Real Time Object Detection using Deep Learning: A Webcam Based Approach

Sandipan Chowdhury¹, Prabuddha Sinha²

¹²Department of Information Technology, Government College of Engineering and Ceramic Technology, West Bengal

E mail: sandipanmyself@gmail.com

Abstract:

Object Detection has come to the forefront as one of the most important applications of deep learning which is characterized by its learning ability of features and depiction of features compared to the traditional object detection methods. Region proposal algorithms form the basis of which Object Detection networks hypothesize object regions. Algorithms that have cut down on the running time of these detection networks are SPPnet and Fast R-CNN. In this paper we propose a method for object detection through webcam. The method used is a combination of Region Proposal Network (RPN) and Fast R-CNN, where high quality region proposals are generated by training the RPN end to end, which are in turn is used by Fast-RCNN for detection. These two modules combine to generate an object detection system called Faster R-CNN. Deep models can automatically act as a classifier and detection device, they do not require hand engineered technologies. Therefore, one of the greatest prospects of deep learning technology is in Object Detection. One of the major focus in computer vision technologies is object detection, which has been applied in the pedestrian detection driverless car, video surveillance robotics and counting techniques. The Faster R-CNN models: InceptionV2 and ResNet50 have been used for object detection using webcam, i.e. real time object detection.

Key Words: Object Detection, Faster R-CNN, Deep Learning, Real Time, Webcam.

1 Introduction

Object detection is one of the most important research direction and focus of computer vision technologies (Erhan et al.,2014) which have real time and important applications in driverless cars, video surveillance, robotics, pedestrian detection (Borji et al.,2015),(Tian et al.,2015), counting and monitoring of areas using Remote Sensing also being a new area for applications. Deep learning technologies has been a revolution in object detection methods, it has changed the traditional modes of object identification and object detection. Recently all advancements in object detection based on deep learning techniques is due to region proposal methods (Uijlings et al.,2013) and region-based convolutional neural networks (R-CNNs) (Girshick et al.,2014). R-CNN’s are computationally expensive as originally developed (Girshick et al., 2014), but their cost has been significantly reduced by sharing convolutions across proposals (He et al., 2014), (Girshick, 2015). Fast R-CNN use very deep networks to achieve near real time rates (Simonyan and Zisserman,2015), ignoring the time spent on proposals for regions. The deep neural network has the strong feature of representation capacity (Ouyang et al.,2015) in image processing and it is very useful as the feature
extraction module in object detection. It has been noticed that Fast R-CNNs use GPUs while regional proposal methods are applied to the CPU.

The goal of deep learning is to simulate functions of brain neural activity to perceive particular images of the data. The simulated neural network combines low-level features to obtain high-level representation to determine the distributed characteristics of the data for analysis and learning. (Krizhevsky et al.,2012). In this paper we apply the two models of Faster R-CNN: Inception V2 and Resnet50 which are the deep learning models for the recognition of objects from real time data that is viewed through a webcam. As there are no available datasets we have had to create our own datasets of training and testing images in order to train the network for object recognition through webcam. Images are only used in training and testing as the live streaming through a webcam, i.e. video which is just a collection of images.

2 Related Work

Object detection is an application that is used to detect the objects that are present from the specific scene by a certain method. Prior to the emergence of deep learning technologies the methods of object detection were based on mathematical models (Tang et al.,2017). The common classical methods in object detection are as follows: Hough transform (Merlin and Farber,1975) method, frame-difference (Singla,2014) method, background subtraction method (Lee,2005), optical flow method (Horn and Schunck,1981), (Barron et al.,1992), sliding window model method (Viola and Jones,2001) and deformable part model (Felzenszwalb et al.,2010), (Felzenszwalb et al.,2008) method.

There were various traditional machine learning algorithms with the focus on image analysis tasks such as hand-crafted features extraction, colour segmentation, normalization and these methods were further supported by classification algorithms such as regression and support vector machines. Due to these methods being incapable of processing high dimensional sets of image features, new methods like neural networks came to the forefront which provided a feasible solution for extraction process in automated high-dimensional image sets. Currently, neural networks are used everywhere and there are many experiments utilizing multi-layered networks (Bezak,2016).

