Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Object Detection Based on Deep Learning

Hang Yin *

Chengdu College of University of Electronic Science and Technology of China, Sichuan, China

* Corresponding Author Email: 100573@yzpc.edu.cn

Abstract. Object detection technology is a crucial research direction in the field of computer vision. Its primary task is to identify and locate objects within images. This technology finds extensive applications across various domains such as autonomous driving, medical diagnostics, and security monitoring. With the development of machine learning, especially the progress of deep learning in related fields of image processing, the accuracy rate of object detection has achieved better results. However, different algorithms have different characteristics, which increases the difficulty of algorithm selection in actual use. This paper starts with the development of object detection, and summarizes three aspects: data set, metrics and object detection algorithm. This paper mainly introduces the commonly used data sets of object detection (such as COCO, VOC), traditional object detection algorithms and deep learning-based object detection algorithms. In addition, the current challenges encountered in object detection, especially in the detection of small objects, and future research directions are discussed.

Keywords: Object detection; deep learning; small object detection; YOLO; Fast R-CNN.

1. Intruduction

Object detection, a fundamental problem in the field of computer vision, involves identifying all relevant objects within an image and determining their categories and locations. It serves as a cornerstone for computer understanding and analysis of images, finding broad applications across various domains including autonomous driving, security monitoring, medical diagnosis, among others. Specifically, the task of object detection can be subdivided into the following aspects:

First and foremost is the identification of objects. Object detection algorithms are tasked with locating all objects belonging to predefined categories within an image. These categories exhibit significant diversity across different scenarios, encompassing entities such as humans, vehicles, animals, and everyday items. For instance, in the context of autonomous driving, categories that may require identification include pedestrians, vehicles, and traffic signs, whereas in security monitoring scenarios, categories to be identified may encompass individuals, vehicles, and suspicious items.

Subsequently, ascertaining the object's position becomes essential. Beyond categorizing objects, effective object detection mandates precise localization of each object within the image. Two approaches prevail in determining object localization: one entails pinpointing the object’s centroid and dimensions, while the other involves specifying the coordinates of its top-left and bottom-right corners, both of which can be encapsulated within bounding boxes. The inherent diversity in object appearances, shapes, and orientations, compounded by imaging challenges such as lighting variations and occlusion, significantly heightens the complexity of object detection. Thus, algorithms must possess robust feature extraction and analytical capabilities to accurately discern and position objects amidst diverse and intricate scenarios.

This paper offers a dual perspective on object detection, focusing on datasets and algorithms. Chapter 2 introduces and elaborates on two pivotal datasets, Coco and Voc, along with their associated evaluation metrics. Chapter 3 delves into the evolution of object detection algorithms, spanning from early approaches to contemporary methodologies driven by deep learning and transformer-based techniques. Chapter 4 addresses the prevalent challenges in modern object detection paradigms and

presents corresponding solutions, while also outlining potential future research directions. Finally, the concluding chapter provides a cohesive summary of the entire paper.

2. Datasets and Metrics

2.1. Datasets

2.1.1. COCO dataset

The COCO (Common Objects in Context) dataset, released by the Microsoft team in 2014, represents a substantial and comprehensive resource for tasks involving object detection, segmentation, and captioning. With scene understanding as its primary objective, this dataset is meticulously compiled from complex real-world scenes, with precise localization of objects achieved through meticulous segmentation. It holds considerable significance within the realm of computer vision. Comprising over 330,000 images, with approximately 200,000 images extensively annotated, the COCO dataset spans 91 categories, encompassing 80 object classes that include a diverse range of common objects like humans, cars, and motorcycles [1]. Each image not only features annotations for object instances but also provides five distinct image descriptions, aiding in image captioning tasks. Furthermore, the COCO dataset is partitioned into training, validation, and test sets, each accompanied by detailed annotation information [2].

