Real-Time Object Detection

Prerna Kaul

Deep Learning, Data Science Department, Georgetown University, Washington D.C.

1. Abstract

Computer Vision has seen a lot of advancement since the introduction of neural networks. Object detection has been a quintessential part of this field and a lot of state-of-the-art techniques exist to tackle tasks in the same area. The rapid expansion of computer processing power combined with the rapid development of digital camera capability has resulted in equally rapid advances in computer vision capability and use. This paper talks about real time object detection using transfer learning from Mask R-CNN and OpenCV programming functions. The main objective of this technique is to segment the different objects in an image clicked in real-time from a webcam, mask those objects distinguishably and classify them.

2. Introduction

Humans glance at an image and instantly know what objects are in the image. Computers have always been good at dealing with numerical data but analyzing huge amount of data in images was not easy until the rise of deep learning techniques. Making the machines to "see" objects is a significantly harder task in Computer Vision than the traditional Image Classification. However, the most successful approaches to object detection are currently extensions of image classification models. Compared to Image Classification, Object Detection is considerably more complicated due to the simple fact that an image can have anywhere from zero to dozens of objects in them. So, the task at hand becomes recognizing all the different objects in the image, producing a bounding box around each of them, giving out their class labels and the probability with which the object is recognized to be a part of that class i.e. the accuracy of the object to belong to that class label.

The techniques of object detection are commonly used to recognize a particular object in an image like presence of a cancer cell, detecting dogs, or cats thereby recognizing only one class. This paper deals with recognizing common objects that we see in our daily surroundings and hence thereby is a multiclass object detection problem.

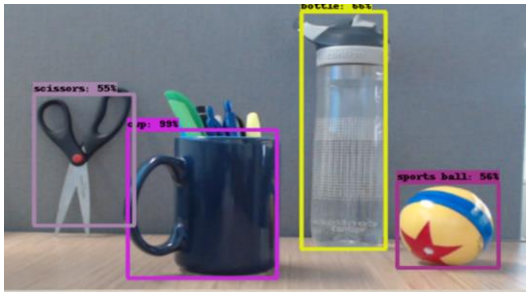

Fig.1 Multi Class Object Detection. Recognizing Scissors with 55% accuracy, cup with 99% accuracy, bottle with 66% accuracy and sports ball with 56% accuracy. Each having a different color bounding box showing different object presence in the image

Taking one step ahead in the world of artificial intelligence: Normally, the images fed to the model are stored on the computer but the method proposed in this paper wants the user to click a picture from their phone or web camera and then feed that to the model in real time and get the results, thus making an interactive object detection tool.

Detecting objects in an image or in a video in real time can be very useful in different domains, for instance in traffic cameras that capture car number plates to get hold of people who break the rules. It can be useful to government authorities, that keep tabs on terrorists by detecting their faces in video surveillance and many more scenarios.

3. *Related Work*

The has been a lot many state-of-the-art works done in this field with different approaches and task at hand. This paper particularly emphasizes on detecting the object and then masking the pixels of that object.

Most studies on transfer learning for object recognition have focused on multiclass recognition without a background class (saying if a crop image contains an object out of M possible classes [4, 5, 6, 9]).

The object localization task requires efficient solutions that can be part of a window scanning. In addition, and more fundamental, the detection task needs to focus on the problem of distinguishing the objects from the background class which requires strong discriminative models in order to get good recognition performance.

Few multiclass object detection systems have been proposed showing improved performance with respect to independently trained algorithms (e.g., [7, 8, 10]).

A different approach proposed is by framing object detection as regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network is used to predict bounding boxes and class probabilities directly from full images in one evaluation [1].

A large step towards masking the objects which means to identify the pixels that belong to the object inside the bounding box, and very relevant to this paper has been prevailing, by using deep convolutional neural network [2].

One of the important tasks of detecting objects in real time is that the response time should be very small. Methods have been established to improve the efficiency of detection task i.e. to make the detection of objects faster as compared to other existing methods by using Region Proposal Network that shares full-image convolutional features with detection network thus enabling nearly cost-free region proposals [3].

4. *Data*

Since this object detection tool is designed for common objects that are in our surroundings the data used for this pupose is the COCO (Common Objects in Context) dataset. It is a large-scale object detection, segmentation, and captioning dataset. It has over 200k labeled images with 80 object classes. Few of the instances are:

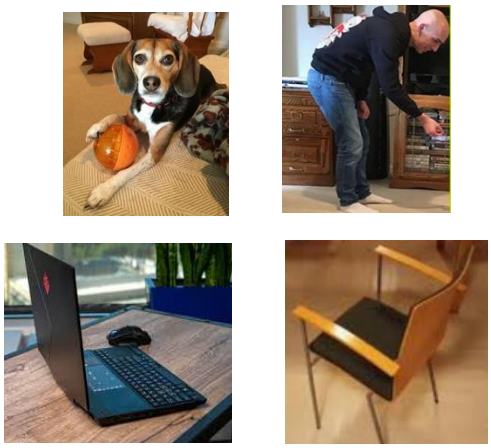

Fig.2 Class Instances (in clockwise direction): Dog, Person, Laptop, Chair

The model is trained on these 80 classes and then the a live image is fed to the model to detect objects in it with the masking and formation of bounding boxes. This model does not require any particular predefined size og images and the image size will solely depend upon the resiltuion of the webcam camera used.

5. *Methods*

Object detection is a process involving certain small processes. It follows:

- *Image classification* which involves predicting the class of one object in an image.
- *Object localization* which refers to identifying the location of one or more objects in an image and drawing a bounding box around their extent.
- *Instance segmentation* which refers to computing a pixel-wise mask for every object in the image.

