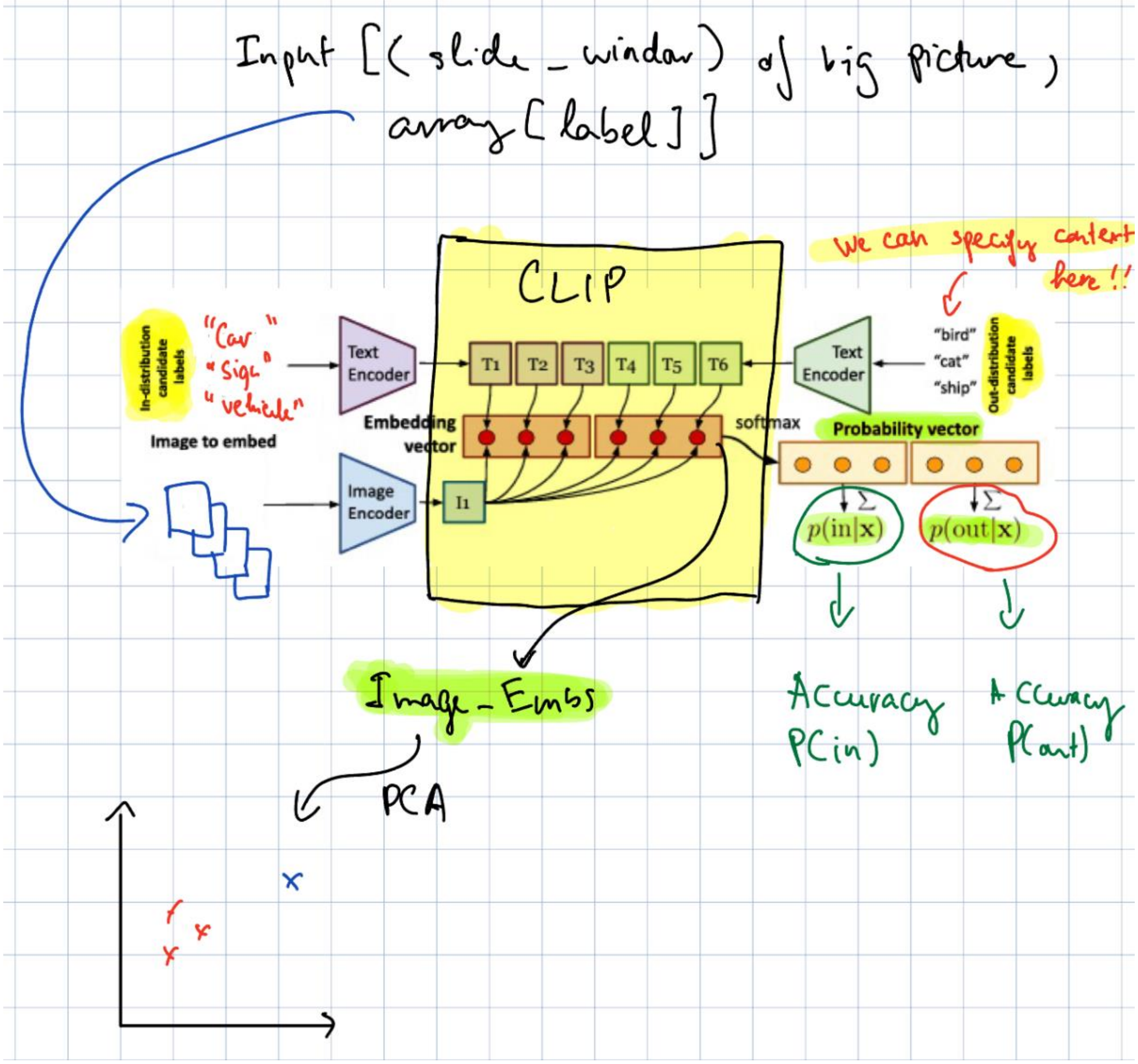


Figure 1.Schema of Zero-Shot prediction from article "Understand CLIP by OpenAI"

