

Application of Deep Learning in Drone Image Detection

Xiang Cheng^{1,*}, Boning Lu² and Shaokang Qu³

¹Guicheng Middle School, Foshan, China

²Xiamen Yingcai School, Xiamen, China

³Jinan Xinhang Experimental Foreign Language School, Jinan, China

*Corresponding author: b20160601413@stu.ccsu.edu.cn

Abstract. As an important research direction in the field of computer vision, image detection is a technology that can localize and identify the image and can accurately discriminate the features of the task image, and its application to UAVs has been a hot topic in recent years. The field of algorithms of deep learning applied to image detection is progressing rapidly, however, there are still many advantages and disadvantages between various algorithms, and the application of UAV image detection focuses on different. Therefore, the research topic of this article is to compare the advantages and disadvantages of various algorithms of image detection and conduct a technical analysis of improved algorithms applied to drones. The research method of this paper is as follows: firstly, introducing the model of each algorithm, analyzing the principle, and then comparing the advantages and disadvantages of detection speed, detection accuracy, and finally choosing improved algorithms that are suitable for drone Image detection. the improved algorithms suitable for UAV image detection are selected. In terms of accuracy, Faster R-CNN and YOLO perform better compared to R-CNN and Fast R-CNN. In terms of speed, YOLO has a higher processing speed and is suitable for application scenarios that require high real-time performance. Faster R-CNN is suitable for scenarios that require high accuracy but relatively low-speed requirements. SSD performs outstandingly in handling multi-scale targets. Therefore, in the field of UAV image detection, choosing the appropriate deep learning model can balance the needs of accuracy and speed according to the application scenarios.

Keywords: Deep learning; algorithm model; drone; image detection.

1. Introduction

Image detection is a technology that can self-locate or recognize objects in videos or photos. Its significant feature is the ability to automatically proofread task images for feature extraction and make accurate judgments. It is the most important direction of research in the field of computer vision nowadays. In the 1950s, this technology started to enter the research stage, but the main objects of recognition were numbers and characters. Entering the 1980s to 1990s, Yann LeCun and his team developed a convolutional neural network (CNN) [1]. The emergence of convolutional neural networks has creatively promoted the development of computer vision technology and created a solid foundation for the development of image recognition in the future. After entering the 21st century, the development of image detection encounters a new turning point. In 2012, AlexNet was developed by Geoffrey Hinton and Alex Krizhevsky. Moreover, they acquired a good score in the ImageNet competition [2]. AlexNet not only makes people attach greater importance to CNN but also accelerates the speed of the research process to deep learning and provides a new technology to support the development of image detection in the future. In 2015, Kaiming He's team developed ResNet which has become a strong tool in the field of computer vision and can build a very deep neural network and help the accuracy of image detection to close to 90% [3]. After this period, the development and application of object detection models such as the series of YOLO and RetinaNet not only optimized performance detection, but also created a significant breakthrough for image detection in accuracy and efficiency [4][5].

As the algorithm model of deep learning becomes more diverse and mature, its application in every field becomes wider, especially in the field of drone detection. On the one hand, the algorithm model

of deep learning refreshed the knowledge about drones to people, so drones are no longer an investigative tool for the military but have become a strong force for collecting, testing, and analyzing data. On the other hand, the function of drone image detection has entered into humans' lives through the assistance of the algorithm model of deep learning. In agriculture, it helps agricultural workers to analyze, calculate, and predict the growth pattern of crops. In the field of environmental protection, it helps environmental inspectors test the extent of pollution and predict natural disasters. In the field of security, it can not only collect information but also monitor the real-time situation of traffic in the cities.

This article will provide a detailed introduction and analysis algorithm model of deep learning related to drone image detection, including YOLO, SSD, and R-CNN series (R-CNN, Fast R-CNN, Faster R-CNN). Then summarize their working principle and benefits of performance in the testing process. In addition, during the research process, the features and performance of the above models of algorithm were comparatively analyzed, including accuracy, running speed, and other indicators, to comprehensively and objectively evaluate their performance in image detection tasks. This means that provides a deep analysis for the application of deep learning in drone image detection.