2.1.2. VOC dataset

The VOC (Visual Object Classes) dataset emerges as a highly valuable and widely applied asset, developed by the computer vision group at the University of Oxford, and routinely utilized as the standard dataset in the annual PASCAL VOC challenge. Encompassing 20 distinct object categories, ranging from transportation modes like cars and airplanes, to household items such as chairs and tables, alongside animals like cats, dogs, and human subjects, the dataset's diverse array of classes positions it as an ideal testbed for evaluating and refining image recognition and understanding algorithms [3].

A distinguishing characteristic of the PASCAL VOC dataset is its varied annotation types. While some images feature annotations at the pixel level, facilitating nuanced image segmentation tasks, others offer only object-level annotations, delineating solely the bounding boxes of objects [4]. This hierarchical variation in annotations grants researchers flexibility, empowering them to select annotation levels tailored to their specific research objectives.

In summation, both datasets represent indispensable resources in computer vision research, containing rich repositories of image and annotation data. They play pivotal roles in the study and advancement of state-of-the-art object detection and image segmentation algorithms.

2.2. Metrics

Evaluation metrics in object detection primarily encompass mAP (mean Average Precision) and Precision. In the realm of object detection tasks, mAP serves as a pivotal gauge of performance. It derives as the mean value computed from the AP (Average Precision) values across different categories. The following offers a detailed elucidation:

True Positives (TP). These signify the count of correctly identified instances, i.e., objects accurately detected. In object detection, if a detection box intersects with any ground truth (GT) box with an Intersection over Union (IoU) exceeding 0.5, it qualifies as a TP. Each GT box contributes to the count only once [5].

False Positives (FP). These denote instances wrongly identified, i.e., objects incorrectly detected. If a detection box has an IoU of 0.5 or lower with all GT boxes, or if it detects the same GT box multiple times (more than once), it is deemed an FP.

False Negatives (FN). These represent instances missed during detection, i.e., undetected objects. If a GT box remains unspotted by any detection box, it is categorized as an FN.

Precision. This metric gauge the ratio of correctly identified positive instances among all instances identified as positive.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

Recall, or sensitivity, refers to the proportion of positive samples correctly detected among all actual positive instances.

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

AP (Average Precision) refers to the area under the curve of precision as recall varies. In practical applications, Precision and Recall values are typically computed at various confidence thresholds, and AP is calculated using interpolation.

mAP(mean Average Precision) represents the average of AP values across all classes.

$$mAP = \frac{(AP1 + AP2 + \cdots + APn)}{n} \tag{3}$$

Where n is the number of categories.

Intersection over Union (IoU), also known as the Jaccard index, represents a fundamental evaluation metric extensively employed in tasks including object detection, semantic segmentation, and tracking. This metric quantifies the overlap between predicted bounding boxes and their corresponding ground truth bounding boxes.

$$IoU = \frac{A \cap B}{A \cup B} \tag{4}$$

MAP, Precision, Recall, and IoU represent various evaluation metrics for object detection. Precision and Recall focus on assessing classification accuracy, whereas IoU measures localization. mAP amalgamates both classification and localization metrics, providing a holistic evaluation of object detection algorithms. Hence, mAP emerges as the principal assessment metric, offering a comprehensive performance evaluation across different classes and IoU thresholds. In real-world applications, employing the COCO dataset for object detection often involves utilizing the COCO API to compute metrics like mAP.

3. Algorithm

The early phases of object detection research predominantly relied on traditional image processing techniques, including approaches grounded in geometry and machine learning algorithms. However, the surge of deep learning in recent years has captivated the interest of researchers, prompting a notable transition towards deep learning-based methodologies. Subsequent sections elaborate on various facets, encompassing geometric-based and machine learning-driven algorithms.

3.1. Traditional Image Processing Methods

In the initial phases of object detection research, feature extraction predominantly relied on template matching and analysis of color, shape, and texture, among other methods, followed by classification using machine learning algorithms. Regarding localization, the prevalent technique involved image segmentation via sliding windows, followed by the derivation of object localization results based on classification outcomes [6].