- (1): CLIP is a deep learning multi-modal between Computer Vision (CV) and Natural Language Processing (NLP). It is trained upon 400,000,000 (image, text) pairs. CLIP has two sub-models (Text Encoder and Image Encoder). The modal combines Text Encoder and Image Encoder to predict how likely/similarity between text and image.
- (2) + (3): Zero-shot detection: by giving a list of prompt text and an image, CLIP will calculate the embeddings, softmax() to get the probability of each class and return the predicted class by taking max(probabilities).

b. **Schema of our pipeline:**



Parameters:

- Text_prompt = ["Image of a car", "Image of object"]. For this parameter, we can have text description based on the context to detect unknown objects in which context that we want.
- Patches = 64
- Window = 2
- Stride = 2
- Threshold = 0.5

3. Embeddings plot with PCA

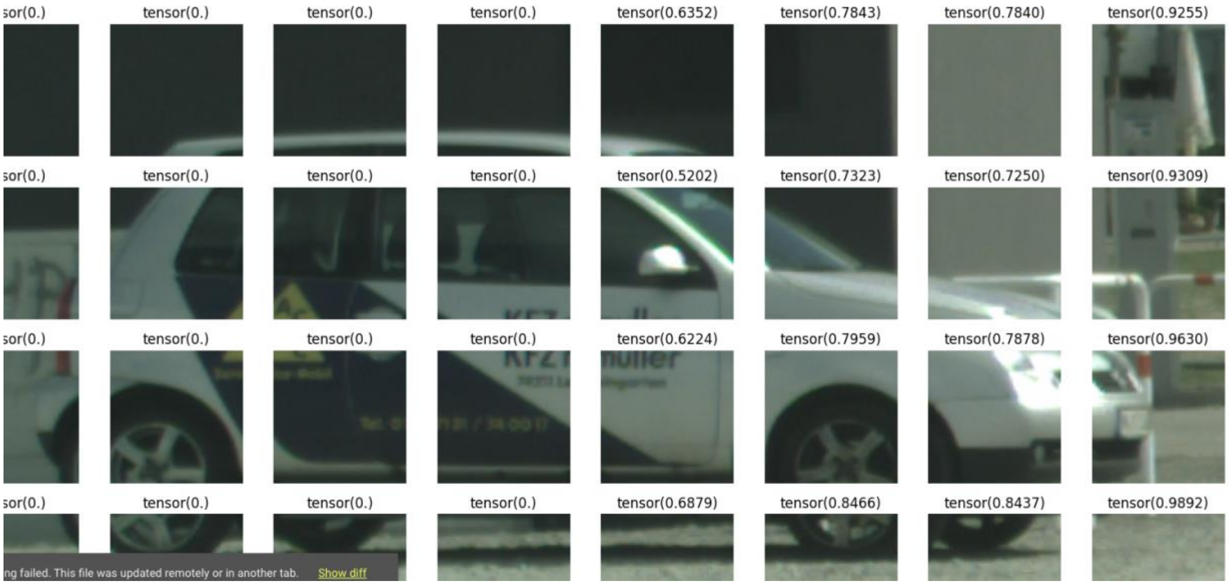
- Input: Dataframe containing image_embs (shape = (128,512)
- Parameters:
    - Components = 2 or 3
- Output: Plot with matplotlib

RESULTS

Since lost-and-found is an unlabeled dataset, we have the result displayed in the picture with the probability from CLIP.

