Real Time Object Detection Using Convolutional Neural Networks

Sanju Raj Prasad, Swapnil Rathod, Jaideep Rathwa

Abstract In recent years, there has been a notable increase in the field of computer vision research. Object detection is a task in computer vision that involves identifying and localising objects in order to detect or classify them. Human–computer interaction, video surveillance, satellite images, transportation systems, and activity identification are some of the most common object detection applications.For visual imagery, a convolutional neural network (CNN) made up of a series of neural network layers is utilised in the larger family of deep learning architectures. Deep CNN architectures produce excellent results when it comes to object detection in digital images. This paper is a thorough examination of recent advances in object detection using convolutional neural networks. It describes the many types of object detection models, the benchmark datasets that are available, and the research that has been done about using object detection models in various applications.

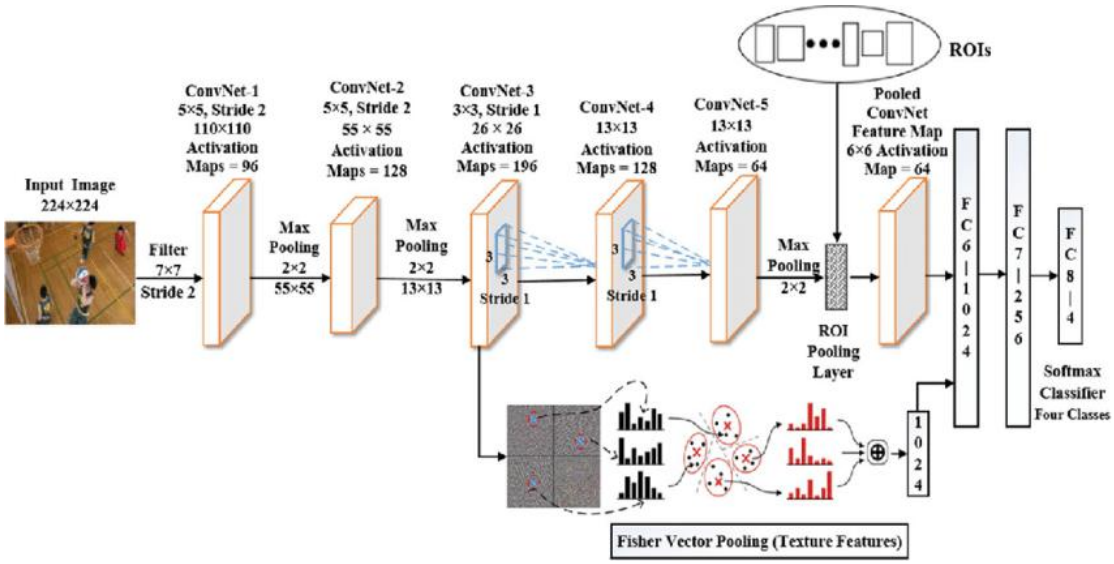
Introduction

Algorithms are a reliable way to handle computer vision problems. Object detection in computer vision is the process of finding instances of objects from a specific class in a digital picture or video. It is a task which involves classifying and locating items in order to detect them. It scales one or more items and determines where the object is presented in the image.
Face recognition system, emotion detection systems, video surveillance, vehicle tracking, and autonomous vehicle driving are all becoming increasingly prominent, therefore fast and precise object detection systems are in high demand. The term "object detection" refers to the process of locating and classifying items within a digital image. CNN-based object detectors are used in a number of applications as a result of the progressive results from deep CNN architectures. It has been classified as a single-stage or two-stage object detection model based on the methodology. R-CNN, Fast R-CNN, Masked R-CNN, SSD, and YOLO are among the CNN-based models covered in this work. Apart from that, it explains the various characteristics of the available datasets. It also goes into the specifics of previous studies that used object detection models in many fields of application.

1. Real Time Object Detection Model Types

The object detection model is extremely dependent on the application it is utilised in, and it must be used effectively. Techniques like transfer learning, on the other hand, make the process of using these models for custom objects / images a lot easier. Object detection frameworks can be grouped into two types: “region proposal” networks and “unified networks”. Multi-stage or two-stage models are those that

are built on region proposal networks. Models that are unified are called single models.The multi-stage models perform the detection task in two stages:
(i) The region of interest (ROI) is generated in first stage and classification is performed on these ROI.
Two-stage object detectors are accurate but somewhat slower.
(ii) Single-stage detectors are faster than two-stage detectors as less computational work is needed.