2. The Features of Imaging of Unmanned Aerial Vehicle (UAV) Technology

There are some main features of drone aerial images below. Imaging of unmanned aerial vehicle technology has features that are high-altitude viewing angles, providing a broad field of vision and comprehensive information [6]. This technology needs multi-scale target detection and recognition abilities due to different scales and small targets. Affected by the aerial viewing angle, the targets may become different in the image. It needs to consider the impact of target detection from an aerial viewing angle. Moreover, dealing with the numerous data is a challenge, imaging of unmanned aerial vehicle technology needs to deal with and store these data efficiently. These features made a challenge for algorithms of target detection because the algorithms have to overcome targets of different scales, recognition of small targets, the impact of viewing angles of aerial photography, and operation of numerous data.

Some features of imaging of UAV technology are high-altitude viewing angles, multi-scale, small targets, affected by viewing angles of aerial photography, and a large amount of data. These features provide challenges for the design and optimization of target detection algorithms for UAV.

3. Analysis of the Algorithm Model

Due to the rise of convolutional neural networks, algorithms of target detection based on deep learning have gradually replaced conventional algorithms of target detection due to their high detection performance and have become the mainstream algorithm in the current target detection application. These algorithms of target detection based on deep learning are divided into two types: two-stage algorithm and one-stage algorithm.

There are common one-stage algorithms, including YOLO, SSD, RetinaNet, and SqueezeDet. The one-stage algorithm does not need to go through the region proposal stage and directly generates the category probability and position coordinate values of the object. After a single detection, the final detection result can be obtained directly. It is mainly based on anchors to directly classify and adjust the bounding box, so the one-stage algorithm has a faster detection speed.

There are common two-stage algorithms, including R-CNN, SPPNet, Fast-R-CNN, faster-R-CNN and FPN, R-FCN. The features of the two-stage algorithm are to use specialized modules to create candidate frames, find foregrounds and adjust bounding boxes, and further classify and adjust bounding boxes based on previously generated candidate frames. The two-stage algorithm does fine-grained object detection, it is more precise but the speed is lower than the one-stage algorithms.

This paper will focus on YOLO, SDD, and R-CNN algorithms. They all belong to algorithms of Anchor-based object detection and depend on predefined anchor boxes also are usually a group of

rectangular boxes at different scales and aspect ratios in the image and are used to advise and predict the targets. YOLO and SSD divide the image into $S \times S$ small grids, and each grid is responsible for detecting targets in the image. For each grid, a specific number of bounding boxes, the confidence and class probability of each bounding box are predicted. Faster R-CNN introduced the Region Proposal Network (RPN), which provided a potential object area by using anchor boxes. These anchor boxes classify and regress based on the extent of overlap between them and real targets in images.

Due to the development of drone technology, Combining the algorithm of UAV target detection with its application has become important research content in the application. The algorithms of UAV object detection based on deep learning have become a hot point in current research, its main benefit is improving the accuracy and efficiency of detection.

Application in the algorithm of UAV target detection is wide, such as disaster rescue, police investigation, agricultural development, traffic monitoring, and other fields. Therefore, there is a significant value in research and application in combining the algorithm of UAV target detection with its application.

4. YOLO

YOLO is a single-stage algorithm of object detection that considers the object detection task as a regression problem and directly predicts the location and category of the object on the entire image [4]. YOLO divides the image into a fixed number of grids, every grid is responsible for detecting the objects on the images. Furthermore, each grid prediction includes the probability of the target, the bounding box coordinates of the target (center point coordinates, width, and height), and the category of the target.

YOLO achieves the detection and category of objects through a single forward propagation, giving a remarkable benefit for its running speed. This means the entire image only needs to operate the network once to obtain all the results of object detection. Due to YOLO considering the entire image at the same time, it has better global context awareness ability. It makes the detection of small objects and the processing of many objects more accurate. YOLO uses a multi-task loss function to comprehensively consider the regression error of the target location, the classification error of whether the target exists, and the classification error of the target category. This helps balance the weight of different tasks in object detection.

The algorithm of YOLO uses CNN to extract features and also uses fully connected layers to get predictive value. It involves 24 convolutional layers and 2 fully connected layers. The convolutional layer mainly uses 1×1 convolution for channel reduction and 3×3 convolution to extract features. The convolutional layer and the fully connected layer use the Leaky ReLU activation function, and the last layer uses the linear activation function.

Compared with using region proposal generates candidate regions, it may generate some situations about missed detection or false detection due to YOLO can directly detect an entire image.

Improvement of different versions of the YOLO series focuses on network structure, prediction technology, model performance, and efficiency, improving the performance and applicability of algorithms of object detection continuously.

