African Master's in Machine Intelligent
Computer Vision
Muhirwa Salomon

Abstract

This session explores Detectron2, a PyTorch-based object detection library, and its instance segmentation capabilities using a Mask R-CNN model. The model performs well on well-lit, separated objects but struggles with inverted orientations and hidden body parts. Contextual information plays a crucial role in recognition, and the model can overcome blurriness to make accurate predictions based on context

Introduction

In this practical session, we will familiarize ourselves with Detectron2, an object detection library written in PyTorch. Detectron2 implements state-of-the-art object detection models. It also provides a "model zoo" a library of models trained on a variety of datasets. We will be going through the Detectron2 tutorials which offer a step-by-step exploration of Detectron2. Throughout this lab we will use colab notebooks to run detectron2 models. This is essential as the models will run much faster on the colab GPU compared to our personal laptops.

1 Instance Segmentation

1.1 Model architecture

The model used is a Mask R-CNN model, it uses a ResNet + FPN backbone with standard conv and FC heads for mask and box prediction, respectively. It's pre-trained on the COCO dataset for object detection and instance segmentation and runs on 12 images from our personal collection.

1.2 Examples

Let's visualize four predictions: 2 corrects and 2.

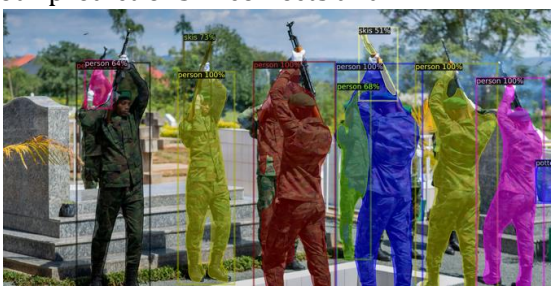

Figure 1: Correct prediction 1

Figure 2: Correct prediction 2

Regarding the correct predictions, we have seen that the two images are very well lit, the objects on these images are well separated, the resolution of these images is good and that the orientation of the images is correct, all this favors an easy recognition by the model.

Figure 3: Incorrect prediction 1

Figure 4: Incorrect prediction 2

The model makes errors in some predictions, taking the first image we can see that the model recognizes the person on this image but not the jerrycan that the person have

Figure 4.

"In this particular prediction, we observed a missed detection where the model failed to recognize the sugar cane in the image. The orientation of the sugar cane appears to be inverted, with the top part of the cane facing downwards instead of upwards as commonly seen. This unconventional orientation likely contributed to the model's inability to correctly identify the sugar cane in this picture, leading to the missed detection. Orientation plays a crucial role in object perception, and deviations from typical orientations can pose challenges for the model's recognition capabilities."

1.3 Observations

2 I have observed that contextual information plays a crucial role in the recognition of objects within an image. In one instance, the model mistakenly identified a large fire as a person, highlighting the influence of surrounding elements on the model's interpretation. Additionally, in another image, the model failed to recognize a person who was completely covered with cans from head to toe. Despite the blurriness in both cases, it is intriguing that the model was able to identify these objects accurately. This suggests that the model's recognition capabilities can sometimes transcend the limitations of image quality and still make meaningful predictions based on contextual cues.Pose estimation

2.1 Model architecture

The model utilized in this scenario is built upon the Mask R-CNN methodology, which extends the Faster R-CNN framework. By combining object detection and instance segmentation, Mask R-CNN can identify objects while also generating precise pixel-level masks for each individual instance. The backbone architecture employed in this model is ResNet-101-FPN, which is pre-trained on the COCO dataset. The COCO dataset comprises more than 200,000 labeled images covering 80 different object categories. Notably, certain object categories, such as humans, include annotations for keypoints, enabling the model to capture spatial information about specific parts of the objects.

2.2 Examples

Analyzing the model's predictions, it becomes evident that the model encounters difficulties in accurately detecting key points of individuals in images with multiple people. Moreover, the model struggles to precisely locate key points when individuals are facing away from the camera. These limitations arise from the model's inability to fully observe and capture the complete structure of limbs (such as arms, hands, and eyes). However, intriguingly, the model showcases the ability to recognize key points when a person is holding a photo and even successfully identify the key points of the person depicted in the photo

2.3 Observations

The model exhibits satisfactory performance when predicting body parts that are visible and not obstructed. However, it encounters challenges when it comes to accurately estimating keypoints for body parts that are partially or fully hidden. These difficulties arise in scenarios where objects or people are obscured by other elements within the image, such as foreground objects, loose clothing, or props. In such cases, the model's ability to precisely locate keypoints may be compromised due to the limited visual information available. Partial occlusion poses a significant hurdle in accurately estimating the positions of hidden keypoints, resulting in potentially less reliable predictions for these particular body parts.

Link Colab :
https://colab.research.google.com/drive/1Fnd2uxB2aKKG2Zj14FYq57I7xZD7062H?usp=sharing