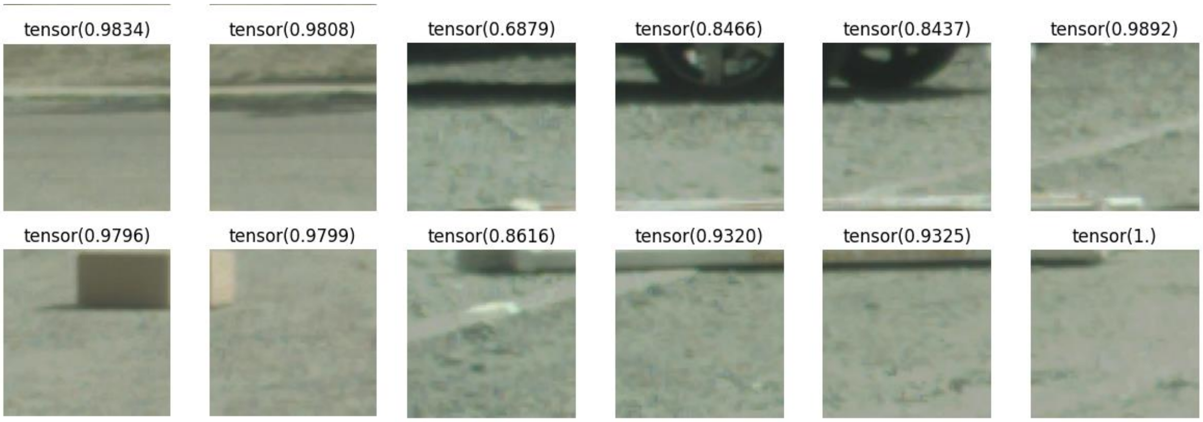
OUTPUT:

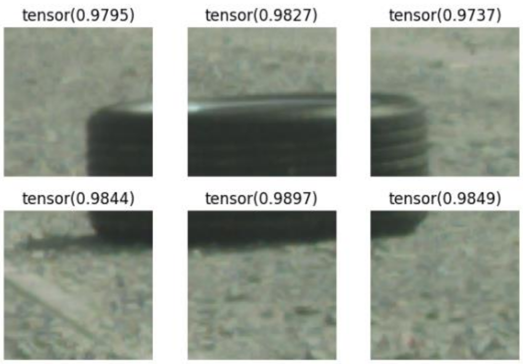
For “Image of a car”



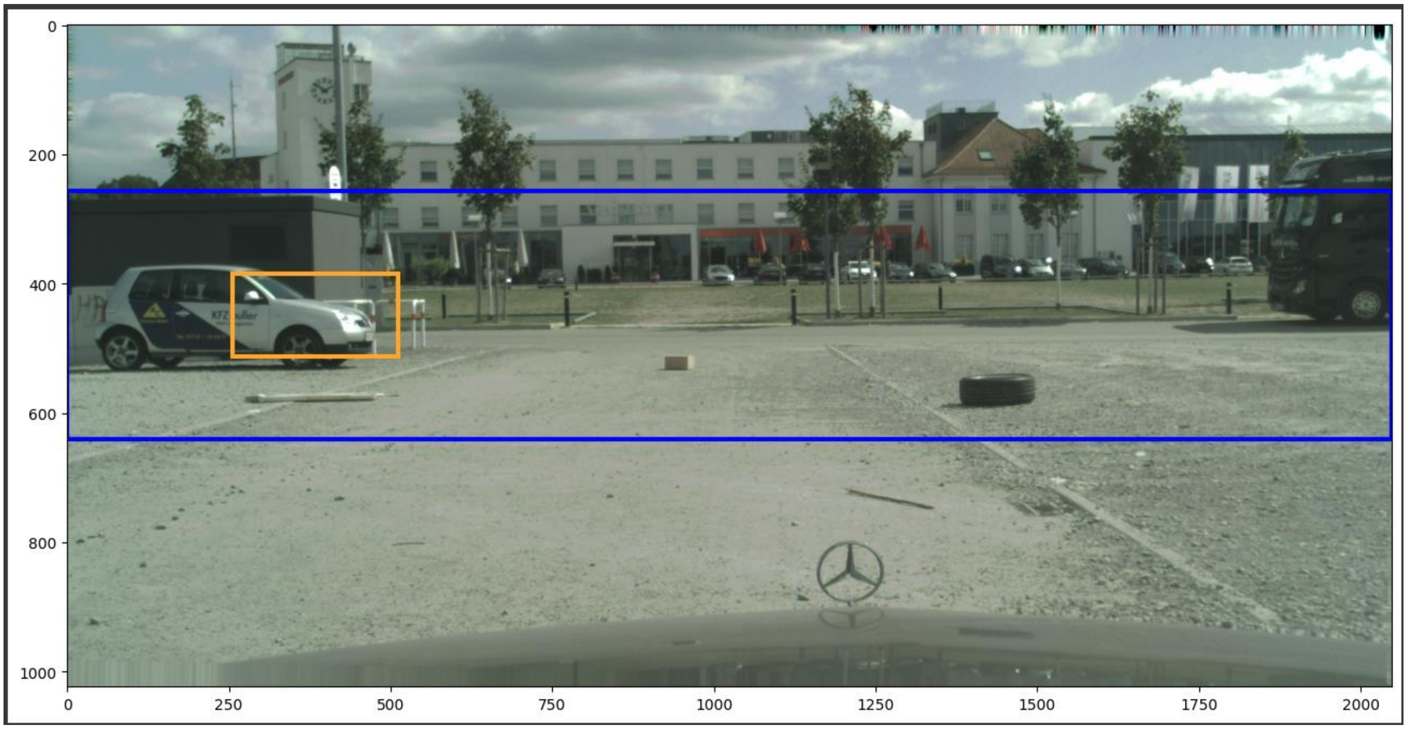
ng failed. This file was updated remotely or in another tab. Show diff

