Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241325

A comparison of deep learning-based object detection for unmanned aerial vehicle

Tinghuan Li

Faculty of Information Management,Shanghai Lixin Accounting and Finance University, Shanghai, China

lth0519@outlook.com

Abstract. Unmanned aerial vehicle (UAV) plays a critical role in the field of object detection, and machine learning techniques have significantly advanced the field in recent years. This paper provides a comprehensive overview of machine learning developments for UAV object detection. The process involves multiple steps. Firstly, deep learning and Convolutional Neural Network (CNN) are widely utilized to extract precise features and train object detection models for accurate classification and efficient recognition of target objects in images and videos. Secondly, classical object detection algorithms such as Faster R-CNN, You Only Look Once (YOLO), and Single Shot MultiBox Detector (SSD), have been enhanced to improve accuracy and real-time performance. In this review, we primarily focus on comparing the principles of CNN and YOLOv5 themselves, as well as their applications in object detection and image recognition. Ultimately, it becomes apparent that CNN is better suited for processing image data, automatically extracting features, and achieving more accurate classification and detection. On the other hand, YOLOv5 directly performs detection on large images, significantly reducing computation time compared to CNN.

Keywords: Unmanned aerial vehicle, object detection, Deep Learning, Convolutional Neural Network (CNN), YOLOv5.

1. Introduction

With the development of unmanned aerial vehicle (UAV) technology and the further increase in UAV market rate, a large number of convenient and functional UAV is used in daily life. UAV, with the pros of high mobility, quick deployment, and wide surveillance range, has developed into an excellent supplement to conventional remote sensing technology with the continued development of high-altitude photography and remote sensing mapping technology [1]. UAV is also of great significance in motion analysis, disaster relief, and security. Their inherent benefits in information monitoring and acquisition make them invaluable tools. In the domain of traffic control, UAV has the ability to capture crucial traffic metrics and high-resolution videos from an aerial perspective, which are not attainable through conventional monitoring techniques. Furthermore, their adaptability to diverse data collection requirements enables continuous monitoring of specific areas, greatly enhancing the efficiency of related tasks [2].

Object detection is an essential technique in the field of computer vision, and its main task is to automatically localize and identify objects in images [1]. The integration of object detection technology

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/47/20241325

on UAV has the potential to greatly enhance their perception capabilities, thereby playing a pivotal role in the advancement and widespread adoption of UAV technology. Considering the challenges posed by the low accuracy of detecting small targets and the slow speed of large-scale object recognition, this paper aims to present an analysis of the implementation process using the Convolutional Neural Network (CNN) and YOLOv5 as exemplars. By summarizing the strengths and limitations of these two approaches, it is anticipated that this research will serve as a valuable reference for future related work.

2. Object detection

Object detection refers to the process of identifying the presence of a specific object in an image and providing information on its category and location. This technique falls within the realm of computerized image recognition and is a subfield of machine learning. It holds significant importance in the domains of deep learning and CNN.

2.1. Applications of UAV target recognition

In recent years, UAV perspective target detection has emerged as a prominent research area in the field of computer vision. As a result, numerous techniques have been extensively employed to develop intelligent solutions for building smart cities. For example, in the case of a pandemic which may spread more quickly in large crowds. UAV that can identify pedestrians can estimate crowd density to address this issue and assist avert pandemics. As another example, according to the new traffic rules, cyclists are required to wear helmets while riding e-bikes. To streamline the detection process via UAV, the integration of e-bike detection technology with helmet detection technology can be considered [3]. In addition, UAV may capture a brief video of the infraction and transmit it to a base station nearby to offer information to the authorities so they can take legal action [4].

2.2. Machine Learning Based Applications in Object Detection

Manual object detection often demands significant manpower and time, and is invariably prone to errors. However, by representing different object features through imagery, and employing machine learning models for training and prediction, the efficiency and precision of the task can be substantially improved. At this time, both of them are greatly improved to help people in decision-making. For example, J.W. James [5] et al. proposed an approach to recognize the shape and sliding contour of an object based on Support Vector Machine (SVM) using TacTip bionic optical tactile sensor. However, in a study of the same bionic tactile sensor, it was found that the performance was good when the data in the task was similar to the training data. It was unreliable for small changes in the task data (e.g., sensor orientation). Lepora N F [6] et al. applied deep learning for the first time to optical bionic tactile sensors and for the first time to edge perception and contour tracking and found that the contour tracking strategy has a strong generalization to the task of the training model. The major advantage of utilizing CNN to address this problem lies in its capacity to generalize beyond the training data, exhibiting robust performance that surpasses previous experiences.

