Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Comparative Study of Target Image Detection using Deep Learning

Neha Vora[1], Divya Shekhawat[2]

[1]Research Scholar, [2]Assistant Professor, Faculty of Computer Science, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India

E-mail: nehavora1989@gmail.com[1], divya.shekhawat23@gmail.com[2]

Abstract

The object detection and deep learning technology has certainly proven effective in surveillance systems, automated driving, and facial recognition. Today, computer vision has given an entirely new perspective. However, when it comes to targeting a particular object within a complex image or video footage, it may seem to be a major challenge. By the rapid developments in the area of computer vision, the detectors have certainly improved greatly. This study presents a comprehensive literature review of various object detection algorithms, and their challenges, including one-stage and two-stage detectors. Finally, based on the current development of target image detection, the future prospects of research have been stated.

Keywords: Convolutional Neural Network (CNN); computer vision; Target Image Detection

1. Introduction

In the surveillance and traffic monitoring systems, there are multiple objects that need to be identified. Traditionally an individual would monitor the CCTV footages and identify objects of interest; this may lead to a lot of negligence due to the inability of a human to monitor constantly without any distraction. The objects would often miss the human eye. However today the object detections technology has proven highly effective and accurate in identifying the objects of suspicion without much human interference. Adam Coates et al., [1] discovered that GPU-based detectors may be trained on millions of instances and are up to 90 times faster than a well-optimized software version.

DOI: https://doi.org/10.36548/jiip.2023.2.001
Received: 24.03.2023, received in revised form: 25.04.2023, accepted: 10.05.2023, published: 20.05.2023

In computer vision, the major challenges arise in processing time, storage, and identifying a single class of object amongst multiple objects in a single frame. The need to create an efficient object detention module has created a great interest among researchers to build highly accurate and efficient object detectors that are able to detect objects in an image or video. Today the Convolutional Neural Networks and deep learning models can detect objects in real time and in large scales [7]. As the technology continues to advance, the applications of image object detection will continue to expand and become increasingly important in our day-to-day lives.

From a set of known labels, object recognition determines an object's identity in an image. [2]. Whereas object detection recognizes the location and movement of the target along with the identity. Image target detection has always been an important area of research. In a complex case such as a crime scene, an image with multiple objects of different class such as sofa, knife, blood, glass, etc. are present. In such a case, an object detection module needs to be efficient and also be able to detect with efficacy.

This research analyses the deep learning-based target detection procedures in two categories: single-stage target detection and multi-stage target detection. Furthermore, the pros and cons of each category and algorithm are discussed. Finally, the survey is concluded with some promising future research directions.

2. Background of Target Image Detection

General target detection focuses on identifying a wider class such as sofa, faces, cars, etc; however, the algorithms may face challenges when detecting a specific category. The algorithms perform a complex task of isolating a specific object from an image, by drawing a bounding box around it. The general object detection performs four tasks to identify an object- identifying the class and features, object localization- aims to locate the target within the image or video, semantic segmentation- every pixel in an image is associated with a label or category, and object instance segmentation- detects instances of objects and demarcates their boundaries. Object detection algorithms use bounding boxes to detect an object. Due to complex image structure and quality, there may be few difficulties at the pixel level [8].

3. TARGET DETECTION STRUCTURE

The two-stage framework uses a Region Proposal Network (RPN) to detect a series of bounding boxes known as region proposals. Each identified region is then fed to the Convolutional Neural Network (CNN) to build a feature map. This feature map is then used to estimate the class and location of the region proposal. This framework is most commonly used in target detection and serves as most commonly practised mechanism for object detection.

When compared to the 2-stage framework, the 1-stage pipeline provides a simpler and more efficient approach to target detection. Unlike the two-stage framework, the one-stage pipeline predicts the target's class and position using a single convolutional neural network. The one-stage pipeline has more accuracy than the latter. This works best in the real-time applications. However, the one-stage pipeline is gaining popularity due to its efficiency and accuracy.