5. SSD

SSD (Single Shot Multibox Detector) is an algorithm of deep learning that is used in detecting objects on images[7]. Its design goal is to complete object detection and position positioning through a single neural network at the same time, thereby making efficient real-time object detection. SSD algorithm' accuracy and speed are greater than the original YOLO algorithm. Based on YOLO, there are three main improvements: multi-scale feature images, using convolution for detection, and setting a priori boxes. SSD uses CNN to detect directly, instead of detecting after the fully connected layer like YOLO; SSD extracts feature images of different scales for detection. Large-scale feature images

can be used to detect small objects, while small-scale feature images used to detect large objects. SSD uses a priori boxes of different scales and aspect ratios.

The detection value of SSD is different from that of YOLO. For each, a priori box of each unit, it outputs a set of independent detection values, corresponding to a bounding box, which is mainly divided into two parts:

SSD also treats background as a special category, so the first confidence level refers to the score that does not contain the target or belongs to the background.

The four values of the bounding box (C_x , C_y , w , h) are only the transformation values of the bounding box relative to the prior box, the process of reverse representation is required, that is, decoding is performed to obtain the true position of the bounding box from the predicted value.

6. R-CNN

R-CNN (Region-based Convolutional Neural Networks) is an object detection algorithm. Its core idea is to first extract regions from the image, then perform feature extraction and classification using Convolutional Neural Networks (CNN) for each extracted region [8].

R-CNN is one of the earliest and most representative deep learning-based algorithms in the field of object detection. Compared to traditional handcrafted features, R-CNN introduces a convolutional neural network for extracting deep features, followed by a classifier to determine whether the search region contains the target and its confidence, which can achieve more accurate detection results.

The process of R-CNN involves first identifying the Region Proposal and then performing classification and regression. R-CNN uses a Selective Search for the region proposal.

Selective Search is only responsible for rapidly generating regions that may contain objects, without performing specific detection. It uses a variety of prior knowledge to make simple distinctions among different regions, avoiding some useless searches, and improving speed and accuracy.

The advantage of this method is that it merges similar image regions, reducing the required computational workload. Furthermore, due to the use of the subregion merging method, it can encompass suspected object boxes of various sizes, making it more suitable for detecting targets at multiple scales, and also increasing the probability of detecting objects.

The R-CNN series has multiple advantages. First, it has achieved high detection accuracy in object detection tasks, especially Faster R-CNN, which has reached a leading level on various standard datasets. Secondly, by introducing the Region of Interest (RoI) pooling layer, the R-CNN series can accurately locate the target objects, thereby improving the localization accuracy of object detection [9]. In addition, the R-CNN series is also suitable for multi-class object detection tasks and demonstrates excellent performance in target classification. Due to the adoption of the CNN structure, the R-CNN series also demonstrates good performance in object classification tasks. These advantages have made the R-CNN series one of the important algorithms in the field of object detection.

The R-CNN series has a major drawback of being slow, especially during the testing phase. This is because it adopts a Two-stage process, which includes extracting candidate regions and classifying these regions. The training process of the R-CNN series is relatively complex, involving multiple stages, including generating candidate regions, performing RoI pooling, conducting object classification, and position regression steps.

Due to its slow speed, the R-CNN series may not meet the requirements in real-time application scenarios, such as autonomous driving and real-time video analysis. It also requires significant computational and memory resources, making it less suitable for resource-constrained environments, such as embedded systems and mobile devices. RoI pooling may introduce some localization errors during region extraction, which to some extent affects the accuracy of target localization. Overall, the R-CNN series has made significant achievements in the field of object detection. However, its speed and computational resource requirements have limited its widespread application in real-time and resource-constrained environments. In R-CNN, the repetitive convolution of the same region by the

CNN leads to a large storage space requirement. In R-CNN, the actual target boxes selected by selective search have various sizes, and scaling the images to the same size may cause the targets to deform.

With the introduction of subsequent algorithms such as SSD and YOLO, some issues have been alleviated, making them more suitable for specific scenario requirements

7. Improved Algorithms for Drone Imagery

In recent years, many object detection neural network methods have been proposed, especially One-stage object detection methods. In contrast to two-stage methods, single-stage methods combine object localization and classification into a single step, enabling real-time object detection on desktop and embedded hardware. In these works, the YOLO series has attracted much attention. In 2023, an algorithm for multi-scale unmanned aerial vehicle image object detection based on the YOLOv8 model was proposed, named Drone-YOLO [10]. Drone-YOLO aims to overcome specific challenges related to UAV image object detection, providing an efficient solution for object detection in drone applications.

