Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 6, May-2019, ISSN 2230-7540

A Survey of Object Detection Methodologies Based on Deep Learning

Parveen Saini[1]* Karambir Bidhan[2] Sona Malhotra[3]

[1] Depertament of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

[2] Depertament of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

[3] Depertament of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Abstract – In past couple of years, it has been observed that digital data is accumulating with astronomical speed. Every day people have been clicking and sharing numerous images with each other. Moreover, video surveillance, vision based security system keep on gathering the raw footage every second. To extract out the desired information from these raw data automatically and precisely in real time, various method has been implemented.

This paper present the comparative study of various methodologies based on deep learning to extract out the desired information from images specifically detection of object from image. Object detection has broad spectrum of application in various artificial intelligence based industry. So it plays a vital role in current technical era.

Keywords—Object Detection methodologies, Computer vision, Deep Learning, Convolutional Neural Network

----------X----------

1. INTRODUCTION

Throughout the history of computer vision, object detection has always been focus of research work. But in past couple of years deep learning has provided considerable improvement in object detection [1], [2].

Krizhevsky et al. in 2012 comes up with "AlexNet" based on deep convolutional network (DCNN), which has achieved exhilarating performance in classification of images in the Large Scale Visual Recognition Challenge (ILSRVC) [3], [4]. Since then all research focus shifts to deep learning based methods. While there are many neural network based methodologies to detect the object, deep convolutional network neural emerged as record breaker. So this paper major focus remains to analyse the deep neural network based object detection methods [1], [5], [6].

The aim of object detection is majorly to discover the instance of desired object presence or absence from provided categories in given image. If it is present it returns the outcome as the spatial location of the instance [4], [7].

Object detection has two main tasks. First detection of certain instance and second is to determine the category of object. The first one focus is to determine occurrence of object such as face of a particular person, cat, a famous building. Whereas second task aim is to determine the different instances of the already defined object classes (categories). For example, car, human, bicycle.

2. BACKGROUND

A. Objective

The major objective of object detection processes to design a object detection algorithm which can achieve high accuracy (quality) and high efficiency. High quality means detection has to precisely locate and recognise the object from different object categories in the real world pictures. Which is also capable of distinguish object instances from the same category and from inter-class

appearances variations. High efficiency means to perform the detection with available computation machines and memory at sufficiently high frame rate. In spite of numerous years of development, research and narratively excellent progress talks the unified target of efficiency and accuracy goals have not been reached if we compare it with human object detection capabilities.

B. Major Challenges

There are numerous challenges to object detection methodologies to achieve the accuracy and efficiency. Major challenges are as below.

For accuracy, challenge is object variation within class and there are numerous number of object categories. Object variation within class has two forms. First object intrinsic factors like object colour, shape, texture and size. Others are imaging condition like pose, light, angle, view, camera, background, weather condition etc. In addition to these digital artefacts like, noise, filter pass, resolution variation etc.

For a note there are 104-105 number of object categories in real world which human vision system can detect easily. But with present advancement we are able to detect with 200 object class only like ILSVRC which has 200 object classes [4]. It is crystal clear those current state models are far behind even in term of object classes as compare to human being system.

Furthermore, there is phenomenal growth in number of images every day, which demands for efficient and scalable object detection system. Social media has engulfs the people digital world with handheld devices and people are seeking to visualize there data there. However these hand held devices has scarce memory and very limited computation abilities, in which scenario efficient detection is imperative.

C. Past Related Work

Before 1990, major methodology to detect the object was based on geometrical representation [8], [9]. Techniques were revolves around template matching and part based models [10].

Later research focus shifts to statistics classifier like SVM and neural networks which were based on appearance features [11], [12]. Then this appearance features based detection in late 1990 shifts from global representation to local representation [13]–[15]. This local representation was stable to various dynamic factors like view angle, rotation, illumination etc. This local invariant features capture the research domain with visual recognition models like SIFT [16], Histogram of Gradients (HOG) [17] etc.

Until 2012, these above models held the market. Then in 2012 Deep Convolutional Neural Networks (DCNN) [3] appeared with breakthrough performance in image classification. This success of DNN in image classification leads to the Region based Convolutional Neural Networks (RCNN) detector which was proposed by Girshick et al. [18].Since then, many deep learning based methodologies has been come into the picture with sincere thanks to the high end computation machines like GPU and large scale datasets such as ImageNet [4], [19] PASCAL VOC [20], MS COCO [21] and ILSVRC [4].

3. DEEP LEARNING BASED OBJECT DETECTION METHODOLOGIES

Before This part will elaborate the object detection methodologies which are based on deep learning. Almost all the approaches which were proposed in past few years are based on DCNN. However researchers are attempting to improve one or other aspect from this system.