There are two categories in classic methods and machine learning applications, the first category, i.e. background subtraction, frame-difference, Hough transform, and optical flow method uses certain data characteristics to create a mathematical model and it is implemented in object detection scenes to obtain the results; whereas in the second category, deformable part model and the sliding window model method integrates classifier algorithms with the supervised features to obtain the object detection result. (Tang et al.,2017).

In deep learning the region selection is done according to some methods, the feature extraction can be obtained by the CNN and the classification can be accomplished by SVM (Tang et al.,2017). In the recent past, in deep learning the development of a model based on region proposal has been brought to the forefront.

The deep learning object detection models based on region proposals is divided into two main parts: first the extraction of regional candidates; the other is a building of a deep neural network.

It has been implemented in the following ways:
1) **R-CNN**

R-CNN (Girshick et al., 2014) is the regional proposal based on Convolution Neural Network which was brought up in 2014 by Girshick who brought to the limelight the concept of region proposal for the first time. The principle that dictates the implementation of R-CNN is the region segmentation method of selective search (Uijlings et al., 2013) for extracting the region proposals present in the image, which consists of the probable objects and loads them into a CNN in order to extract the feature vectors. On completion of this step, the SVM classifier is used to classify the feature vectors for obtaining the classification results in each region proposals. Merging by non-maximal suppression (NMS) yields the outputs of the model with accurate object classification and object bounding boxes in order to achieve object detection. The architecture is as given in Fig. 1.

![Figure 1 R-CNN Architecture](image)

2) **SPP-net**

SPP-net (He et al., 2014) is a deep ANN implemented using spatial pyramid pooling, which rids the method of the warping operation on the input image that was implemented in R-CNN. This allows the input of various sizes for connecting to the fully connected layer of the feature vector of the same dimension after it has passed through the convolution layer. SPP-net eliminates the problem of image incompleteness and object deformation that could be a possibility in R-CNN, but it has a huge drawback because of its poor real-time computation methodology having similarity to R-CNN.

3) **Fast R-CNN**

Fast R-CNN (Girshick, 2015) is the upgraded version of R-CNN which is capable of solving the repeated problem of calculation of the 2000 regional proposals that pass through the CNN. In comparison to the R-CNN the improvement obtained using Fast R-CNN is such that it marks out the region proposals by extricating using selective search algorithm from input images to the feature layer of the CNN and pooling is conducted on the mapped region proposal which forms the ROI pooling feature layer. Fast R-CNN is helped by the ROI pooling in order to extract the stable sized feature vectors, which is
important to connect with the CNN successfully. The performance of ROI pooling is like SPP-net's spatial pyramid pooling. The way that Fast R-CNN operates is displayed in Fig. 2

Mapping of region proposals from the images input to the Fast R-CNN feature layer distributes the computation of the convolution, which in turn brings down the calculation. Furthermore, in order to minimize parameters related to full connection, a truncated SVD is implemented by Fast R-CNN to ensure that two small fully connected layers replace a single fully connected layer, which in turn reduces the calculation of the network. It has been tabulated that during training the converging rate of Fast R-CNN is 8.8 times that of R-CNN and 2.58 times that of SPPnet, while during testing it is 146 times faster than R-CNN without truncated SVD and 213 times faster than R-CNN with truncated SVD and for SPPnet it is 7 times faster without truncated SVD and 10 times faster with truncated SVD (Tang et al., 2017).

3 Architecture

In this study the object detection that has been carried out through webcam has been implemented using Faster R-CNN (Ren et al., 2015), which is the upgraded version of the Fast R-CNN. In order to find a solution to the huge computational; costs and poor real time applications incurred by the selective search method of Fast R-CNN and R-CNN, region proposal networks (RPN) has been implemented by the Faster R-CNN to overcome the obstacle. An end to end framework is implemented by the Faster R-CNN which trains the model in a simpler and faster way than R-CNN and Fast R-CNN.

The Faster R-CNN consists of two modules: firstly, a fully CNN that is used for proposing regions, and the second is the Fast R-CNN (Girshick, 2015) which is a detector that uses the regions proposed. The whole system consisting of both the modules is a single, unified network for object detection. Fig. 3 shows the architecture of the Faster R-CNN.
The main usage of RPN is that it uses the method of attention mechanism (Chorowski et al., 2015) in neural networks in order to tell the Fast R-CNN module where to look. The way that Faster R-CNN operates is displayed in Fig 3.