2.3. Machine Learning Based Applications in UAV's Object Detection

Similarly, target detection has been a long-standing concern in the UAV field. Traditional approaches for UAV image target detection primarily rely on statistical machine learning techniques, such as AdaBoost, SVM, and Random Forest. However, these approaches have several limitations. They exhibit high algorithmic complexity, as they need to process numerous sliding windows. Additionally, constructing reliable feature extractors manually presents difficulties. As a result, traditional target detection algorithms fail to meet the demands of practical applications. Moreover, traditional methods necessitate manual parameter setting and are limited to specific scenarios, lacking effectiveness in recognizing large-scale data. Therefore, there is a need for more advanced and efficient techniques to address these shortcomings and enhance UAV target detection capabilities.

In the past few years, deep learning techniques have revolutionized many aspects of computer vision, making significant advances in robustness and detection accuracy, and have begun to be widely used in

DOI: 10.54254/2755-2721/47/20241325

image classification and detection. CNN achieved great success in the image classification competition in 2012.In 2014, Girshick [7] proposed R-CNN, and CNN were applied for the first time to target detection. Based on this, many R-CNN-based target detection techniques have been proposed, including fast R-CNN [8], faster R-CNN [9], and R-FCN[10]. Redmon [11] created the first real-time CNN-based target detection model, YOLO, which treats the target detection as a regression problem in order to improve the effectiveness of target recognition. However, YOLO has low detection accuracy for dense and multi-scale objects due to the fact that only the latest feature mapping is used for detection. In 2016, Liu [12] developed the SSD algorithm that combines the advantages of YOLO and Faster R-CNN. SSD operates at the speed of YOLO while maintaining the accuracy of Faster R-CNN. In 2018, Redmon [13] proposed the yolov3 algorithm, which draws on the idea of SSD to predict different feature mappings, introducing residual layers and jump connections to improve the detection performance.

3. Algorithms

3.1. CNN

CNN is a classical deep learning model that is widely used in image recognition tasks. CNN extracts local features of an image by utilizing components such as convolutional, pooling, and fully connected layers, and gradually abstracts and integrates the features, ultimately outputting a prediction of the object category [14][15].

In the case of further optimization of the activation function, the neural network structure often chooses the Rectified Linear Unit (ReLU) function rather than the Sigmoid or tanh functions. Its function expressed at eq(1):

$$f(x) = \max(0, x) \tag{1}$$

In neural networks, the ReLU function serves as the activation function of the neuron for the nonlinear output result of the neuron after a linear transformation. In other words, for the input vector x from the previous layer of the neural network that enters the neuron, the neuron using the ReLU function outputs at eq (2):

$$\max(0, w^{\mathrm{T}} x + b) \tag{2}$$

to the next layer of neurons or as an output of the whole neural network which depends on where the present neuron is located in the network structure.

3.1.1. Input Layer. The input layer is the first and initial layer of the neural network. Its role is to perform the initial sampling of information, and this sampling is done by convolution. In the field of computer image processing, for color images, three channels of RGB (Red, Green, Blue) are generally used as input, while grayscale image modes are single channel images.

3.1.2. Hidden Layer. Since the input and output layers are visual image representations, while the middle layers constitute a network structure with black-box properties, they are referred to as hidden layers. The hidden layers primarily consist of convolution, pooling, and density. The primary role of these hidden layers is to further extract image features and learn more comprehensive feature structures by performing operations such as convolutional filtering and pooling.

a. Convolution

The function of the convolutional layer is to perform feature extraction on the input image. Convolutional layer performs convolutional operations on images and other 2D data. There is an important parameter in the convolutional layer, the convolution kernel. It uses a set of learnable filters (also known as convolution kernels) to perform convolution operations with the input to extract different features from the input. The convolutional layer is effective in capturing spatial correlations.

DOI: 10.54254/2755-2721/47/20241325

b. Pooling

The main purpose of a pooling layer is feature selection and information filtering. It is used to reduce the spatial dimensions of the neural network output while retaining important feature information. The pooling operation follows a similar step as the convolutional kernel, determined by the pooling size, stride, and padding. Pooling approaches primarily include LP pooling, random/mixed pooling, and spectral pooling. The most common pooling operation is max pooling, which extracts the most significant feature values from the input.

c. Dense

A dense layer is one of the most common neural network layers, where each neuron is connected to all the neurons in the previous layer. It multiplies the input vector with a set of weights and applies an activation function to generate the output. Its function is to nonlinearly combine the extracted features to obtain the output, i.e., the fully connected layer is not expected to have feature extraction capability itself but rather tries to utilize the existing higher-order features to accomplish the learning objective. This layer helps the neural network make predictions and assign input data to specific categories or classes based on the learned weights and biases.

