Comparative Analysis of Two-Stage and One-Stage Object Detection Models

Yibo Han[a]
School of Artificial Intelligence, South China Normal University, Guangzhou, China

Keywords: Two-Stage Detectors, Object Detection, Real-Time Applications, One-Stage Detectors.

Abstract: The background and basic ideas of object detection are examined in this essay, with a focus on contrasting one-stage and two-stage detectors. The main goal is to compare and contrast the accuracy and speed of different object identification methods. The study covers a thorough examination of models including Single Shot Multi-Box Detector (SSD), You Only Look Once (YOLO), and the Region Convolution Neural Network (RCNN) series. The study involves preprocessing data from the Pattern Analysis, Statistical modelling and Computational Learning Visual Object Classes (PASCAL VOC), Image Network (ImageNet), Microsoft Common Objects in Context (MS COCO), and Open Images datasets, followed by the application and training of different detection algorithms. Performance metrics, including Mean Average Precision (MAP) and Frames Per Second (FPS), are used to assess the models. The findings indicate that two-stage detectors, such as Faster R-CNN, process information more slowly yet achieve better detection accuracy, especially in complex scenarios. On the other hand, one-stage detectors, such YOLO and SSD, have quicker inference times and are therefore better suited for real-time applications, but precision is usually lost, particularly when identifying smaller objects. This study holds significant implications for fields requiring high-performance object detection, like medical imaging and driverless driving.

1 INTRODUCTION

With the use of associated computer vision and image processing techniques, object detection is a particularly promising field of computer technology. Its main objective is to recognize and process specific semantics (such as glasses, traffic signals, birds, and fish) in static images or videos (Jiao et al., 2019). The distinction between image classification and object detection is that the former requires the exact location of recognized objects in addition to their classification; the latter can be accomplished simply drawing the bounding boxes around the objects in the image (Liu et al., 2020). Today, object detection techniques are employed in many different fields. Autonomous driving, where helps vehicles identify and react to obstacles, pedestrians, and traffic signs in real¬-time (Feng et al., 2020). In traffic safety, these techniques are used to monitor and analyze traffic patterns, detect accidents, and improve overall road safety (Razi et al., 2023). In addition, object identification in medical imaging can precisely locate the tumor's location on the tissue, which helps in early diagnosis and treatment planning (Yang and Yu, 2021).

There are now two primary categories of object detection techniques. One is the conventional approaches, which depend on conventional machine learning and features that are manually created (Zou et al., 2023). Among these are the Histogram of Oriented Gradients (HOG), Sliding Window Method (SWM), and Deformable Part Models (DPM) (Zou et al., 2023). Because deep learning is developing so quickly, another Convolutional Neural Network (CNN)-based object detection method has been gaining popularity (Pathak et al., 2018). One-stage, two-stage, and anchor-free object detectors are the three categories into which this technique falls. One-stage object detectors are quick and ideal for real-time detection since they can regress and classify directly on the picture (Zhang et al., 2021). One-step object detectors include RetinaNet, which enhances the performance of small or obscured item recognition by lowering the weights on samples that are easily

[a] https://orcid.org/0009-0007-0864-6382

Han and Y.
Comparative Analysis of Two-Stage and One-Stage Object Detection Models.
DOI: 10.5220/0013515900004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 289-294
ISBN: 978-989-758-754-2

Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

observable; You Only Look Once (YOLO), which directly predicts bounding boxes and categories using a single neural network; Single Shot MultiBox Detector (SSD), which detects objects of various sizes using multiscale feature maps (Carranza-García et al., 2020). The basic idea behind two-stage object detectors is to first identify a candidate region, after which they do bounding box regression and classification on it. For two-stage object detectors, there are Region-based Convolutional Neural Networks (R-CNN), Fast R-CNN, and Faster R-CNN. R-CNN is computationally intensive despite having a high accuracy rate because it creates candidate regions using selective search and then classifies each region using a CNN (Zhao et al., 2019). Fast R-CNN, which speeds up detection by spreading convolutional features over the whole image and applying region suggestions on the feature map (Zhao et al., 2019). Faster R-CNN, which generates candidate areas straight from the feature map by introducing a Region Proposal Network (RPN), significantly increases detection speed (Zhao et al., 2019). Anchor-free object detectors are of two types, CornerNet, which predicts the object's upper left and bottom right corners to establish the bounding box, and CenterNet, which calculates the bounding box by regressing the height and width and forecasting the object's center (Tian et al., 2020). Along with the already mentioned, there are Detection Transformers (DETR), Multi-task Learning, and Ensemble methods for object detection.