For “Image of an object”





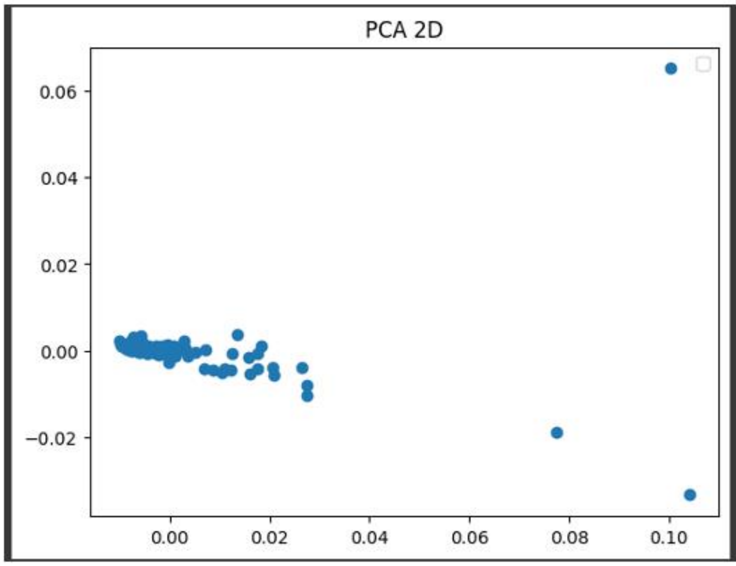
FULL OUTPUT:

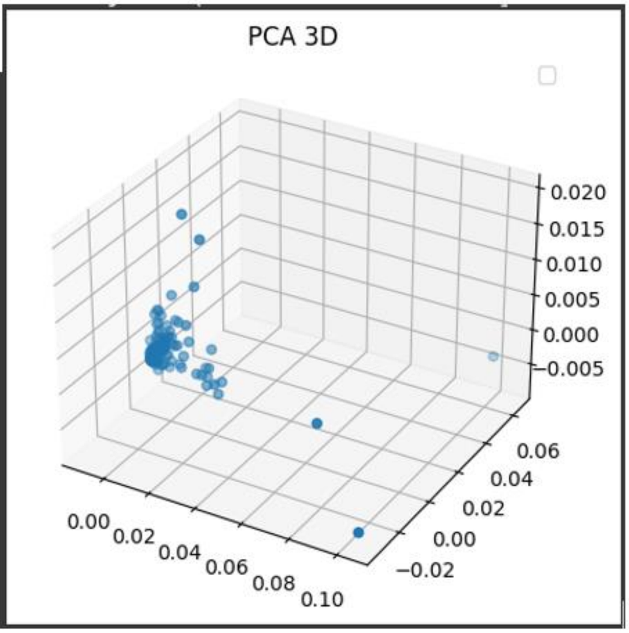


With orange bounding box for “car” and blue bounding box for “object”