- *Object detection* combines all these tasks and localizes, masks and classifies one or more objects in an image.

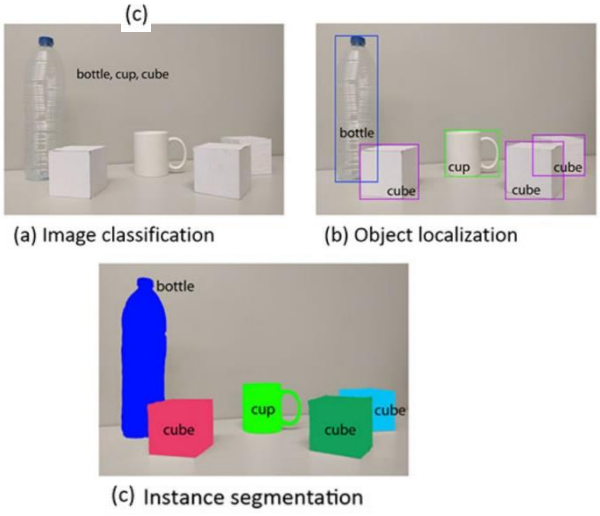

Fig3. Process of object detection

This paper uses transfer learning from Mask R-CNN (Region Based CNN) to detect and segment objects in an image to enhance the accuracy of this interactive tool. It has 2 modules:

1. Generating proposals about regions of interest (ROI) using deep fully convolutional network
2. Predicting class of the object, refining the bounding box and generating a mask in pixel level of the object based on proposed ROI through Faster RCNN

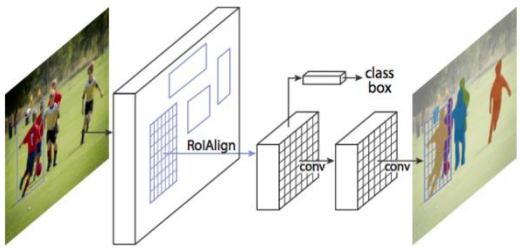

Fig.4 A snapshot of the Mask RCNN

6. Results

The images fed into the model were clicked in real-time through webcam of a laptop. The experimented images included people, bottle, bag, remote, cell phone, cars and many more objects. A few result points to be noted are:

- The general overall accuracy of classifying and bounding objects in boxes was pretty good
- Objects that were overlapping were masked appropriately, recognizing background and foreground properly and keeping parts of an object intact (in same color)
- The overall time to respond with results after capturing was on an average 19 seconds
- Sometimes the model used to get confused between visually similar objects like remote and cell phone, or detected the gap between an open shirt as a tie

A few result instances:

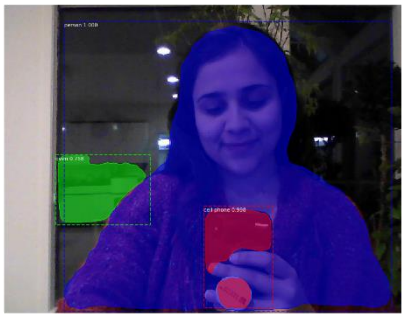

Fig.5 The Real-Time Object Detector identifying (a) Person with 100% accuracy, (b) Oven with 76.8% accuracy and (c) Cell phone with 99.8% accuracy

Fig.6 The object detector detects cars, truck, bicycle and person and misclassifies a car as truck and incorrectly detects a tie

7. Discussion Of Results

As mentioned earlier, this object detector in an interactive tool keeping common man as a user to detect objects in their routine surroundings.

Not much of work is done in detecting multi objects in a single image. This tool successfully detects multiple objects in an image with the model running in the background when the image is captured through the webcam i.e. in real time.

A lot of object detection exists restricted only to indentifying the classes and including the objects in bounding boxes. The model used in this paper adds onto these works by including instance segementation by masking the pixels of every object. Even if same object is present multiple times in the image, each of those are masked in a different color showing that they are all separate objects and are detected with their respective probabilities.

8. Conclusions and Future Work

This paper successfully constructs a model to detect objects in a real time envisronment through the use of tranfer learning from Mask R-CNN. Real time evironment is created using OpenCV tools that allow to take the pictures from webcam of a device.

This model works on any resolution of images and gives out fast responses with the masked objects and bounded boxes.

In future, a more targeted object detection could be prepared for specific environments like for medical domains to identify what could be possible abnormalities or cysts in a scan or for crime domains to identify what kind of weapons a person is carrying hidden (this would require additional screening to detect images under cover through maybe temperature maps) or astronomical purposes.

The real time can also be modified by rather than using images, object detection can be performed on live video streaming which can be useful in catching law offenders. Another area where it can be used is a night time object detector since it would be difficult to see in the dark atmosphere.

9. References

[1] Joseph Redmon , Santosh Divvala, Ross Girshick , Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. University of Washington, Allen Institute for AI, Facebook AI Research, 2015.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask R-CNN. Cornell University, 2018.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Cornell University, 2016.

[4] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In CVPR, 2008.

[5] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In CVPR, 2008.

[6] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In CVPR, 2010.

[7] S. Krempp, D. Geman, and Y. Amit. Sequential learning of reusable parts for object detection. Technical report, CS Johns Hopkins, 2002.

[8] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In CVPR (1), 2006.

[9] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In ECCV, 2010.

[10] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In CVPR, 2004.

[11] Matterport Inc. https://github.com/matterport/Mask_RCNN