3.2. YOLOv5

3.2.1. Composition of YOLOv5. YOLOv5 consists of four main parts which include Input, Backbone network, Neck network and Output. YOLOv5s is the most basic network among them which is the network with the smallest depth and the smallest width of the feature map in the Yolov5 family. In addition, based on YOLOv5s, YOLOv5 changes the width and depth of the model by adjusting the parameters, and according to its size, from small to large, it is divided into YOLOv5m, YOLOv5l, and YOLOv5x[16].

3.2.2. The detection process of YOLOv5. Input: Firstly, the Mosaic data enhancement algorithm is used to select four images from the dataset, randomly scale and crop them, and then combine them into a single image according to a certain ratio, which is beneficial for the model to recognize the target in a smaller range, thus increasing the complexity of the dataset and enhancing the robustness of the model. Secondly, during the training process, adaptive anchor frame computation and adaptive image scaling techniques are utilized for different datasets to calculate the anchors. The network then makes predictions based on these anchors, generating prediction frames that are subsequently compared with the labeled frames. Finally, the parameters within the network are iteratively updated through the back propagation of the gradient, optimizing the model's performance.

Backbone: The Focus module is employed to perform a slicing operation on the image. The resulting sliced image is then convolved with the selected operation, resulting in a new image. Finally, a two-fold downsampled feature map is obtained, preserving all relevant information without any loss.

Neck: Built on the basis of Feature Pyramid Networks (FPN) and Pyramid Attention Network (PAN), the FPN conveys strong semantic features from top to down, while the PAN conveys localization features from bottom to up, and also uses Cross Stage Partial Network(CSPNet) to fuse information from higher layers and the backbone network, which further improves the feature information of the network and enhances the feature fusion capability of the network.

Output: It processes target detections of different sizes and uses Non-Maximum Supression to remove redundancy and duplicate prediction frames. It retains outputs with the highest confidence level [17].

3.3. Comparison

Apparently, compared with other algorithms, CNN has better robustness and generalization ability to adapt to different datasets and tasks. CNN can obtain richer image information through multi-level feature extraction, and has higher accuracy in tasks such as object detection and segmentation, so it is easier to recognize small objects. However, CNN requires a lot of computational resources and time,

DOI: 10.54254/2755-2721/47/20241325

and the training and testing process is time-consuming, especially on large-scale datasets. On the other hand, YOLOv5 adopts a direct detection approach on large-scale images, which greatly reduces the computational complexity and processing time compared to the chunking method of CNN. Similarly, the ability of YOLOv5 to handle multi-category target detection tasks needs to be improved. Although the model complexity YOLOv5 of and computational effort are low, this limits its ability to generalize to different datasets and tasks, as shown in Table 1.

Table 1. Comparison of CNN and YOLOv5

Names	CNN	YOLOv5
Principle	Processing data with spatial structure; Extracting features of input data through multiple convolutional and pooling layers; Performing tasks such as classification or regression through fully connected layers	Real-time target detection model; Employing a single-stage detector-based approach; Transforming the object recognition task into a bounding box regression problem
Advantages	High recognition accuracy, High robustness, Much more application scenarios	Simple, Efficient, Fast detection
Disadvantages	Complex, Slow detection	Poor detection accuracy for multi-scale objects and dense objects
Application Scenario	Suitable for large-scale datasets, Better at recognizing small objects	Real-time target detection

4. Algorithms analysis

In this study, CNN and YOLOv5 were employed for verification, utilizing the same UAV-captured image depicted in Figure 1, with the primary emphasis on identifying a car within the image.

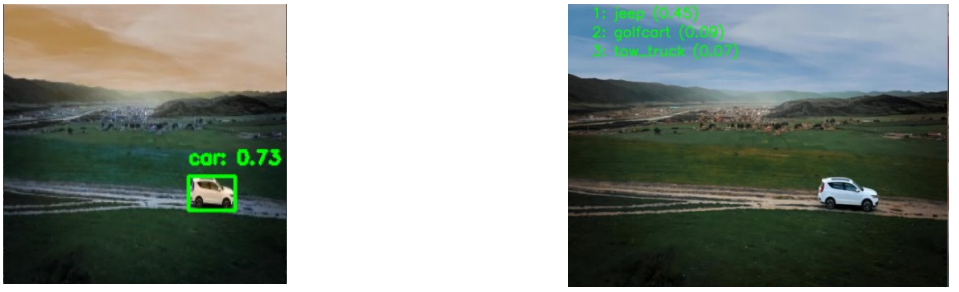


Figure 1. The original UAV-viewed image.

First, we resize the image, adjusting the original image from 1280×720 to 640×640 to enhance the object detection model's ability to recognize the car.