EMBEDDINGS PLOT:

EMBEDDINGS RESULTS:





DISCUSSION / CHALLENGE

1. Discussion:

For computer vision (CV) models, the Lost-and-Found dataset is an intriguing collection of images that presents a special challenge. The collection consists of low-resolution photos of a variety of unexpected or challenging-to-identify things. The dataset is particularly valuable for examining the performance of CV models because of this feature. Researchers can evaluate the models on this dataset to assess the benefits and drawbacks of various CV algorithms as well as which ones are better able to handle unexpected objects in low-resolution photos. Overall, the Lost-and-Found dataset offers a great chance to advance computer vision research and boost the precision of picture identification algorithms.

Contrastive Language-Image Pre-Training, or CLIP, is a potent computer vision method for zero-shot object detection. With CLIP, it is unnecessary to invest money on retraining the model to find new items. Instead, CLIP uses natural language processing (NLP) to its full potential to provide the model the ability to identify things solely from their textual descriptions. This makes CLIP very helpful for zero-shot object detection because it can identify things that were not part of the initial training set. CLIP is a crucial tool for developing the area of computer vision since it allows researchers to save time and money while still obtaining precise object detection.

It seems that our model's accuracy at this time does not produce expected results, even the pca_plot is showing clearly the dissimilarity of the objects. We must analyze the picture preprocessing methods that we are employing for this specific dataset to address this problem. To make sure that we are accurately capturing the features of each image, we may also need to re-evaluate how we derive the embeddings array of text-image. Finally, normalizing the probabilities in our model may assist to increase accuracy. By examining these crucial areas, we may find and fix any potential flaws in our strategy and aim to increase the general accuracy of our model.

2. Challenges:

- Challenging dataset.
- Basic knowledge in Computer Vision, I will need more experiences/experiments to improve my approach.
- Computer resources.

CONCLUSION

In this paper, we have presented a basic approach for detecting unknown objects for self-driving dataset. CLIP is a potential multi-modal that opens a lot of opportunities for practical applications for both researchers and businesses to utilize. It has a high performance to detect objects as well as to generate caption based on image. Our model can detect unknown objects but at a basic level. Further work is necessary to make the modal more robust and to build benchmarks for more detailed evaluation of our modal. Personally, this is my very first steps to explore this field, and I am happy that I have a chance to learn and to apply my knowledge into the production of something that I am very interested in □

For future contribution, this work can be a part of Continual-Learning, where we can also use CLIP to label the unknown object that we detect. This potential model can help us to save a lot of money and resources for pre-trained steps. The only obstacle to achieve efficiency is to improve data preprocessing and the pipeline of the dataset application.

REFERENCE

1. https://arxiv.org/pdf/2106.03004v3.pdf
2. https://arxiv.org/pdf/2207.09775.pdf
3. https://openaccess.thecvf.com/content/CVPR2022W/L3D-IVU/papers/Zheng_Towards_Open-Set_Object_Detection_and_Discovery_CVPRW_2022_paper.pdf
4. https://arxiv.org/pdf/2210.15996.pdf
5. https://arxiv.org/pdf/1905.12019v5.pdf
6. https://analyticsindiamag.com/top-8-algorithms-for-object-detection/#h-8-yolo-you-only-look-once
7. https://paperswithcode.com/task/open-set-learning
8. https://huggingface.co/docs/transformers/model_doc/clip
9. https://towardsdatascience.com/using-transformers-for-computer-vision-6f764c5a078b
10. https://medium.com/analytics-vidhya/15-best-open-source-autonomous-driving-datasets-34324676c8d7
11. https://blog.roboflow.com/openai-clip/
12. https://cv-tricks.com/how-to/understanding-clip-by-openai/