To address the problems of large-scene sizes and small detected objects, Drone-YOLO, compared to the YOLOv8 model, has improved its neck component by adopting a three-layer PAFPN structure. It has also combined a detection head that uses large-scale feature maps to tailor small objects, significantly enhancing the algorithm's ability to detect small objects. The architecture of the Drone-YOLO network model is an improvement based on the YOLOv8-l model. The main body of the network uses reparameterized convolutional modules based on the RepVGG structure as the down-sampling layers. During the training process, this convolutional structure simultaneously trains 3x3 and 1x1 convolutions. During the inference process, the two convolutional kernels are merged into a single 3x3 convolutional layer. This mechanism allows the network to learn more robust features without affecting the inference speed or increasing the model size. The neck of Drone-YOLO extends the PAFPN structure to three layers and attaches a small-sized object detection head. The combination of the proposed sandwich-fusion extracts spatial and channel features from three different layers of the network backbone. This optimization enhances the multi-scale detection head's ability to gather spatial localization information for the objects to be detected.

Drone-YOLO has been evaluated in ablation experiments and compared with other state-of-the-art methods on the VisDrone2019 dataset. The results show that Drone-YOLO outperforms other baseline methods in terms of object detection accuracy. The algorithm has demonstrated superior performance in detecting and tracking small objects, particularly in the context of UAV applications. Compared to YOLOv8, Drone-YOLO has achieved a significant improvement in the mAP_{0.5} metric, with an increase of 13.4% in the VisDrone2019 test and 17.40% in the VisDrone2019-val. The experiments have verified the effectiveness of the Drone-YOLO method in the task of object detection in UAV images.

8. Conclusion

In the field of unmanned aerial vehicle image detection, a comparative analysis of the performance of deep learning models, including the R-CNN series, SSD, and YOLO, on standard image detection datasets, revealed that in terms of accuracy, Faster R-CNN and YOLO outperform R-CNN and Fast R-CNN. This indicates that both the two-stage approach, which involves introducing a Region Proposal Network, and the single-stage approach, which employs a single forward pass, have advantages in improving accuracy.

In terms of speed, YOLO has a higher processing speed due to its design of a single forward pass, making it suitable for applications that require real-time performance. Due to its two-stage design, Faster R-CNN is relatively slower, making it suitable for scenarios that require relatively lower processing speeds but higher accuracy. SSD performs well in handling multi-scale objects, and its

unique multi-scale feature fusion mechanism gives it an advantage in object detection tasks at different scales. Therefore, in the field of unmanned aerial vehicle image detection, choosing the appropriate deep learning model can balance the requirements of accuracy and speed according to different application scenarios.

It is worth noting that the object detection algorithms introduced above are all based on the Anchor-based method, which relies on predefined anchor boxes. Drones may encounter targets with significant scale variations in different scenarios and require lightweight models and real-time performance. In such cases, Anchor-free methods may be more suitable for drone target detection applications. Therefore, for drone target detection tasks, Anchor-free object detection algorithms such as CornerNet, CenterNet, FCOS, and FoveaBox could be considered. The research in this article is not comprehensive, and future expansion of more algorithm research in the field of drone applications is needed to continuously optimize and promote the development of this field.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] LeCun Y. Boser B. E. Denker J. S. Henderson D. Howard R. Hubbard W. & Jackel L. D. (1989). Backpropagation applied to handwritten Zip code recognition. *Neural Computation*, 1(4): 541–551.
- [2] Krizhevsky A. Sutskever I. & Hinton G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, 60(6): 84–90.
- [3] He K. Deep residual learning for image recognition, 2016.
- [4] Redmon J. Divvala S. et al. You Only Look Once: Unified, real-time object detection computer vision & pattern recognition. *IEEE*, 2016.
- [5] Lin T. Focal loss for dense object detection, 2017.
- [6] Chen Liu. Shixin Li. Fanrun Meng. & Fankai Chen. Survey of UAV target detection based on deep learning. *Computer Science and Application*, 2023, 13: 1092.
- [7] Liu W. Anguelov D. Erhan D. Szegedy C. Reed S. Fu C. & Berg A. C. SSD: Single shot multibox detector. in *lecture notes in computer science*, 2016: 21–37.
- [8] Girshick R. Donahue J. Darrell T. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. *IEEE Computer Society*, 2014.
- [9] Girshick R. Fast R-CNN[J]. *Computer Science*, 2015.
- [10] Zhang, Z. Drone-YOLO: An efficient neural network method for target detection in drone images. *Drones*, 2023, 7(8): 526.