2. Two-Stage Detectors

2.1 Region Based CNN

R-CNN implements a selective research approach that generates 2000 region proposals from the image. These regions are sent into CNN, which generates a 4096-dimensional feature vector. To classify the entities, SVM is applied to the produced vector. Furthermore, the bounding boxes are also drawn which surrounds to an object. A big problem with “R-CNN” is its speed, as it works very slow around 400ms - 900ms. Furthermore, the algorithm employed to create the proposals, selective search, is very fixed, limiting the risk of learning. The primary disadvantage are that the selective search method suggests 2000 regions per image, generates a CNN feature vector for each region of image, and there is no shared execution between these phases. “R-CNN” achieved a “mAP” value of 53.3 percent, which is a significant improvement above preceding PASCAL VOC 2012 work.

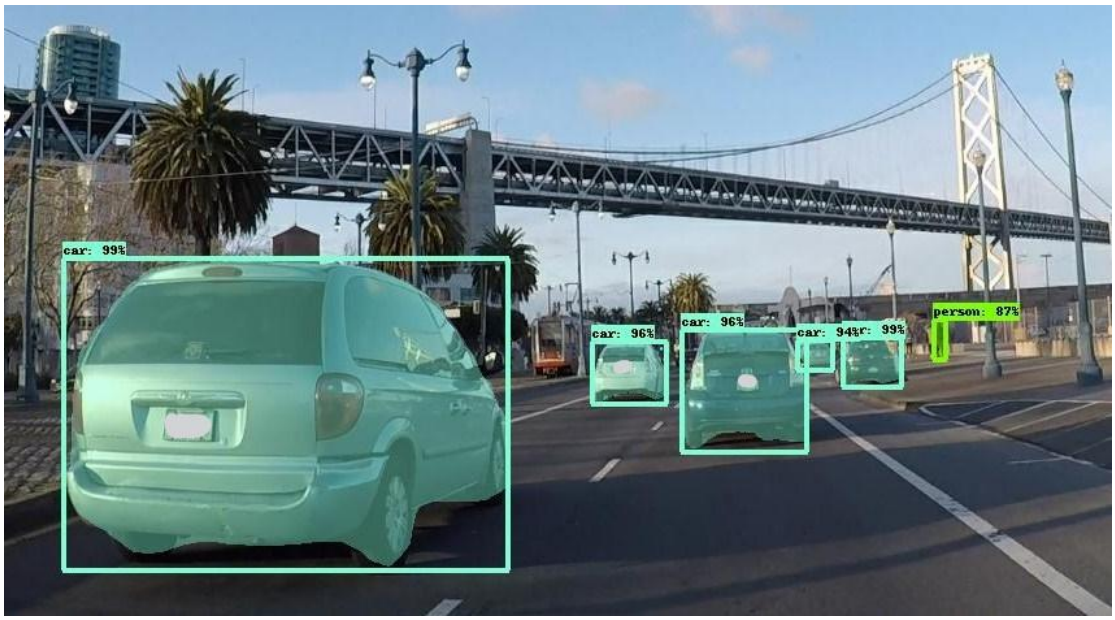
3.2 Fast Region Based CNN

S-Convolution is performed to each region proposal by CNN, and the detection procedure takes longer. The principles of region of interest (ROI) are used in Fast R-CNN to lower the time consumption. A fixed-sized feature map with fixed height and width is retrieved using the “ROI pooling layer”. It uses the max pooling operation to exchange features from the regions. Before applying a max pooling operation, the **h x w** layered window is divided into set of small and fixed sub-windows, i.e., **H X W**. The size of each generated sub-window in the grid is **h * w**. Experimental results

showed that Fast R-CNN obtained a mAP score of 66.9% whereas; R-CNN obtained 66.0%. The experiment was conducted on PASCAL VOC 2007 dataset .