3.1.1. Geometric-based methods

Template matching is a technique to identify objects by comparing specific regions of an image with predefined templates to assess similarity. While this method performs well in dealing with objects of fixed shapes and sizes, its efficacy may be constrained when faced with variations in shape and size. Color, shape, and texture analysis entail identifying objects by analyzing color distributions, shape features, and texture patterns within images [7]. These methods prove highly effective in handling objects with specific colors, shapes, or textures.

Model-based and rule-based approaches rely on predefined rules and models to identify objects. Typically, these methods necessitate extensive manual design and adjustment to adapt to diverse application scenarios.

While these conventional methods were useful given the computational resources available at the time, they often required substantial manual feature engineering and exhibited limited generalization capabilities for new, unseen datasets.

3.1.2. Feature extraction and descriptors

Utilizing feature descriptors for object detection represents a simplistic methodology, predominantly reliant on manually crafted feature operators for extracting crucial points and edge details from images. Operators like ORB and SIFT undergo a meticulously designed process to exploit various geometric attributes within images, encompassing gradients, colors, and textures, for feature extraction purposes. Descriptors, meanwhile, furnish mathematical representations of the adjacent regions surrounding these feature points. By sampling and encoding these regions, fixed-length vectors are formulated, thereby streamlining subsequent image-processing endeavors [8].

3.1.3. Early object detection algorithms

Before the widespread integration of deep learning techniques in object detection, earlier methods such as the Viola-Jones detector, Histogram of Oriented Gradients (HOG), and Deformable Part Models (DPM) were prominent and extensively applied. These methods played pivotal roles in the field of object detection during their time.

However, as technology progressed, these conventional approaches began to exhibit certain limitations. For example, they demonstrated relatively weak generalization capabilities, leading to decreased performance when encountering novel, unseen data. Furthermore, these algorithms suffered from suboptimal computational efficiency, resulting in slower processing speeds, thus constraining their widespread utilization in practical applications [9].

3.2. Object Detection Based on Deep Learning

As deep learning continues to progress, object detection models rooted in this paradigm, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), undergo training on extensive datasets to autonomously discern salient features within images. Departing from traditional reliance on manually engineered feature extractors, these models harness multi-layer neural network structures to inherently learn feature representations across hierarchical levels. This data-centric approach fosters the heightened adaptability of object detection models to a myriad of intricate scenarios and tasks [10].

3.2.1. Two-stage object detection: R-CNN series

Within the realm of object detection based on deep learning, the R-CNN series algorithms serve as emblematic examples of two-stage object detection, comprising R-CNN, Fast R-CNN, and Faster R-CNN.

Initially, R-CNN emerged as a trailblazer in this field, being the earliest algorithm to integrate convolutional neural networks into object detection. Nevertheless, despite its technical breakthroughs, R-CNN exhibits notable computational inefficiencies, characterized by protracted training durations and relatively suboptimal efficiency [11].

To address this issue, Fast R-CNN emerged. This algorithm significantly improves accuracy by employing a multitask loss function, enabling all layers of the network to be simultaneously updated in a single forward pass, eliminating the need to store large amounts of feature data. This enhancement greatly enhances training efficiency. However, Fast R-CNN still suffers from a dependency issue, as it still relies on the Selective Search algorithm to generate candidate regions, thereby somewhat limiting its optimization potential in terms of computational speed.

Following this, Faster R-CNN has further propelled the advancement of object detection technology. This algorithm introduces the Region Proposal Network (RPN), enabling the sharing of convolutional features, which not only substantially reduces computational costs but also decreases model complexity. Through such design, Faster R-CNN achieves significant improvements in both speed and performance. However, when confronted with challenges such as small object detection or class imbalance, Faster R-CNN still requires some targeted enhancements.

The developmental trajectory of the R-CNN series algorithms has been instrumental in propelling advancements within the domain of object detection. Progressing from R-CNN to Fast R-CNN, and subsequently to Faster R-CNN, each iteration has cemented foundational elements crucial for further exploration and practical implementation. As deep learning technology advances persistently, these algorithms undergo continual refinement and enhancement to accommodate the evolving complexity and variability of application scenarios.

