Deep Learning Based Object Detection

Subjects: Others

Contributor: Khurram Azeem Hashmi

Object detection is a complex problem due to underlying high intra-class and low inter-class variance. High intra-class variance is the consequence of different objects belonging to a single class, for instance, different poses of humans or humans wearing different clothes in an image. Low inter-class variance is the outcome of similar-looking objects belonging to different classes such as samples of class chair can easily be misclassified into the class bench and vice versa.

object detection challenging environments low light image enhancement complex environments state of the art deep neural networks computer vision performance analysis

1. Introduction

Object detection is considered as one of the most important and elementary tasks in the field of computer vision. The problem of object detection deals with the identification and spatial localization of objects present in an image or a video [1]. The task of object detection covers a wide range of many other computer vision tasks, such as instance segmentation [2][3][4], visual question answering [5], image captioning [6][7], object tracking [8], activity recognition [9][10][11] and so on.

One of the earlier approaches for object detection algorithms relied on sliding windows, applying classification on each window to find objects [12][13][14]. Later, the sliding window concept was replaced with region proposals to narrow the search before applying classification [15][16][17][18][19]. The recent surge in deep learning has given rise to object detection systems along with other fields.

The prior published work in object detection can be further classified into three categories which are explained below. **Figure 1** depicts the basic difference between them: 1. Object Detection (OD) : OD aims at detecting objects regardless of their class category [15][20]. OD algorithms [21][22][23][24] generally propose a large number of possible region proposals, from which, later on, the best possible candidates are selected according to certain criteria. 2. Salient Object Detection (SOD) : SOD algorithms use the human attention mechanism concept to highlight and detect the objects in a picture or video [25][26]. 3. Category-specific Object Detection (COD) : COD aims at detecting multiple objects. Unlike OD and SOD, COD has to predict the category class and the location of the object in the image or video [14][27].

The deep learning-based object detection algorithms are categorized into two-stage object detectors and one-stage object detectors. Two-stage object detection architectures such as R-CNN [14], Fast R-CNN [28] and Faster R-CNN [21] segregate the task of object localization from the object classification task. They employ region proposal techniques to find possible regions where the likelihood of an object's existence is maximum. Later segmentation output and better detection pooling [21] techniques were introduced with Mask R-CNN [23]. On the other hand, one-stage object detection algorithms first generate candidate regions, and then these regions are classified as object/no-object. For instance, one-stage detectors such as YOLO [22][29][30][31] and SSD [24] work with feature pyramid networks (FPNs) [32] as a backbone to detect objects at multiple scales in a single pass rather than first predicting regions and then classifying them.

2. Related Surveys on Object Detection

There are many surveys carried out on the topic of object detection [33][34][35][36]. This section covers some of the prior surveys.

Han et al. [37] organized the survey in which deep learning techniques for salient and category-specific object detection are reviewed. In 2019, Zou et al. [38] performed an extensive survey on object detection methods that have been proposed in the last 20 years. The authors discussed all the types of object detection algorithms proposed over the years and highlighted their improvements.

Another survey organized by Jiao et al. [39] discussed various deep learning-based methods for object detection. The proposed work provided a comprehensive overview of traditional and modern applications of object detection. Moreover, the authors discussed methods for building better and efficient object detection methods by exploiting existing architectures. Arnold et al. [40] surveyed 3D object detection methods for autonomous driving. The proposed work compared various 3D object detection-based approaches.

It is vital to mention that all of the prior surveys have focused on the general problem of object detection. Although these surveys explain how object detection has improved over the years, they do not cover the challenges and solutions to improve object detection performance in a challenging environment such as low light, occlusions, hidden objects, and so on. To the best of our knowledge, we provide the first survey that reviews the performance of deep learning-based approaches in the field of object detection in a challenging environment.

3. Experiments

We have investigated the performance of current state-of-the-art object detection algorithms on the three most challenging datasets. The idea is to conduct an analysis that explains how well object detection algorithms can perform under harsh conditions. We employed Faster R-CNN [21], Mask R-CNN [23], YOLO V3 [30], Retina-Net [41], and Cascade Mask R-CNN [42] to benchmark their performance on the datasets of ExDARK [43], CURE-TSD [44], and RESIDE [45].

