Introduction to Object Detection and Segmentation

Dr. Mehmet Kerem Turkcan, Columbia University
October 24, 2024

Contents

Introduction

We will learn about object detection and segmentation problems, taking an experimentation-first approach.

This first handout for the course is focused on giving an overview of the field. The focus on *reproducibility* and *replicability* in machine learning allows people to test models easily, and for newcomers to contribute to the field.

- Learn about the current state-of-the-art in object detection and segmentation from a high-level view.
- Learn about popular datasets used for these approaches.
- Run demos online to see these algorithms in action.
- Through experimentation and reading, try to figure out some common issues with the current models and existing datasets.

Definitions

- In *object detection*, we aim to build machine learning models that seek to localize each *instance* of a subset of objects in an image spatially. In practice, for each object, we usually seek to find the coordinates of the rectangle (x-coordinate, y-coordinate, width and height) as well as the class of the object (person, car, cat, dog, etc.). That is, for each object we will seek to find 5 values.
  - An emerging approach today is *zero-shot* object detection; in these approaches, the model takes in a list of *object classes* to detect, and only detects those. These approaches use recent strides in large language models, such as ChatGPT.
  - *Pose estimation* is an important subproblem that supports many different use cases today. In pose estimation, we seek to estimate the position of some landmark points (usually corresponding to important human joints and face landmarks) along with an object, to try to understand their movements and gestures in a succint way.
- In *segmentation*, we aim to build machine learning models that seek to segment a scene, on a per-pixel basis. That is, for each pixel of the image, we want to find a single value.
  - In *object segmentation*, we only focus on foreground objects, such as people or animals, as opposed to sidewalks and building facades.
  - In *instance segmentation*, we seek to further separate each *instance* of an object with a separate identifier, so that we know which pixel which entity belongs to, in addition to their class.
  - In *panoptic segmentation*, we further seek to separate background pixels into their classes. *Panoptic segmentation* also has its uses: in a streetscape, knowing the location of roads, sidewalks and building facades could be crucial for mapping, and spatial understanding.

Getting Started

To train computer vision models, we first have to understand some datasets that exist in this space. What our models can learn is limited by what people have annotated in the past.

Figure 1: The VisDrone dataset contains aerial views of pedestrians and vehicles.

Figure 2: Example images from the MS-COCO dataset.

Detection

VisDrone The VisDrone dataset contains aerial views of pedestrians and vehicles. The dataset was collected from 14 cities in China. Link to Samples: https://www.kaggle.com/datasets/wwj0510/visdrone

Segmentation (And Detection)

COCO The COCO dataset contains a large variety of scenes and object classes, which has made it a baseline object detection and segmentation dataset. Pretraining a model on MS-COCO allows it to have good priors for other real world tasks. Link to Samples: https://cocodataset.org/#explore

OpenImages Open Images is a dataset of 9M images annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives. Pretraining a model on MS-COCO allows it to have good priors for other real world tasks. Link to Samples: https://storage.googleapis.com/openimages/web/visualizer/index.html

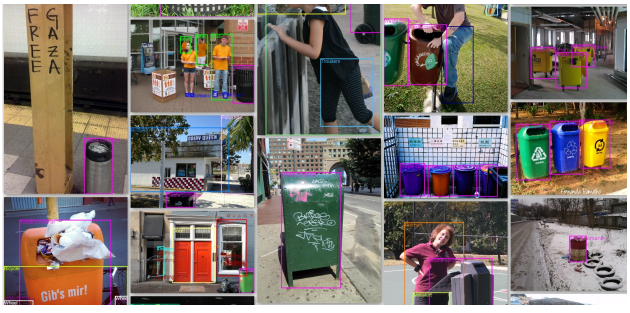

Figure 3: Examples from the OpenImages dataset.

Demos

Here, you will get to play with and learn from some of the state-of-the-art computer vision models.

- **Oneformer** OneFormer is a high-quality segmentation model. *Link:* https://huggingface.co/spaces/shi-labs/OneFormer
- **YOLO-World** YOLO-World is a real-time open-vocabulary object detection model. *Link:* https://huggingface.co/spaces/stevengrove/YOLO-World
- **YOLOv8 for Object Detection and Segmentation** Performs object detection and segmentation in real-time, trained on large datasets. *Link:* https://huggingface.co/spaces/fcakyon/yolov8-segmentation
- **Grounding Dino and OWL-ViT2** These are slower, higher-quality open-vocabulary object detection models. *Link:* https://huggingface.co/spaces/merve/GroundingDINO_OWL
- **Florence2** A vision foundation model trained to address vision tasks. *Link:* https://huggingface.co/spaces/SkalskiP/better-florence-2

Exercises

- Look online to find the classes in the COCO dataset. Can you point to any obvious classes from your daily lives that is missing? What is the most similar object in the dataset?
- Compare the streetscapes in the VisDrone dataset against the streetscapes of New York. What is missing? What would be surprising to a hypothetical entity that learned about the world only from VisDrone images? How can we collect data from New York to build a model that will work in New Orleans, or in Boston? What are some differences in the objects between these cities?