In general, we can divide the detection approach based on deep learning in two categories.

1. Two Stage Detection approach

In this approach at first stage region proposals are represented and then the detection part is performed on image.

2. One Stage Detection approach

In one stage detection method there is no region proposals step all detection steps performed in single go.

A. Two Stage Detection approach (Region Proposals Based)

In two stage detection approach, at first level region proposals are developed from the image which are independent of category, then with CNN features are pull out from the regions. Then the category labels are determined from the region proposals with category specific classifiers. Initially based on this approach Detector Net [22], Over Feat [23], Multi Box [24] and RCNN [18] were proposed. Then various parts have been tweaked to enhance the performance. One by one this paper will review the all major methodologies based on two stage detection approach in this section.

1) Region Proposal + CNN (RCNN)

Girshick et al. used the “AlexNet” [3] and “selective search” to develop the RCNN. He integrate the AlexNet with selective search (more precisely used the region proposal method). RCNN achieved the

high quality in object detection. It has multistage pipelines.

- Class- Sceptical region proposal
- First with selective search [6], candidate regions are obtained which might have the objects.
- Then region proposal are cropped and warped into identical size and input to the CNN to fine tune it. CNN model here trained in advance with the large scale dataset like ImageNet.
- By using constant length features, number of linear SVM classifiers is trained.
- In last with CNN features, Bounding box regression is observed for each class object.

Though RCNN has reached high quality in object detection, it has limitation also.

Here is RCNN each and every stage need to train separately, which makes the training slow and complex. Further to it, all regional proposals give only rough localization. This is needed to be externally detected.

Bounding box regression and SVM classifier training is costly both in term of memory space and time. Because for each region proposal for every image, CNN features is pulled out independently. That further poses great challenges to large scale detection, particularly with deep CNN networks like AlexNet [3] and VGG [25]. Testing in RCNN is too slow as for every test image, CNN features are pulled out per object.

2) Fast R-CNN

Fast RCNN addressed the few disadvantages of RCNN along with improvement in detection quality and speed. It was proposed by Girshick [26].

In fast RCNN training is developed in stream line manner. In training process it congruently learn a class-specific bounding box regression and softmax classifier using a multitask loss instead of training SVMs, BBRs and softmax classifier in three different phases as in RCNN.

Fast RCNN shares the computation of convolution layers across region proposals. To pull out the constant length feature for every region it puts a Region of Interest (RoI) pooling layer between last convolutional layer and the first fully connected (FC) layer. After the features are passed through series of FC layers it eventually divides into two output layers i.e. softmax and class-specific bounding box regression. Where objects category prediction is performed by softmax probabilities and region proposal for further refinement is performed with class-specific bounding box regression offsets.

In summary, Fast-RCNN performs the training 3 times faster and testing almost 10 times faster than RCNN. Further it's concept of single-stage training process which updates all layers simultaneously lead to higher detection quality and without need of extra memory to cache the features.

3) Faster R-CNN

Though fast RCNN considerably improves the speed of detection, yet it depends on external resources for region proposals.

In recent research work, it is revealed that CNNs have capability to localize object in convolution layers [27]. A ability which is absent in FC layers. Thus an internal tool has been found to replace the external localization tool selective search to produce the region proposals. In addition to it, RPN is basically a Fully Convolutional Network (FCN) [28], [29]

Ren et al. [30] proposed the Faster R-CNN framework with an effective and accurate region proposal network (RPN). He utilises the single network to get region proposal by RPN and region classification with fast RCNN. Thus Faster RCNN is a purely CNN based network which does not use hand crafted features.

RPN work flow is as, at every CONV feature map position it initiate L n*n references boxes with variant scale and aspect ratios. These reference boxes are called anchors. Then each anchor is mapped to lower dimensional vector like VGG. Further, these are fed into two parallel FC layers (box regression layer and object category classification layer)

In the end faster RCNN, has efficient computed the region proposal with CNN and achieved high accuracy to detect the object. It was successfully tested with PASCAL VOC 2007 by using 300 region proposals for each image.

Additional information, while Faster RCNN was evolving, concurrently Lenc and Vedaldi [26] proved that CNNs has enough geometric information to detect the object in CONV layers as compare to FC layers. This further provided the concrete proof that an integrated, simpler, faster object detection model is just possible with CNNs without any external region proposals.

4) Region based Fully Convolutional Network (RFCN)

Faster RCNN proved to be efficient than Fast RCNN .But it still needs the region wise sub networks per ROI. (Numerous ROIs per image)

Dai et al. [31] come up with RFCN detector, which has no masked fully connected (FC) layers, was purely based on fully convolutional layers, where almost all computation for entire image is shared. In RFCN RoI sub network was dropped as compare to the faster RCNN.

