Dude Where's my Mug? An Image Driven Dialogue

Tayyub Yaqoob
MSc in Big Data

Introduction

This project uses state-of-the-art Computer vision and Speech recognition libraries (programmed in python) along with Natural language processing to create a framework capable of answering simple questions about an image based on object labels & location.

Dataset & Methodology

Preliminary analysis has been done by using open source dataset (labelled) , MSCOCO (Microsoft Common Objects in Context) which is a large scale object detection, segmentation and captioning dataset with 330k images.

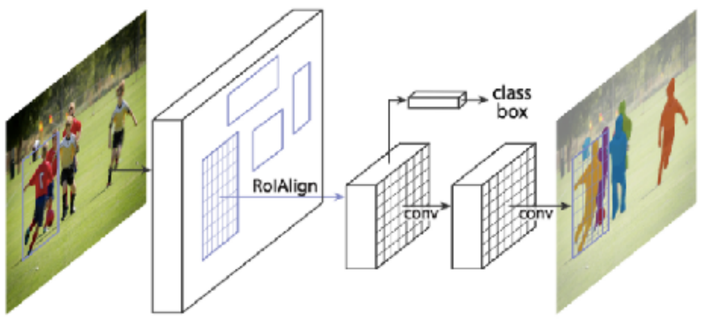

Figure 1: Instance Segmentation with MaskRCNN

For object detection, pre-trained Tensorflow model Mask RCNN is used. Other libraries includes, but not limited to: Numpy, gTTs (Google text-to-speech), Pygame (Playing audio), Spacy (Natural Language Processing) and Tkinter (User interface).

Results

Figure 2: Object Detection and Localization

This system is able to answer these kind of simple questions about images:

Q. What is closest to Book11?
Ans. Laptop8
Q. Is the Book11 below the Person2?
Ans. Yes
Q. Where is the Person2 in the image?
Ans. Person2 is to the right of Person1

Moreover, this framework is also able to generate descriptions about image based on their location.

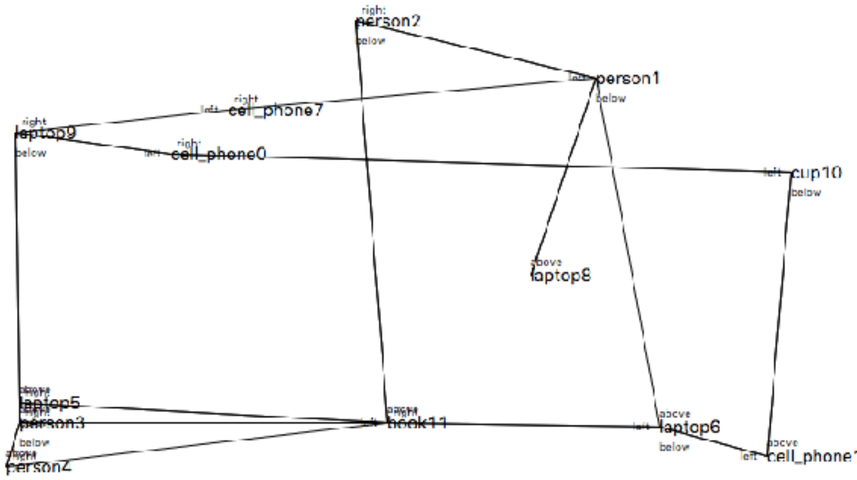

Figure 3: Spatial Graph by Location and Tag

Implementation

This systems makes graphical data more accessible by describing scenes to a blind person using a camera. Moreover, it's cloud based solution could be integrated to social media sites to build up knowledge bases and to improve their exploration of a photo downloaded from internet.

For further information contact: *Tayyub Yaqoob* *Tay00011@Students.stir.ac.uk*