3.2.2. Single-stage object detection: YOLO series and SSD

The YOLO series and SSD are two widely used object detection algorithms, both belonging to the category of single-stage object detection algorithms. The notable advantage of the YOLO series lies in its high speed and excellent real-time performance. To simplify the structure, the YOLO series divides the input image into a grid and performs bounding box regression and classification within the same network. While this approach enhances speed, it sacrifices some localization accuracy. Particularly in small object detection, the performance of the YOLO series is not ideal and tends to generate false positives [12].

Unlike the YOLO series, SSD utilizes multi-scale feature maps for detection, resulting in improvements in both speed and accuracy, particularly when compared to YOLOv1. However, SSD's performance may falter when handling extremely small objects or large-sized images.

Hence, when selecting an appropriate object detection algorithm, it is crucial to consider the trade-off between speed and accuracy based on specific application scenarios and needs. For example, if real-time requirements are high and precision demands are relatively modest, then the YOLO series may offer a better fit. Conversely, if precision is paramount while real-time considerations are less critical, SSD might be more appropriate.

3.2.3. Transformer-based object detection

Within the current landscape of deep learning-based object detection, Transformer-based methodologies have garnered significant interest due to their unique strengths. These approaches feature global attention mechanisms, enabling models to factor in the entire image during processing, as opposed to concentrating solely on local regions. Moreover, Transformers exhibit exceptional prowess in feature learning, proficiently extracting valuable features from images. Their end-to-end

detection capability facilitates seamless progression from raw data to conclusive results, eliminating the necessity for complex intermediary procedures. Simultaneously, they employ efficient label-matching strategies, thereby augmenting detection precision.

The merits outlined above enable Transformer-based object detection approaches to excel in detecting small objects and managing diverse datasets. Notably, in the realm of small object detection tasks, Transformers showcase a capacity for precise identification and localization of diminutive, often neglected Notably, in the realm of small object detection tasks, Transformers showcase a capacity for precise identification and localization of diminutive, often neglected Notably, in the realm of small object detection tasks, Transformers showcase a capacity for precise identification and localization of diminutive, often neglected objects. Moreover, across heterogeneous datasets, Transformers demonstrate versatility in accommodating different image types, thereby achieving accurate object detection [13].

Although Transformers offer considerable advantages in object detection, they also come with noteworthy drawbacks. Primarily, Transformers demand significant computational resources, thereby requiring high-end hardware configurations to execute such models efficiently. This demand escalates notably with high-resolution image processing, owing to the additional detail involved, necessitating even more computational resources. Secondly, the elevated complexity of Transformer models not only complicates model training but may also exacerbate overfitting concerns. Finally, real-time performance poses a formidable challenge, particularly in applications demanding swift responses, like autonomous driving or real-time monitoring.

4. Challenges and Solutions in Object Detection

4.1. Small Object Detection and Object Detection in Dense Scenes

In real-world applications, object detection technology faces several challenges, particularly in its performance within scenarios involving small objects and dense environments.

Small object detection is hindered by the diminutive size of the objects, resulting in less distinctive visual features and thereby escalating detection difficulty. Furthermore, regions containing small objects often exhibit significant noise originating from various interferences during image acquisition, transmission, or processing, thereby further complicating the detection process. Concurrently, existing datasets may lack comprehensive coverage of small objects, thereby inadequately providing sufficient samples to train effective detection models.

Meanwhile, within dense scenes, the overlap between objects and their resemblance to the background presents another formidable challenge for detection tasks. Distinguishing between individual objects becomes notably challenging when multiple objects are spatially close or partially obscured. Additionally, complex backgrounds may share similar features with the objects, thereby disrupting the judgment of detection algorithms.

4.1.1. Solution to small object detection and dense scene

In response to these challenges, researchers have proposed several approaches to enhance the performance of small object detection. Among them, multi-scale learning emerges as a potent strategy, facilitating feature extraction across various scales to better capture the intricacies of small objects. Contextual information learning, on the other hand, harnesses the relationships between objects and their surrounding context to enrich the model's comprehension of object contexts, thereby refining detection accuracy. Augmenting data techniques and Generative Adversarial Networks (GAN) contribute to improved model generalization by augmenting training data diversity and realism. Anchor-free methodologies, by contrast, discard conventional anchor-based detection paradigms, thus streamlining computational overheads and heightening detection efficiency [14].