We have leveraged the capabilities of transfer learning in our experiments. All the object detection networks are incorporated with a backbone of ResNet50 [46] pre-trained on the COCO dataset [47]. We fine-tuned all the models for 15 epochs with a learning rate of 2× 10− 5 and used Adam [48] as an optimizer. We resized images to 800 × 800 during the training and testing phases.

4. Evaluation

This section discusses the well-known evaluation criteria essential to standardize state-of-the-art results for object detection in difficult situations. Moreover, this section analyzes the performance of the approaches discussed in Section 3 with quantitative and qualitative illustrations. Finally, we will present the outcome of our experiments on the three most widely exploited challenging datasets.

The standardization of how to assess the performance of approaches on unified datasets is imperative. Since object detection in a challenging environment is identical to generic object detection, the approaches appraise similar evaluation metrics.

Precision [49] defines as the percentage of a predicted region that belongs to the ground truth. **Figure 1** illustrates an the difference between precise object detection and imprecise object detection. The formula for precision is explained below: (1) Predicted area in ground truth Total area of predicted region = TP TP + FP where TP denotes true positives and FP represents false positives.

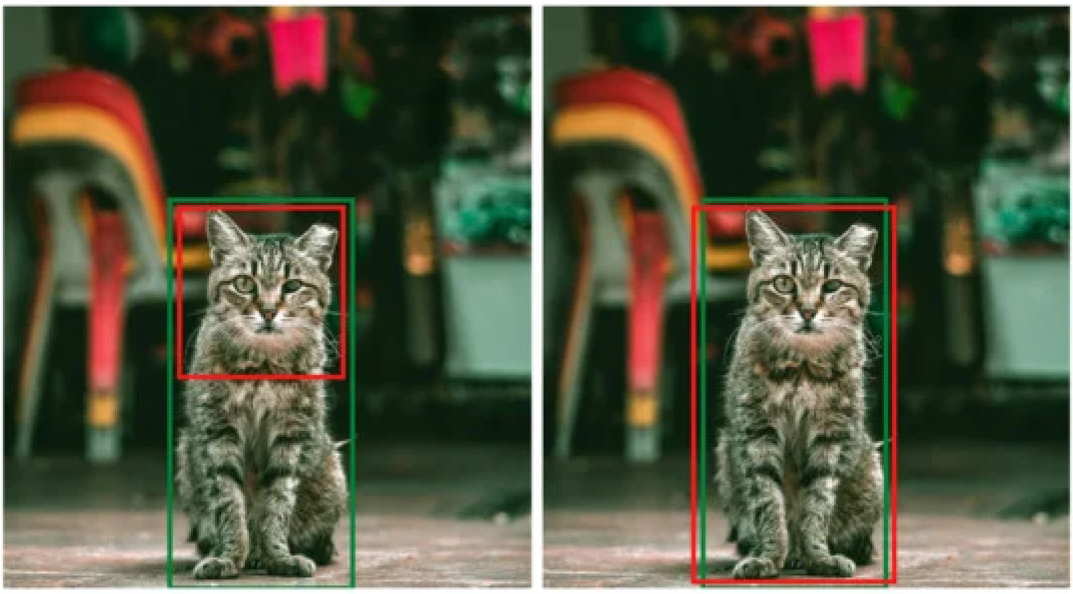

Figure 1. The image explains the visual difference between precise and imprecise prediction in object detection. The green color represents the ground truth, and the red color depicts the predicted boundary. Considering the IOU threshold value equals 0.5, the left prediction is not precise because the IOU between the ground truth and the inferred bounding box is less than 0.5. The bounding box prediction on the right side is precise because it covers almost the complete ground truth area.

The main reason for the low performance of these state-of-the-art generic object detection algorithms is that they are not trained on challenging datasets that include low-light images or occluded images. Furthermore, the backbone network of these architectures cannot optimally extract the spatial features necessary for detecting objects in challenging environments. Hence, it is empirically established that generic object detection algorithms are not ideal for resolving object detection in challenging images.

References

1. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 2009, 32, 1627–1645.
2. Dai, J.; He, K.; Sun, J. Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1July 2016; pp. 3150–3158.
3. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 447–456.
4. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2014; pp. 297–312.
5. Alberti, C.; Ling, J.; Collins, M.; Reitter, D. Fusion of detected objects in text for visual question answering. arXiv 2019, arXiv:1908.05054.
6. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; 2015; pp. 2048–2057, PMLR 37:2048-2057.
7. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; Van Den Hengel, A. Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 1367–1381.
8. Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Trans. Circuits Syst. Video Technol. 2017, 28, 2896–2907.
9. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; pp. 1112–1121.