Among those YOLOv5 models, this paper opts for the YOLOv5s detection model due to its lower complexity to ensure higher detection accuracy while occupying smaller memory occupation and faster recognition speed, thus meeting the requirements for UAV object detection. The detected result of YOLOv5s on this UAV-viewed image is below as shown in Figure2(a):

DOI: 10.54254/2755-2721/47/20241325



(a) (b)

Figure 2. Detected results of different methods on UAV-viewed picture. (a)Detected results of YOLOv5s; (b)Detected results of ResNet50.

In regard to CNN, this paper adopts ResNet50, which has deeper layers compared to traditional CNN models. Residual learning is introduced to address the gradient problem and enhance training performance. The detected result of YOLOv5s on the UAV-viewed picture is displayed above as presented in Figure 2(b).

After experiments, the object detection results of the two models for the same UAV-viewed image were compared, and the findings are presented in Table 2. The results clearly indicate that YOLOv5s outperforms CNN in terms of recognition accuracy and operational speed, demonstrating higher detection accuracy and faster processing. The lightweight network structure design and optimization of YOLOv5s contribute to its efficiency, speed, and high accuracy with excellent recall.

Table2. Comparison of the performances of the two models

	YOLOv5s	ResNet50
Detection Accuracy	73%	45%
Run Time(s)	0.73	2.66

5. Conclusion

Both CNN and YOLOv5 have been widely used in the field of object detection, but they differ in their application areas and approaches. CNN is commonly used in image classification tasks, leveraging local features to achieve accurate categorization. It can also serve as a fundamental component or sub-module in object detection. A common approach involves applying CNN to image regions at various locations and scales, using sliding windows or image pyramids to detect and classify target objects.On the other hand, YOLOv5 excels in real-time object detection. It is a faster algorithm that maintains high accuracy and real-time performance which can directly process an image in a single forward pass, generating multiple bounding boxes and category predictions simultaneously for real-time detection of targets. Efficient and accurate image recognition models are crucial in the development of UAV object detection capabilities. Future research should focus on further improving the model's speed, accuracy, applicable datasets, and other related issues.

References

[1] Yuanzhu Liu, Zhiming Ding, Yang Cao, and Mengmeng Chang. 2021. Multi-scale Feature Fusion UAV Image Object Detection Method Based on Dilated Convolution and Attention Mechanism. In Proceedings of the 2020 8th International Conference on Information Technology: IoT and Smart City (ICIT '20). Association for Computing Machinery, New York, NY, USA, 125–132. https://doi.org/10.1145/3446999.3447023

[2] Liu X, Zhang Z. A vision-based target detection, tracking, and positioning algorithm for unmanned aerial vehicle[J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-12.

DOI: 10.54254/2755-2721/47/20241325

[3] G. Yan, Q. Sun, J. Huang, and Y. Chen, ‘‘Helmet detection based on deep learning and random forest on UAV for power construction safety,’’ J. Adv.Comput. Intell. Intell. Informat., vol. 25, no. 1, pp. 40–49, Jan. 2021.

[4] N. A. Khan, N. Z. Jhanjhi, S. N. Brohi, R. S. A. Usmani, and A. Nayyar, ‘‘Smart traffic monitoring system using unmanned aerial vehicles (UAVs),’’ Comput. Commun., vol. 157, pp. 434–443, May 2020.

[5] J. W. James, N. Pestell and N. F. Lepora, "Slip Detection With a Biomimetic Tactile Sensor," in IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3340-3346, Oct. 2018, doi: 10.1109/LRA.2018.2852797.

[6] Lepora N F , Church A , Kerckhove C D , et al. From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor[J]. 2018.

[7] Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013.

[8] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[9] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 91–99.

[10] Dai, Jifeng et al.“R-FCN: Object Detection via Region-based Fully Convolutional Networks.” ArXiv abs/1605.06409 (2016): n. pag.

[11] Redmon, Joseph et al. “You Only Look Once: Unified, Real-Time Object Detection.”2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2016): 779-788.

[12] Liu, W. et al. “SSD: Single Shot MultiBox Detector.” ECCV (2016).

[13] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[14] Krizhevsky A , Sutskever I , Hinton G .ImageNet Classification with Deep Convolutional Neural Network[J].Advances in neural information processing systems, 2012, 25(2).DOI:10.1145/3065386.

[15] Simonyan K , Zisserman A .Very Deep Convolutional Networks for Large-Scale Image Recognition[J].Computer Science, 2014.DOI:10.48550/arXiv.1409.1556.

[16] Gao M T, Sun H, Tang Y Q, et al. Fingerprint Second-order Minutiae Detection Method Based on Improved YOLOv5[J]. Laser & Optoelectronics Progress, 2023, 60(10): 1010006.

[17] T. Bie, K. Fan and Y. Tang, "UAV Recognition and Tracking Method Based on YOLOv5," 2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA), Chengdu, China, 2022, pp. 1234-1239, doi: 10.1109/ICIEA54703.2022.10006200.