For the detection of dense scenes, contextual models assist algorithms in comprehending the spatial relationships between objects, while instance segmentation techniques accurately delineate the boundaries of individual objects, thus improving detection precision.

With ongoing research and technological innovation, these challenges are progressively being surmounted. Continuous optimization of algorithms and model architectures, coupled with the utilization of superior quality datasets for training, are enhancing the performance of object detection technology in scenarios involving small objects and dense scenes, thereby offering more precise and dependable detection services across diverse application domains.

4.2. Emerging Trends and Challenges in Object Detection

4.2.1. Cross-domain and few-shot learning

Within the realm of object detection, scholars have been exploring avenues to bolster model generalization amidst limited data scenarios. To tackle this challenge, cross-domain learning and few-shot learning have emerged as two pivotal research trajectories.

The core idea of domain adaptation is to transfer knowledge learned in a specific domain (source domain) to another domain with different data distributions (object domain). The challenge of this process lies in the significant disparities between the source and object domains, which may encompass distinct feature spaces, label spaces, or data distributions. To address these challenges, researchers have proposed various techniques such as adversarial training and distribution matching. Adversarial training introduces a discriminator to distinguish between data from the source and object domains, thereby encouraging the model to learn a more generalized representation. Distribution matching, on the other hand, aims to align the data distributions of the source and object domains to reduce their disparities.

Sample-efficient learning focuses on how to effectively learn from a limited number of annotated samples. The challenge in this research field lies in extracting sufficient information from these scarce samples so that the model can accurately identify new, unseen data. To address this challenge, researchers have developed various methods, including meta-learning, transfer learning, and data augmentation. Meta-learning is an approach that involves learning how to quickly adapt to new tasks, enabling the model to rapidly learn to recognize new categories after seeing a small number of samples. Transfer learning, on the other hand, applies knowledge learned from one task to another related task to improve the model's performance on the new task. Data augmentation involves increasing the diversity of existing data to simulate more training samples, thereby enhancing the model's generalization ability [15].

Both cross-domain learning and few-shot learning methodologies are geared towards augmenting a model's adaptability and efficacy in addressing novel environments and tasks. Through these methodologies, researchers seek to cultivate more intelligent object detection systems capable of precisely discerning and pinpointing object objects within images, even amidst data scarcity. Such endeavors not only propel the progress of object detection techniques but also furnish invaluable lessons and innovation for other domains grappling with data limitations.

4.2.2. Adaptive and online learning

Adaptive learning and online learning aim to enable models to adapt more flexibly and effectively to continuously changing and evolving data patterns.

Adaptive learning represents an advanced machine learning approach involving automatic adjustments of model parameters to better fit new data distributions. The challenge lies in devising algorithms capable of swiftly adapting to new environments while maintaining model performance stability. To address this challenge, researchers have proposed various solutions, including but not limited to incremental learning and transfer learning. Incremental learning allows models to retain knowledge of old tasks while learning new ones.

In tandem with adaptive learning, online learning emerges as a method enabling models to learn and update their parameters in real time as they process continuous data streams. The challenge inherent in online learning is to ensure stability and accuracy amid ongoing learning processes. To address these challenges, researchers have devised various strategies, such as active learning and meta-learning. Active learning entails selectively extracting the most beneficial information from the data for learning purposes.

Both adaptive learning and online learning methodologies have significantly enhanced the adaptability and generalization capabilities of models. They enable models to not only perform exceptionally well on static datasets but also maintain efficient and accurate performance in dynamic and evolving environments. Such flexibility and adaptability are highly valuable in the field of object detection, as they imply that models can better accommodate the varied scenarios present in the real world.