On implementation of RPN in Faster R-CNN the region proposals are reduced from 2000 in selective search to 300 only, which drastically reduces the computational time of the model and speeds up the whole network. Experiments indicates that the speed of Faster R-CNN is 5fp/s, i.e. 10 times faster than that of Fast R-CNN and its accuracy is also improved.

4 Result

The experiments were conducted using NVIDIA platform and implemented TensorFlow API on Windows 7. The GPU NVIDIA GeForce GTX 660 with 2GB of RAM was used in accelerating the learning process. We used CPU NVIDIA Quadro FX 4800 with 8GB RAM. There are many datasets available like MNIST, MSCOCO, CIFAR and PASCAL VOC using which we could dry run our models before implementing them on our own datasets. Also there are several standard models publicly available under the CC licence that we were able to use like LeNet, AlexNet, GoogleNet, ResNet, Inception, MobileNet and NAS. These models have a lot of differences in the architecture and storage requirements. 2GB of RAM is available to us and is present on the GPU. It is very complex to find the best fit models to fully utilize the memory of the GPU. Through trial and error we have identified two models which provide optimal performance on the available memory. These two models are: Inception V2 and ResNet50.

A. Inception V2 model

The Inception V2 model is the first model that we have trained. This model has been based on 3000 training images and 600 validation images (20%) with image dimension of 500x500 pixels. The image was coloured image in RGB of the jpg encoding format which is not a lossless format and does not hold much details.

The outputs obtained for image detection through webcam has been as given below in Fig.4, 5, 6, 7, 8. Validation with a moderate accuracy of 80% has been achieved with the increment in the number of learning epochs. There has been a loss of 10% and 5% during training and validation respectively. The model obtained a prediction accuracy of 88.6% and 12.4% was
the rate of unsuccessful prediction. Deep learning models that successfully win Kaggle competitions nowadays have over 90% accuracy in implementations of object detection from simple RGB imagery.

So, therefore the need arises to modify, change and implement our neural network model in a better way such that we can achieve the required prediction rate. After increasing the number of training images, changing the model, parameter tuning, and creating and testing several different models we get successful result from ResNet50.

B. *ResNet50 model*

The final model we have trained is the ResNet50 model. This model has been based on 5000 training images and 1000 validation images (20%) with image dimension of 500×500 pixels. The image was coloured image in RGB of the jpg encoding format which is not a lossless format and does not hold much details.

The outputs obtained for image detection through webcam has been as given below in Fig. 9, 10, 11,12,13. Validation with a high accuracy of 95% has been achieved with the increment in the number of learning epochs. There has been a loss of 5% and 3% during training and validation respectively. The model obtained a prediction accuracy of 95% and 5% was the rate of unsuccessful prediction.
5 Discussion

Here we present a deep learning model for the automated object recognition and object detection in webcam streaming video which gives us object detection in real time. The latest deep learning models form the key aspects of this system; i.e. Faster R-CNN and detailed design of processing methods for the image classification. In the past, object detection in photographic images and video has mainly been restricted to low-level image analysis and classical machine learning methods, which utilize only a small set of features of an image. Deep Learning based approaches focus on automatic extraction of image features and these approaches show the best performance. Furthermore, in recent years it has been noted that the computational power and the accuracy has undergone a huge improvement due to Faster R-CNN which implements the RPN. Image blur and low/high light are mainly responsible for the errors in detection made and also due to limitation of the training sets can further contribute in errors. It is the area of possible improvements. The fast convergence of deep learning models is a must for real time object detection which is available in Faster R-CNN.

6 Conclusion

On comparison of the two different models it has been deduced that the ResNet50 model has better performance than the Inception V2 model. The proposed model is capable of object detection in the live streaming videos through webcam, which is Real Time object detection. The model is ascendable and capable to learn from a lot of training datasets. It is required to arrange a significant amount of training datasets of images from various angles of various different objects that are able to be thought of as the bottom line of comfortable verification method for testing and validation of the projected model. The advantage of the implemented ResNet50 is the possibility to implement it in real time and further implementations across all platforms for future applications as an example in movable sensible devices. Earlier. The drawback of less obtainable memory (2GB on GPU) represents the likelihood to use the model by a memory restricted platform, i.e. small smart devices and drones, etc.

References