Examining deep learning-based object detection techniques is the primary objective of this work. An outline of the main ideas and background information pertaining to object detection is given in the first part. The second section examines the key technologies that underpin deep learning-based object detection and provides a thorough description of the underlying ideas. An assessment of these technologies' performance is given in the third section, which also covers their benefits, drawbacks, and possible future advancements. The fourth and last section provides a summary of the research findings as well as suggestions for the field's future directions.

2 METHODOLOGY

2.1 Dataset Description

The development of object detection and the solving of several difficult and complicated problems that come up in this field depend heavily on datasets. Datasets can provide an abundance of pictures of different scenes and objects for training various object detection models to learn to recognize different objects in an image, where a substantial quantity of annotated data is essential. At the moment, ImageNet, Microsoft COCO (MS COCO), Open Images (OI), Pattern Analysis, Statistical modelling and ComputAtional Learning Visual Object Classes (PASCAL VOC), and ImageNet are the primary datasets utilized in the object detection sector. Table 1 provides a summary of these datasets' characteristics. A benchmark dataset for object detection and classification is PASCAL VOC. Deep learning in vision has advanced thanks in large part to ImageNet and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). ILSVRC has improved the object detection algorithm standardization training and evaluation of PASCAL VOC by a factor greater than ten (Liu et al., 2020). MS COCO has richer image understanding compared to ImageNet, which has complex life scenes and some common objects in nature. The Open Image Challenge Object Detection (OICOD) belongs to OI, which is the world's largest item detection dataset that is available to the public. Unlike ILSVRC and MS COCO, OICOD only annotates human-identified positively labelled objects.

Table 1: Popular databases for object recognition (Liu et al., 2020).

Name of dataset	Total pictures	Classifications	Picture per category	Objects per picture	Picture dimensions	Started year
PASCAL VOC 2012 (Liu et al., 2020)	11, 500	19	300–4090	2.5	470 x 380	2005
ImageNet (Liu et al., 2020)	14 million+	21,841	-	1.5	500 x 400	2009
MS COCO (Liu et al., 2020)	328,000+	91	-	7.3	640 x 480	2014
Open Images (Liu et al., 2020)	9 millions+	5999+	-	8.3	Various	2017

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

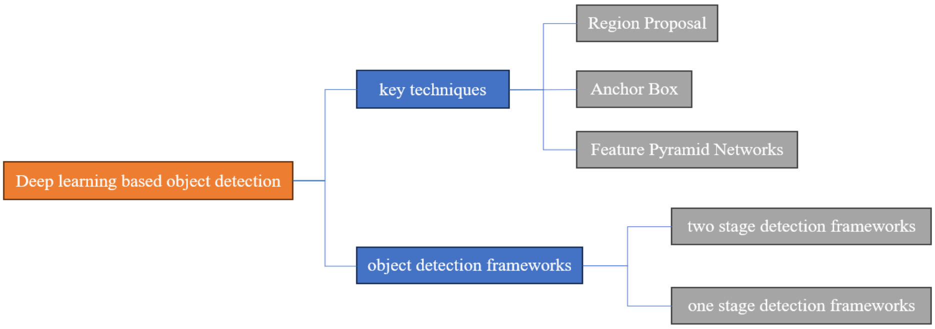
Deep learning based object detection
key techniques
Region Proposal
Anchor Box
Feature Pyramid Networks
object detection frameworks
two stage detection frameworks
one stage detection frameworks

Figure 1: The research overview (Picture credit: Original).