3.3 Masked Region Based CNN

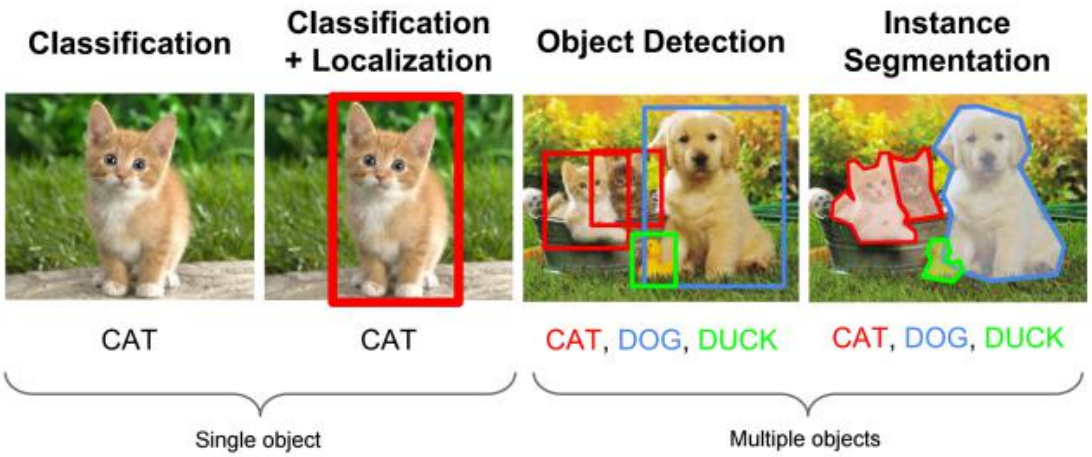
Mask R-CNN, introduced by He et al. [15], enhances Faster R-CNN by emphasizing on instance segmentation from an image. It is a Faster RCNN extension that creates the object mask in addition to the class name and bounding box. In order to complete an instance segmentation task, reliable detection is required.Therefore, it combines the two important aspects of computer vision task, object detection, which classifies and localizes the objects from an image and semantic segmentation, which classifies and assigns each pixel into a fixed set of category. Mask R-CNN introduced RoIAlign layer that maps the regions more precisely by fixing the location misalignment.



3. Single Stage Object Detectors

4.1 YOLO (You Only Look Once)

Yolo is a single-stage object detection paradigm that is much faster than masked region object recognition, although accuracy and precision suffer as a result.
In a single evaluation, it generates the border boxes and class predictions. It's also known as a unified network, and it's much faster than the Faster R-CNN because it only utilizes one convolutional neural network. The CNN in YOLO was originally based on the GoogLeNet model, and the revised version is known as DarkNet based on VGG. It divides the input image into a grid of cells, where each cell explicitly classifies the item and predicts a bounding box, as seen in the diagrams below. As a result, a vast number of boundary boxes are formed and merged into a final product. There are several variants of YOLO present “YOLOv1”, “YOLOv2”, “YOLOv3”.



4.2 Single Shot Detector (SSD)

SSD, like YOLO, is a single stage detection technique, which means it just needs one shot to distinguish many objects in a picture. Both object localization and classification were done in one pass. SSD's fundamental model for retrieving relevant picture characteristics is the '**VGG-16 model**', which had been pre-trained on the ImageNet dataset. It has additional convolutional layers for object detection at the conclusion of the base model. A score is generated for each object category provided in a picture while predicting using the default box. It also performs alterations to the box to better item shape matching. Predictions made from several feature maps with differing resolutions are likewise pooled in the SSD network. This technique aids in the manipulation of things of multiple lengths.

4. Applications of Object Detection Model

Object detection model can be applied in various technical and no technical field such as agriculture, scientific research facilities, manufacturing areas etc.

Below table summarizes the work done using different object detection model, some for pretrained models and some for transfer learned models using various datasets of hand signs.

Object Detection Model	Dataset Size and Type	Image Types
YOLO v2 real time object Detection model (for customized objects)	3500+ hand signs images of size 512x512.	Hand symbols and gestures.

MobileNetSSD for real time object detection (for customized objects)	3200+ hand signs images of size 512x512	Hand symbols and gestures.

5. Conclusion

Face recognition system, emotion detection systems, video surveillance, vehicle tracking, and autonomous vehicle driving are all becoming increasingly prominent, therefore fast and precise object detection systems are in high demand. The term "object detection" refers to the process of locating and classifying items within a digital image. CNN-based object detectors are used in a number of applications as a result of the progressive results from deep CNN architectures. It has been classified as a single-stage or two-stage object detection model based on the methodology. R-CNN, Fast R-CNN, Masked R-CNN, SSD, and YOLO are among the CNN-based models covered in this work. Apart from that, it explains the various characteristics of the available datasets. It also goes into the specifics of previous studies that used object detection models in many fields of application.

6. References

1. Wikipedia Object detection (https://en.wikipedia.org/wiki/Object_detection)
2. Wikipedia Region based detections (https://en.wikipedia.org/wiki/Region_Based_Convolutional_Neural_Networks)
3. David Cocard MobileNetSSD (https://medium.com/axinc-ai/mobilenetssd-a-machine-learning-model-for-fast-object-detection-37352ce6da7d).
4. Object Detection through Modified YOLO Neural Network (https://www.researchgate.net/publication/342419696_Object_Detection_through_Modified_YOLO_Neural_Network).
5. Datasets (https://www.kaggle.com/grassknoted/asl-alphabet)