However, at later Dai et al. observed that his new design becomes inferior in term of accuracy. Later he clubbed the RFCN with ResNet and achieved the comparable accuracy and faster running time as compare to faster RCNN.

B. One Stage Detection approach

As, we have discussed the progress of region based pipeline methodologies in last section. Despite the fact and its achievements, it has very limited approach with mobile devices. Because region based pipeline methodologies are computationally expensive for hand held or wearable devices. Thus researchers focus shifted to develop some unified method, instead of improving the individual modules of region based method.

Here unified methodology basically refer to a framework in which, it predict the object class and bounding box in straight forward manner without region proposal and after math of classification from image. This methodology is quite simple and performs all steps in single stage without need to region proposals.

1) Overfeat

Sermanet et al. [23] proposed the first of kind single stage detector based on CNNs. It won the ILSVRC2013 localization competition and become the most successful method in single stage detector.

It uses the multi scale sliding window to detect the object. It performs the forward single pass through CNN network which is based on convolution layers only (excluding final classification and regression layers).

It develops the grid of features vectors. Each vector helps to predict the object in image as each vector represent the different view location of object in image and it share the computation between the overlapping regions. As soon as object is predicted it uses these features for bounding box regressor.

RCNN has been outperformed by the OverFeat in term of speed, but it is less accurate as compare to it.

2) You Only Look Once (YOLO)

Redmon et al. [32] shown up with YOLO (You Only Look Once), a unified detector. It performs the object detection as regression issue in all sorts of steps like in spatially segregated bounding boxes, image pixels and associated class expectations. It straight forwardly predicts the object detection by using small chunk of candidate regions. It completely drops the region proposal generation stage. Region proposals based methodology such as Faster RCNN, predicts the object detection as per features from confined regions. While YOLO use the whole region. In other words complete image globally.

YOLO works on small grids. It divides the images into small grids of P*P size. These grids are used to predict the class expectations (C), bounding box positions and scores for these boxes. Entire prediction is performed as P*P*(5B+C) tensor. YOLO runs in real time, faster by dropping out the region proposal generation phases completely.

It commits more localization errors. Major reason behind it is that it sees the entire picture and to make forecast it encodes contextual information about the object classes. Moreover it is not able to forecast the false positives which present on image base.

Other possible downside of YOLO is that it may fail to localize some objects. Because grid division performed at broad level, in ever grid by architecture it covers one object.

YOLOv2 and YOLO9000 are further advance version based on basic of YOLO.

3) Single Shot Detector (SSD)

Liu et al. [33] clubbed the ideas from RPN of Faster RCNN, multi scale CONV features and YOLO and offered the Single Shot Detector (SSD). Resultant acquired the faster detection speed along with the high detection quality. SSD attains the real time speed with improved accuracy. It has satisfactory speed as per YOLO and has accuracy equivalent with region-based detectors such as Faster R-CNN.

SSD architecture is as follow. It uses the fully convolutional CNN network. Its starting layers are based on standard framework like VGG [25], which act as base network.

To this base network then it add multiple auxiliary CONV layers, with decreasing sizes as move deep in network. At last information with low resolution is present which is too coarse spatially for precise localization. By the use of multiple scales it operates on multiple CONV features map. Each of that predicts scores for object category and bounding boxes offset as per size. It makes use of thicker layers to detect the small objects.

4. CONCLUSION

It is impractical to compare all available detectors. Furthermore, there is no standard platform to differentiate them in consolidate manner. Reason is that each has their own fundamental framework, innovative technique for feature and training procedure. This paper presented the study of three major families of detector, Faster R-CNN, R-FCN and Single Shot Detector (SSD) successfully with speed and accuracy parameters which varies in image resolution, base network and number of box for region proposals. It has highlighted the achievement of particular method along with downside.

In the end, a word of thanks is required for the deep learning techniques which dramatically change the picture of image detection with absolute performance. At the same time, object detection has to be improving more to satisfy the practical applications requirements. It is far away from human vision system speed and accuracy so there is enough room for improvement in this direction.

5. FUTURE WORK

A. Context Based Object Detection

Deep learning has done very limited to exploit the information present in image context. Real world objects generally associated with surrounding and other objects. These object relation, their associations; global scene could help to detect the object more accurately and specially with poor quality images. So in this direction work is need to be done.

B. Robust Learning Mechanism

Current deep learning methods based on fully supervised learning, which requires fully labeled data. There are 10 4- 10 5 different classes of object in the world .It would be impractical to label them for training. So, there is need to power the CNNs with robust training methods instead of fully supervised learning method.

C. 3D Object Detection

The research domain of object detection with existing state of art mechanisms/ methodologies remains far away from 3D object detection. There is urgent urge to start research in this direction.