A. Regions with Convolutional Neural Networks

Regions with Convolutional Neural Networks (RCNN) is widely popular due to its accuracy in detecting smaller objects. It can detect multiple objects in the same image as well. RCNN is based on RPN, and can help in reducing the processing time.

RCNN algorithm first divides the image into 1,000 to 2,000 boxes by using the selective search method. These boxes are then sent to the CNN to extract the desired properties. Regression is then used to process these characteristics in order to modify the position of the bounding boxes associated with each individual feature. The major challenges in RCNN is the space and processing time it takes to detect an object. It is necessary to extract the matching photos for several places beforehand. Given that each test image takes around 0.78 minutes to process, RCNN can’t perform in real time. Moreover, the selective search strategy is a static algorithm. As a outcome, at that point, no learning takes place, which could consequence in the creation of poor candidate region proposals. Conventional CNN requires fixed-size input images, and the crop/warp (normalised) procedure will cause objects to be truncated or stretched, which will result in the loss of inputted information.

B. Fast R-CNN

It is a quicker form of R-CNN because it doesn't require CNN to receive 2000 region suggestions on a consistent basis. Instead, a feature map is produced from the convolution kernel pooling, which is only performed once per image. Fast RCNN [3] can address RCNN's

drawbacks while simultaneously enhancing both its speed and accuracy. Fast RCNN uses multi-task loss, has a one-stage training method [4], has higher Mean Average Precision (MAP), shares parameters across all network layers, and doesn't require disc space to act as a temporary feature cache [10][11].

The candidate region and the complete image make up the Fast R-CNN input. Fast R-CNN conducts multiple convolution kernel pooling for the image once and generates a feature map. The algorithm then identifies a Region of Interest (ROI) within the image. Fixed dimension feature extracting ROI pooling is used for every input ROI region to extract feature vectors. The full connection layer is where the feature vectors are subsequently delivered in order, with the FC branching into 2 output layers at the same level [5]. The first layer's function is to divide the target into K object classes and give the probability distribution for every ROI just to calculate the SoftMax probability. The bounding box position of each of the four real values output by the second layer, representing a distinct class of K objects, which is precisely encoded, is real. In the entire framework, complete training using multitasking losses is used (apart from the RP extraction stage).

Fast RCNN has significantly increased in both speed and accuracy, but it still has a lot of drawbacks because it employs selective search, which takes a long time. The candidate area can be found in roughly 2-3 seconds, whereas the feature classification only needs 0.32 seconds. The need for real-time applications cannot be satisfied by this.

C. Faster RCNN

The faster version of Fast-RCNN is Faster RCNN. There are two modules in it: the Fast RCNN detection module and the RPN target box extraction module for the RGN. RPN is a full convolutional neural network that has a full CNN layer. Based on the extraction of RPN, the target is identified by the faster RCNN [17]. Once the image is given as input, it generates a candidate region through RPN, then the features are extracted and sent to the classifier to classify. Regression and fine-tuning the regressor's position come next as the final phase [24].

D. Mask RCNN

For every target, target image detection is performed and a top-notch segmentation result is produced. It is also simple to adapt for different jobs, like character key point identification.

4. Unified Pipeline (One Stage Pipeline)

A. YOLO

'You Only Look Once' is referred to as YOLO [6]. This method (in real-time) finds and recognises objects in an image. The YOLO procedure of object identification, which is done as a regression problem, provides the class probabilities of the detected images. When an input image is provided, it promptly yields the appropriate bounding box and its classification categories in various areas of an image.

As compared to other object identification methods, the YOLO detection speed is extremely quick, reaching 155 FPS. Unlike other object detection algorithms, YOLO's input is a complete image, allowing it to make effective use of the entire information while detecting objects. It also has a low likelihood of predicting the incorrect object information on the backdrop. YOLO is more mobile and has the ability to learn extremely generalised traits. Unfortunately, object detection's accuracy is subpar, and positioning mistakes are simple to make. A grid cell also performs poorly at recognising little things because it can only forecast two objects at a time.