4.2.3. Adversarial attacks and defenses

Within the domain of object detection, adversarial attacks and defenses stand as central concerns regarding model security. Adversarial attacks, leveraging model vulnerabilities through intricately designed inputs, deceive models. The challenge therein lies in formulating efficient attack strategies to exploit model vulnerabilities maximally. To tackle this issue, researchers have proposed various solutions, including FGSM (Fast Gradient Sign Method) and PGD (Projected Gradient Descent). These technologies proficiently assail models, thus uncovering their latent vulnerabilities.

Adversarial defense mechanisms are deployed to mitigate the influence of adversarial attacks on models while maintaining their performance. The key challenge lies in bolstering model robustness without sacrificing efficiency. To address this challenge, researchers have proposed several strategies, including adversarial training, input transformation, and feature extraction. Adversarial training incorporates adversarial examples into the training process to enhance the model's ability to recognize and withstand adversarial attacks. Input transformation modifies input data to help the model better handle adversarial examples, while feature extraction aims to improve the model's resilience by extracting more robust features.

A thorough grasp of the principles underlying adversarial attacks and defense is essential for effectively safeguarding models and ensuring their stability in the face of various threats. Therefore, whether in theoretical exploration or practical implementation, a comprehensive understanding of these concepts is indispensable.

5. Conclusion

This paper comprehensively examines various aspects of object detection, encompassing its significance in the field of computer vision, specific tasks, as well as the challenges and solutions currently encountered. Initially, the paper outlines the pivotal task of object detection: identifying objects within images and determining their respective locations. Subsequently, it elaborates on two widely utilized datasets, COCO and VOC datasets, along with associated evaluation metrics such as mAP, Precision, Recall, and IoU.

Regarding algorithms, this paper meticulously traverses from conventional image processing techniques like template matching and geometric methods to deep learning-based object detection frameworks such as the R-CNN series, single-stage YOLO, and SSD. Each approach is extensively expounded upon, delineating its inherent strengths and weaknesses, while emphasizing their applicability across varied scenarios and demands.

Confronting the challenges of small object detection and dense environments, this paper introduces remedies such as multi-scale learning, contextual information acquisition, and data augmentation techniques. Furthermore, it deliberates on the significance of emerging trends like cross-domain learning and few-shot learning.

Lastly, the manuscript underscores the perpetual evolution of object detection technology. Through algorithmic refinements, structural enhancements of models, and the utilization of high-fidelity datasets, the performance of object detection is steadily ascending across diverse scenarios, furnishing precise and dependable services for a myriad of applications.

References

[1] Cao Y, Lu X, Zhu Y, et al. Context-based fine hierarchical object detection with deep reinforcement learning. 2020 7th International Conference on Information Science and Control Engineering (ICISCE), 2020.

[2] Pang Y, Cao J. Deep learning in object detection. Deep Learning in Object Detection and Recognition, 2019, 26(7): 19-57.

[3] Song D, Liu B, Li Y. Based on end-to-end object detection algorithm with Transformers for detecting wafer maps. 2022 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2022.

[4] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[6] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252.

[7] Han G, Zhang X, Li C. Revisiting faster R-CNN: A deeper look at region proposal network. Neural Information Processing, 2017: 14-24.

[8] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Williams K, Winn J, Zisserman A, et al. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2014, 111(1): 98-136.

[10] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. Computer Vision – ECCV 2014, 2014: 740-755.

[11] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: A review. arXiv e-prints, 2018, 30(11): 3212-3232.

[12] Lu S Y, Wang B Z, Wang H J, et al. A real-time object detection algorithm for video.Computers & Electrical Engineering, 2019, 77(1): 398-408.

[13] Sharma V, Mir R N. A comprehensive and systematic look up into deep learning based object detection techniques: A review. Computer Science Review, 2020, 38(11): 1-29.

[14] Aziz L, Salam M S B H, Sheikh U, et al. Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. IEEE Access, 2020, 8(1): 170461-170495.

[15] Shin H C, Lee K I, Lee C E. Data augmentation method of object detection for deep learning in maritime image. 2020 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2020.