2.2 Proposed Approach

This study's main goal is to investigate deep learning-based object detection methods in-depth, as shown in Figure 1. The rapid breakthroughs in deep learning, particularly in the field of CNN, have resulted in notable improvements in the performance and capabilities of object detection systems in recent years. This paper focuses on key techniques such as Region Proposal (RP), Anchor Box (AB), and Feature Pyramid Networks (FPN), providing a detailed analysis of each. Additionally, the study will look at the one-stage and two-stage object detection frameworks, which are the two main types of frameworks. These are the two primary methods in object detection; each has unique benefits and works well in various application contexts. One-stage frameworks, known for their speed, are often preferred for real-time detection, while two-stage frameworks, offering higher accuracy, are typically used in applications where precision is critical. This paper will compare and contrast these frameworks, highlighting their unique contributions and use cases.

2.2.1 Introduction of Basic Technologies

In object detection, three methods are crucial: RP, AB, and FPN. The primary goal of RP is to produce candidate regions for object detection, and the detection's speed and accuracy are directly impacted by the caliber of the candidate regions. RP adopts two methods, Selective Search (SS) and RPN. SS creates candidate boxes by combining areas of an image that are similar in terms of color, texture, and size. By merging comparable areas of an image with characteristics like size, color, and texture, SS creates candidate frames.

The disadvantages of this method are obvious: the number of candidate regions generated is large and the speed is slow. RPN can produce excellent candidate regions by sliding straight into the feature map. RPN has the speed advantage over convolutional neural networks since it shares computation with them. RP is mainly used in the two stages detector. AB is the predefined bounding box in the object detection model, which has different sizes and box aspects. In the target detection model, AB is a predetermined bounding box with varying sizes and aspect ratios. The presence of an object is detected by placing the bounding box at various locations within the feature map. This allows the model to function at different scales and aspect ratios, which improves its detection performance for objects of different shapes and sizes. AB is frequently used in conjunction with one stage detectors like SSD and YOLO. FPN is a multi-scale feature extraction technique. Because FPN fuses information at different feature map scales, FPN not only excels in detecting small objects, but also efficiently detects large objects. In order to recognize objects at multiple scales, FPN uses top-down feature fusion and bottom-up feature extraction techniques. This allows the model to have a variety of semantic information at each size. FPN is utilized in detectors that are one stage or two stages.

2.2.2 Introduction of Two Stage Detectors

In order for two stage detectors to function, a set of candidate regions must first be created. These candidate regions are then the subject of additional target categorization and exact localization. Its workflow is divided into two stages. The workflow diagram of two stage detectors can be referred to Figure 2. Initially, RPN generates a collection of candidate regions from the feature map that are most likely to contain targets. RPN scans over the feature map by means of a sliding window and predicts the possible target locations. Two stage detectors further

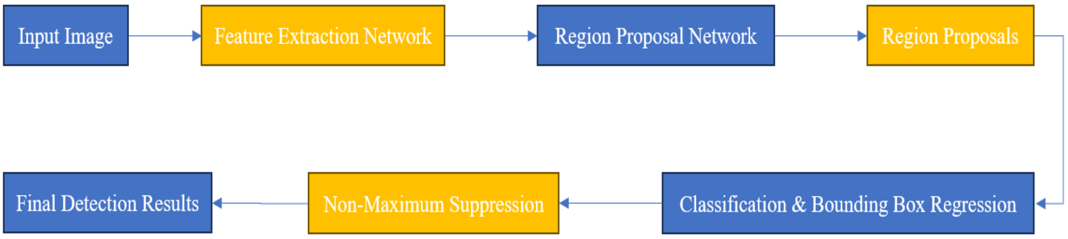

Figure 2: Workflow diagram of two stage detectors (Picture credit: Original).