B. Single Shot Detector

Single Shot Detector (SSD) has a speed that is similar to YOLO's but faster than Faster RCNN, because it uses complete, step-by-step operation mode, which is different from Faster RCNN but similar to it. This comparatively fast speed can be used in a variety of conditions and has good real-time performance. With regard to accuracy, SSD employs many feature layers for detection (multi-scale), allowing it to handle a wide range of problems with various sizes and statuses [25]. As a result, SSD is better as compared to YOLO for detecting small objects. However, the total accuracy is approximately equal or greater than Faster RCNN. Hence, it requires both speed and precision.

5. Evaluation Index

A. Accuracy

It is defined as the proportion of accurate samples to all samples, and it is commonly used to assess the detection model's accuracy. It has limited data, therefore it is not possible to assess the model thoroughly.

B. Confusion Matrix

The expected number of categories is on one axis of the confusion matrix, and the real number of labels is on the other. The sum of accuracy divided by the number of images in the diagonal test set of confounding matrices can also be determined because the diagonals show how many consistent model predictions and data labels there are.

C. Precision, Recall & PR curves

One typical illustration is the test set that contains basketball & volleyball images only, expecting that the classification system's ultimate purpose is to get all volleyball images instead of basketball images. Then it can be specified as:

True Positives (TP) are when the positive sample is accurately recognized as the positive sample & volleyball image is properly recognized as the volleyball.

True Negatives (TN) are when the negative samples are accurately recognized as negative samples, however basketball images are not recognized since the algorithm incorrectly believes they are basketball.

False Positive (FP) are when a negative sample is by mistake identified as a positive sample, i.e., an image of a basketball is by mistake identified as a volleyball.

False Negatives (FN) occur when positive samples are mistakenly recognized as negative samples, images of volleyballs are not recognized, & the system incorrectly believes they are basketballs.

Precision is the proportion of images accurately recognised, and here is where True excels. This is the proportion of all recognised volleyballs in this theory that are genuine volleyballs.

Percentage of positive samples in the test set that are accurately identified as positive is referred as Recall rate.

P symbolizes precision and R symbolizes recall in the PR curve, which depicts the connection between P & R. In general, recall is assigned the x-co-ordinate,whereas precision is assigned the y-coordinate.

D. Average Precision and Mean Average Precision

Using recall rate and accuracy, a curve may be formed for each class in object identification, where AP is the area of the curve and MAP is the mean value of each category acquired by AP. Greater the MAP value, better the detector's accuracy

6. Summary and Analysis

Detection of objects is a significant & difficult computer vision task which has drawn significant public interest. This article provides a thorough explanation of the various target detection techniques. The following areas may be the subject of target detection research in the future:

In addition to operating reliably on mobile devices, lightweight object detection also greatly reduces working hours. It has uses in facial recognition and smart cameras [14]. The speed between the computer and human vision, especially when identifying relatively small things, is still extremely different when detecting targets [15][16].

Video object identification: When detecting video targets, there are numerous circumstances that lead to high precision in the detection process, including quick motion that blurs the target, out-of-focus video, small target, occlusion, and more. Future studies will concentrate on complex data and sporting goals.

Weakly supervised detection: Deep learning-based detectors are often trained on a substantial amount of annotated picture data. The annotating procedure is labour-intensive, costly, & ineffective [29]. For training the detector, weakly supervised detection techniques simply use image level annotation or a portion of boundary box, and it can both lower costs and provide more accurate models.