REFERENCE

[1] Y. LeCun, Y. Bengio, and G. Hinton (2015). “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444.

[2] G. E. Hinton (2006). “Reducing the Dimensionality of Data with Neural Networks,” Science, vol. 313, no. 5786, pp. 504–507.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton (2017). “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90.

[4] O. Russakovsky et. al. (2014). “ImageNet Large Scale Visual Recognition Challenge,” arXiv:1409.0575 [cs].

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Ha (1998). “Gradient-Based Learning Applied to Document Recognition,” p. 46.

[6] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders (2013). “Selective Search for Object Recognition,” Int J Comput Vis, vol. 104, no. 2, pp. 154–171.

[7] C. Galleguillos and S. Belongie (2010). “Context based object categorization: A critical survey,” Computer Vision and Image Understanding, vol. 114, no. 6, pp. 712–722.

[8] J. L. Mundy (2006). “Object Recognition in the Geometric Era: A Retrospective,” in Toward Category-Level Object Recognition, vol. 4170, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3–28.

[9] R. Girshick (2015). “Fast R-CNN,” arXiv:1504.08083 [cs].

[10] M. A. Fischler and R. A. Elschlager (1973). “The Representation and Matching of Pictorial Structures,” IEEE Trans. Comput., vol. C–22, no. 1, pp. 67–92.

[11] H. A. Rowley, S. Baluja, and T. Kanade: “Neural Network-Based Face Detection,” p. 28.

[12] M. J. Swain and D. H. Ballard (1991). “color indexing.pdf,” International Journal of Computer Vision, vol. 32, pp. 11–32.

[13] H. Murase and S. K. Nayar (1995). “Visual learning and recognition of 3-d objects from appearance,” Int J Comput Vision, vol. 14, no. 1, pp. 5–24.

[14] Y. V. Lata, C. K. B. Tungathurthi, H. R. M. Rao, D. A. Govardhan, and D. L. P. Reddy

(2009). “Facial Recognition using Eigenfaces by PCA,” vol. 1, no. 1, p. 4.

[15] G. E. Hinton (2007). “Learning multiple layers of representation,” Trends in Cognitive Sciences, vol. 11, no. 10, pp. 428–434.

[16] D. G. Lowe (2004). “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110.

[17] N. Dalal and B. Triggs (2005). “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), San Diego, CA, USA, vol. 1, pp. 886–893.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014). “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580–587.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei: “ImageNet: A Large-Scale Hierarchical Image Database,” p. 8.

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2010). “The Pascal Visual Object Classes (VOC) Challenge,” Int J Comput Vis, vol. 88, no. 2, pp. 303–338.

[21] T.-Y. Lin et. al. (2014). “Microsoft COCO: Common Objects in Context,” arXiv:1405.0312 [cs].

[22] C. Szegedy, A. Toshev, and D. Erhan: “Deep Neural Networks for Object Detection,” p. 9.

[23] N. Chavali, H. Agrawal, A. Mahendru, and D. Batra (2016). “Object-Proposal Evaluation Protocol is ‘Gameable,’” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 835–844.

[24] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov (2014). “Scalable Object Detection Using Deep Neural Networks,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 2155–2162.

[25] K. Simonyan and A. Zisserman (2014). “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv:1409.1556 [cs].

[26] K. Lenc and A. Vedaldi (2015). “R-CNN minus R,” arXiv:1506.06981 [cs].

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba (2014). “Object Detectors Emerge in Deep Scene CNNs,” arXiv:1412.6856 [cs].

[28] E. Shelhamer, J. Long, and T. Darrell (2017). “Fully Convolutional Networks for Semantic Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651.

[29] C. M. Bautista, C. A. Dy, M. I. Manalac, R. A. Orbe, and M. Cordel (2017). “Convolutional neural network for vehicle detection in low resolution traffic videos,” in 2016 IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia, 2016, pp. 277–281.

[30] C. Kaensar (2013). “Analysis on the Parameter of Back Propagation Algorithm with Three Weight Adjustment Structure for Hand Written Digit Recognition,” in 2013 10th International Conference on Service Systems and Service Management, Hong Kong, China, pp. 18–22.

[31] J. Dai, Y. Li, K. He, and J. Sun (2016). “R-FCN: Object Detection via Region-based Fully Convolutional Networks,” arXiv:1605.06409 [cs].

[32] S. Ren, K. He, R. Girshick, and J. Sun (2017). “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149.

[33] W. Liu et. al. (2016). “SSD: Single Shot MultiBox Detector,” in Computer Vision – ECCV 2016, vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, pp. 21–37.

Corresponding Author

Parveen Saini*

Depertament of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

pkmantra@gmail.com