10. Vaswani, N.; Chowdhury, A.R.; Chellappa, R. Activity recognition using the dynamics of the configuration of interacting objects. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; Volume 2, p. II-633.
11. Motwani, T.S.; Mooney, R.J. Improving Video Activity Recognition using Object Recognition and Text Mining; ECAI; Citeseer: Forest Grove, OR, USA, 2012; Volume 1, p. 2.
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
13. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. Cascade object detection with deformable part models. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2241–2248.
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 17–19 June 2014; pp. 580–587.
15. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 34, 2189–2202.
16. Carreira, J.; Sminchisescu, C. CPMC: Automatic object segmentation using constrained parametric min-cuts. IEEE Trans. Pattern Anal. Mach. Intell. 2011, 34, 1312–1328.
17. Rahtu, E.; Kannala, J.; Blaschko, M. Learning a category independent object detection cascade. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1052–1059.
18. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. Int. J. Comput. Vis. 2013, 104, 154–171.
19. Zitnick, C.L.; Dollár, P. Edge boxes: Locating object proposals from edges. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2014; pp. 391–405.
20. Kuo, W.; Hariharan, B.; Malik, J. Deepbox: Learning objectness with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 2479–2487.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv 2015, arXiv:1506.01497.

22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
23. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. Available online: http://gitlinux.net/assets/SSD-Single-Shot-MultiBox-Detector.pdf (accessed on 21 July 2021).
25. Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; Shum, H.Y. Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. 2010, 33, 353–367.
26. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.; Hu, S.M. Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. 2014, 37, 569–582.
27. Ouyang, W.; Wang, X.; Zeng, X.; Qiu, S.; Luo, P.; Tian, Y.; Li, H.; Yang, S.; Wang, Z.; Loy, C.C.; et al. Deepid-net: Deformable deep convolutional neural networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2403–2412.
28. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
29. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
30. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. arXiv 2018, arXiv:1804.02767.
31. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
33. Agarwal, S.; Terrail, J.O.D.; Jurie, F. Recent advances in object detection in the age of deep convolutional neural networks. arXiv 2018, arXiv:1809.03193.
34. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7310–7311.

35. Grauman, K.; Leibe, B. Visual object recognition. Synth. Lect. Artif. Intell. Mach. Learn. 2011, 5, 1–181.

36. Andreopoulos, A.; Tsotsos, J.K. 50 years of object recognition: Directions forward. Comput. Vis. Image Underst. 2013, 117, 827–891.

37. Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. Advanced deep-learning techniques for salient and category-specific object detection: A survey. IEEE Signal Process. Mag. 2018, 35, 84–100.

38. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. arXiv 2019, arXiv:1905.05055.

39. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. IEEE Access 2019, 7, 128837–128868.

40. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A survey on 3d object detection methods for autonomous driving applications. IEEE Trans. Intell. Transp. Syst. 2019, 20, 3782–3795.

41. Jaeger, P.F.; Kohl, S.A.; Bickelhaupt, S.; Isensee, F.; Kuder, T.A.; Schlemmer, H.P.; Maier-Hein, K.H. Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. In Proceedings of the Machine Learning for Health NeurIPS Workshop, Durham, NC, USA, 7–8 August 2020; pp. 171–183.

42. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 43, 1483–1498.

43. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the exclusively dark dataset. Comput. Vis. Image Underst. 2019, 178, 30–42.

44. Temel, D.; Chen, M.H.; AlRegib, G. Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics. IEEE Trans. Intell. Transp. Syst. 2019, 21, 3663–3673.

45. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Benchmarking single-image dehazing and beyond. IEEE Trans. Image Process. 2018, 28, 492–505.

46. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

47. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollar, P. Microsoft COCO: Common objects in context (2014). arXiv 2019, arXiv:1405.0312.

48. Zhang, Z. Improved adam optimizer for deep neural networks. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada, 4–6

June 2018; pp. 1–2.

49. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv 2020, arXiv:2010.16061.