Small-object detection: It has never been easy to find small objects in images. Future applications could incorporate the construction of high-resolution lightweight networks and incorporate visual attention methods

7. Conclusion

In conclusion, object detection and deep learning technology have proved to be effective in various fields such as surveillance, automated driving, and facial recognition. The development of computer vision has opened up new perspectives and possibilities. However, detecting a particular object in a complex image or video footage remains a challenge. This study provides a comprehensive review of various object detection algorithms, their challenges, and their effectiveness, including one-stage and two-stage detectors. As the technology continues to advance, the applications of image object detection will continue to expand and become increasingly important in our day-to-day lives. The one-stage pipeline is gaining popularity due to its efficiency and accuracy in real-time applications. The future prospects of research in target image detection have also been discussed, and it is expected that the development of more efficient and accurate object detectors will be a focus of future research.

References

[1] Scalable Learning for Object Detection with GPU Hardware, Adam Coates, Paul Baumstarck, Quoc Le, and Andrew Y. Ng

[2] Object Recognition, Ming-Hsuan Yang University of California at Merced

[3] R. Girshick, “Fast r-cnn,” in Proc. of the IEEE Int. Conf. on Computer Vision, Santiago, Chile, pp. 1440–1448, 2015.

[4] Q. Zhu, M. C. Yeh, K. T. Cheng and S. Avidan, “Fast human detection using a cascade of histograms of oriented gradients,” in Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, New York, NY, USA, pp. 1491–1498, 2006.

[5] S. Maji, A. C. Berg and J. Malik, “Classification using intersection kernel support vector machines is efficient,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, pp. 1–8, 2008

[6] You Only Look Once: Unified, Real-Time Object Detection J Redmon ,et. a, University of Washington, Allen Institute for AI, Facebook AI Research

[7] J. Redmon and A. Angelova. Real-time grasp detection using convolutional neural networks. CoRR, abs/1412.3128, 2014

[8] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu et al., "Attentive contexts for object detection," IEEE Transactions on Multimedia, vol. 19, no. 5, pp. 944–954, 2017.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR, 2014.

[10] R. Girshick, "Fast r-cnn," in ICCV, 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "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, 2017.

[12] P. Druzhkov and V. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," Pattern Recognition and Image Anal., vol. 26, no. 1, p. 9, 2016.

[13] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in ICCV, 2015.

[14] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," in CVPR, 2017.

[15] Z. Hao, Y. Liu, H. Qin, J. Yan, X. Li, and X. Hu, "Scale-aware face detection," in CVPR, 2017.

[16] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face r-cnn," arXiv:1706.01061,2017.

[17] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," arXiv:1701.08289, 2017.

[18] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," arXiv:1509.04874, 2015.

[19] Y. Li, B. Sun, T. Wu, and Y. Wang, “face detection with end-to-end integration of a convnet and a 3d model,” in ECCV, 2016.

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, 2016.

[21] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao,“Yolov4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020.

[22] A.Peng, N. Wang, X. Gao, and J. Li, “Face recognition from multiplestylistic sketches: Scenarios, datasets, and evaluation,” in ECCV, 2016.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in CVPR, 2016.

[24] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards realtime object detection with region proposal networks,” in NIPS, 2015,pp. 91–99.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed,C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multiboxdetector,” in ECCV. Springer, 2016, pp. 21–37.

[26] T.-Y. Lin, P. Doll´ar, R. B. Girshick, K. He, B. Hariharan,and S. J. Belongie, “Feature pyramid networks for object detection.” in CVPR, vol. 1, no. 2, 2017, p. 4.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, “Focal loss for dense object detection,” IEEE transactions on pattern analysis and machine intelligence,2018.

[28] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 734–750.

[29] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, “Object detection with deep learning: A review,” IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212–3232, 2019.

[30] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov,and S. Zagoruyko, “End-to-end object detection with transformers,” in European Conference on Computer Vision. Springer, 2020, pp. 213–229.

[31] D. G. Lowe, “Object recognition from local scaleinvariant features,” in ICCV, vol. 2. Ieee, 1999, pp. 1150–1157.

[32] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov,“Scalable object detection using deep neural networks,” in CVPR, 2014, pp. 2147–2154.

[33] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian,“Centernet: Keypoint triplets for object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6569–6578.