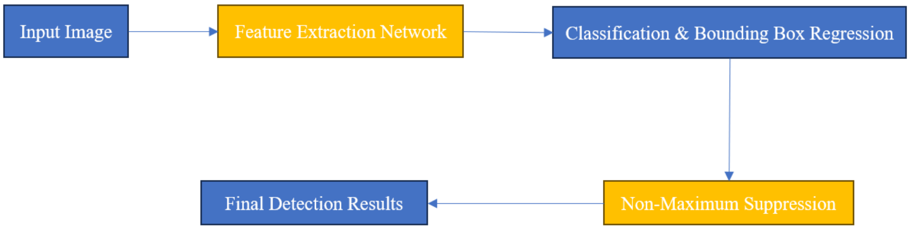
Input Image
Feature Extraction Network
Classification & Bounding Box Regression
Non-Maximum Suppression
Final Detection Results

Figure 3: Workflow diagram of one stage detectors (Picture credit: Original).

classify and fine-tune the bounding box of these regions in the second stage, which is based on the candidate regions created in the first stage. In order to extract data from each candidate region and identify the target class and its specific location within it, this stage usually employs a classification network-like design. Dual-stage detectors have high detection accuracy, especially when dealing with complex scenes or small objects.

The separation design of the two-stage makes the model more robust to complex backgrounds and dense targets. The two stage detectors' classical model is called Faster R-CNN. It introduces RPN, which allows region suggestion and detection to share the feature extraction part, dramatically increasing the detection speed. Although faster R-CNN strikes a better balance between speed and accuracy, it is still slower than one stage detectors.

2.2.3 Introduction of One Stage Detectors

The basic idea behind one stage detectors is to complete all object detection stages in a single forward propagation, eliminating the requirement for an area suggestion stage and accurately predicting the class and location of the object directly from the image. The workflow of one stage detectors mainly consists of feature extraction, detection and classification. Workflow diagram of one stage detectors can be referred to Figure 3. Convolutional neural networks are required for feature extraction in order to extract the image's features. The model must produce numerous anchor frames at each location on the feature map, each with a distinct size and aspect ratio, in order to perform detection and classification. The model predicts whether or not these anchor frames contain targets and performs classification and positional regression on them. The network structure of the one stage detectors is simple and easy to train and deploy.

And since one stage detectors require only one forward propagation, it is fast. YOLO and SSD are two devices that are exemplary of one stage detectors. The YOLO model makes predictions directly on various image grids, each of which projects a predetermined number of bounding boxes and category probabilities. SSD uses multi-scale feature maps for detection. As a result, it can process objects of different sizes more efficiently by generating anchor frames of different sizes from feature maps of different scales.

3 RESULT AND DISCUSSION

3.1 Result Analysis

As shown in Table 2, the table summarizes the two stage detectors of the RCNN series, as well as YOLO and SSD, which belong to the one stage detectors. RCNN uses SS as the RP, with a fixed input image, and an accuracy of 58.5% for VOC07 and 53.3% for VOC12, which is very slow (<0.1 FPS).Fast RCNN also uses SS, but supports an arbitrary size of the input images, the accuracy improves to 70.0%

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 2: An overview of the characteristics and functionality of detection frameworks for general object detection (Liu et al., 2020).

Name of detector	RP	Input ImgSize	VOC2007 outcomes	VOC2012 outcomes	Speed (FPS)
Two stages					
RCNN (Liu et al., 2020)	SS	Fixed	58.5(07)	53.3(12)	< 0.1
Fast RCNN (Liu et al., 2020)	SS	Random	70.1(VGG) (07+12)	68.5(VGG) (07++12)	< 1
Faster RCNN (Liu et al., 2020)	RPN	Random	73.3(VGG) (07+12)	70.3(VGG) (07++12)	< 5
One stage					
YOLO (Liu et al., 2020)	-	Fixed	66.4(07+12)	57.9(07++12)	< 25(VGG)

SSD (Liu et al., 2020)	-	Fixed	76.8(07+12)
			81.5(07+12+CO)
				74.9(07++12)
				80.0(07++12+CO)
					< 60

(VOC07) and 68.4% (VOC12), and is slightly faster (<1 FPS).Faster RCNN introduces RPN, and the accuracy further improves to 73.2% (VOC07) and 70.4% (VOC12), and the speed also improves (<5 FPS).YOLO doesn't use RP, and directly performs the detection on the fixed-size images, which is substantially faster (<25 FPS) but with slightly lower accuracy of 66.4% (VOC07) and 57.9% (VOC12).SSD is similar to YOLO, being the fastest (<60 FPS) and with accuracy of 76.8% and 74.9% on VOC07 and VOC12, respectively. The accuracy can be further improved to 81.5% and 80.0% by including the COCO dataset. The two stage detectors and the one stage detectors differ significantly in terms of speed and accuracy. The main reason for this difference is that the two stage detectors requires separate processing of the candidate region. RCNN series of algorithms are slower but more accurate, and are suitable for tasks requiring high accuracy. Because of their speed, YOLO and SSD can complete real-time detection jobs and attain a greater speed-accuracy ratio.

3.2 Discussion

Despite their reputation for great accuracy, two-stage detectors are typically slower. In contrast, one-stage detectors provide a noticeable speed benefit. For instance, because Faster R-CNN uses a Region Proposal Network (RPN) to produce candidate regions, it performs better than both R-CNN and Fast R-CNN. However, it is still not as fast as YOLO or SSD, as two-stage detectors must process each candidate region individually. Among one-stage detectors, SSD is more adaptable to complex scenes than YOLO, largely due to its use of a feature pyramid network, which improves detection accuracy across different object scales. Most current object detection methods involve a trade-off between accuracy and speed. Future research could focus on developing algorithms that enhance speed without compromising accuracy, such as optimized versions of YOLO or Faster R-CNN. This balance is especially critical in applications like autonomous driving, where both high accuracy and fast processing are required. Increasing the speed of object detection models without compromising performance can be achieved through the use of quantization techniques, hardware acceleration (such as Tensor Processing Unit (TPU) and Graphic Processing Unit (GPU)), and better network architectures.

4 CONCLUSIONS

This study offers a thorough examination of the background and core algorithmic models in object detection, with a particular emphasis on comparing two-stage and one- stage detectors. The analysis shows the trade-offs between various methods: two-stage detectors, such Faster R-CNN, perform slower overall but are able to attain higher accuracy thanks to their region proposal processes. In contrast, speed-focused one-stage detectors like SSD and YOLO are more effective but less precise.

These findings are supported by many trials, which demonstrate that although two-stage detectors offer more accuracy, they are slower than one-stage detectors. Bridging the gap between high accuracy and fast speed will be a crucial focus for future research. In the future, network topologies will be optimized in an effort to better balance efficiency and precision. Additionally, advancements in hardware, such as GPUs and TPUs, and techniques like quantization will be explored to enhance detection speed without sacrificing performance. Efforts will also be directed towards improving the robustness of detectors in diverse real-world conditions, including varying lighting, occlusions, and cluttered environments. The project intends to increase object identification skills by merging these advancements and making them acceptable for increasingly demanding and real-time applications.

REFERENCES

Carranza-García, M., Torres-Mateo, J., Lara-Benítez, P., & García-Gutiérrez, J., 2020. On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data. Remote Sensing, 13(1), 89.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., & Dietmayer, K., 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.

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

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M., 2020. Deep learning for generic object detection: A survey. International journal of computer vision, 128, 261-318.

Pathak, A.R., Pandey, M., & Rautaray, S., 2018. Application of deep learning for object detection. Procedia computer science, 132, 1706-1717.

Razi, A., Chen, X., Li, H., Wang, H., Russo, B., Chen, Y., & Yu, H., 2023. Deep learning serves traffic safety analysis: A forward ‐ looking review. IET Intelligent Transport Systems, 17(1), 22-71.

Tian, Z., Shen, C., Chen, H., & He, T., 2020. FCOS: A simple and strong anchor-free object detector. IEEE transactions on pattern analysis and machine intelligence, 44(4), 1922-1933.

Yang, R., & Yu, Y., 2021. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Frontiers in oncology, 11, 638182.

Zhang, Y., Li, X., Wang, F., Wei, B., & Li, L., 2021. A comprehensive review of one-stage networks for object detection. In 2021 IEEE International Conference on Signal Processing, Communications and Computing, 1-6.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X., 2019. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. 2023. Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS