Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 7, No. 3, 2023

Improving the Yolov5 Detection Accuracy Through Enhancing The K-means Algorithm

Yuxi Liu[1, *]

[1]Arizona State University, Tempe, AZ 85281, United States
[*]Corresponding author's e-mail: Yliu866@asu.edu

Abstract: As technology advances, various cutting-edge innovations have brought greater convenience to human life. One such influential advancement is autonomous driving technology, which revolutionizes the automotive industry. By harnessing image recognition techniques, autonomous vehicles now possess unprecedented perception and environmental interpretation capabilities. However, over time, the proliferation of autonomous driving has given rise to several challenges. The increasing traffic volume, complex road conditions, intersecting pedestrian pathways, and an array of traffic signs have intensified the issues that autonomous driving technology must address. Consequently, a growing number of individuals have also become actively engaged in the pursuit of enhancing autonomous driving technology.

Keywords: Yolov5, Automatic driving, K-means model.

1. Introduction

The application of image recognition technology in autonomous driving is particularly prominent, especially with the breakthroughs achieved through Convolutional Neural Networks (CNN). By employing computer vision algorithms and deep learning models, autonomous vehicles can heavily rely on visual perception to analyze and comprehend road conditions, making informed decisions. [1]CNN technology efficiently identifies obstacles on the road, enabling the system to accurately discern specific objects such as road signs, fences, and buildings through model training, thus providing comprehensive perception capabilities to the vehicle.

The rapid development of deep learning and advancements in computer hardware have made traffic sign recognition an active research area. Deep learning has been widely applied in traffic-related studies due to its excellent feature extraction capabilities. Domestic researchers tend to favor Faster R-CNN and Cascade R-CNN, while foreign researchers lean towards using RetinaNet. Faster R-CNN introduced a novel method for traffic sign detection, providing new insights to the field despite its accuracy rate of 0.3449.

Cascade R-CNN adopts multi-scale attention and tackles sample imbalance, achieving high accuracy and recall rates on domestic traffic sign detection datasets. On the other hand, RetinaNet, based on the RetinaNet detector, achieves an accuracy rate of 96.7% and an inference speed of 5.07 frames per second in the domain of traffic sign recognition.[2]

The rapid advancement in deep learning and computer hardware has led to a significant shift towards single-camera vision-based object detection in the field of traffic sign recognition. Among the prominent representatives in this area is the You Only Look Once (YOLO) approach, with its most widely used version being YOLOv5. YOLO, short for "You Only Look Once," performs feature extraction, object classification, and regression within the same CNN.

YOLOv5 has gained popularity in autonomous driving due to its high accuracy and lightweight network architecture. By combining efficient feature extraction and object detection within a unified framework, YOLOv5 has emerged as a powerful tool for traffic sign recognition and plays a crucial role in enhancing the safety and performance of autonomous driving systems.

Although YOLOv5 is currently one of the most popular algorithms, it faces challenges in detecting road signs and vehicles accurately due to the complexity of modern societal road conditions. This results in detection omissions and errors. To address the issue of scanning traffic signs, we propose modifying the K-means++ algorithm to shorten bounding boxes, thereby allowing YOLOv5 to focus on the images within these boxes.

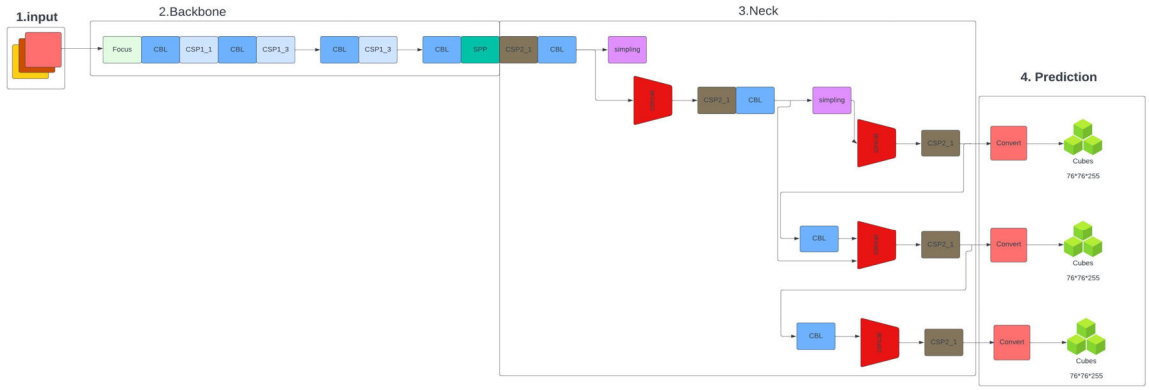

Figure 1. Yolov5 introduction

Figure1 illustrates the overall framework of the YOLOv5 object detection algorithm. The algorithm comprises four primary modules, namely Input, Backbone Network, Neck Network, and Head Output, represented by the four red modules in the image above. YOLOv5 has different versions which are YOLOv5s/m/l and x. In this paper, the focus will be on YOLOv5s, while the other versions are variations with deeper and wider networks based on this version.

The Input module is responsible for receiving the image input, with the network's input image size set to 608×608. During this stage, image preprocessing is typically performed, including resizing the input image to the network's input size and normalization. Throughout network training, YOLOv5 incorporates the Mosaic data augmentation operation, which enhances both model training speed and network accuracy. Additionally, the algorithm introduces adaptive anchor box calculation and adaptive image scaling methods to further improve performance.[3]

The Backbone Network functions as a high-performance classifier network responsible for extracting general feature representations. In the case of YOLOv5, it implements the CSPDarknet53 structure and the Focus structure as the backbone network to bolster its feature extraction capabilities.

The Neck Network is positioned between the Backbone Network and the Head Output. Its purpose is to further enhance feature diversity and robustness. Although YOLOv5 also incorporates SPP (Spatial Pyramid Pooling) and FPN+PAN (Feature Pyramid Network + Path Aggregation Network) modules, the specific implementation details differ from other approaches.

The Head Output module is responsible for generating the object detection results, usually consisting of two branches which are called classification and a regression. YOLOv5 utilizes GIOU_Loss (Generalized Intersection over Union Loss) instead of the Smooth L1 Loss function used in YOLOv4 to further improve the algorithm's detection accuracy.[4]

2. Experiment

In Yolov5, the K-means algorithm is used in object detection to determine a set of anchor box sizes, which are a group of predefined bounding boxes representing different object sizes and aspect ratios. These anchor boxes help adapt to the variations in object sizes and aspect ratios, thereby improving the performance of the object detector.

Traditional K-means algorithm typically starts with randomly selecting K data points as initial centroids, which can lead to sensitivity to the choice of initial centroids and result in unstable clustering outcomes. To address this issue, the K-means++ algorithm is employed as a smarter alternative. In the context of Yolov5, the K-means++ algorithm selects initial centroids in a more intelligent way. It begins by randomly selecting one data point from the dataset as the first centroid. The subsequent centroids are then chosen based on the farthest distance from the existing centroids, iteratively continuing this process.[5]

By adopting this selection strategy, the initial centroids are relatively distant from each other, which increases the likelihood of finding a globally optimal solution. By replacing the traditional K-means algorithm in Yolov5 with K-means++, the clustering process is expected to be more stable and yield better anchor box sizes, ultimately leading to improved object detection performance.[6]

Step 1: To begin with, pick a sample randomly as the initial cluster center from the dataset.

Step 2: Afterwards, we need to determine the shortest distance from each sample to the current cluster centers, denoted as T(x). Then, we calculate the probability for each sample to be chosen as the next cluster center.

$$\frac{T(x)^2}{\sum_{x \in x} T(x)^2}$$

Step 3: Finally, we use the roulette wheel selection method to choose the next cluster center.

Step 4: Repeat Steps 2 and 3 until the next cluster center is selected.

3. Conclusion

By comparing the experimental results, we find that K-means++ performs better in the YOLOv5 model than the traditional K-means algorithm, which has significant advantages. K-means++ uses a smarter initialization method that enables better selection of the initial center of mass, thus avoiding problems with local optimal solutions. This allows K-Means ++ to better capture the differences and features between target bounding boxes, resulting in more accurate prior boxes.

References

[1] ZEILER M D, FERGUS R. (2014) Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. Berlin. 818-833.

[2] TAN M, LE Q V. (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: 36th International Conference on Machine Learning. Long Beach. 10691-10700.

[3] RAJENDRAN S P, SHINE L, PRADEEP R, et al. (2019) Fast and accurate traffic sign recognition for self driving cars using retinanet based detector. In: 4th International Conference on Communication and Electronics Systems. Piscataway. 784-790.

[4] TRAN A C, DIEN D L, HUYNH H X, et al. (2019) A model for real-time traffic signs recognition based on the yolo algorithm–a case study using vietnamese traffic signs. In: 6th International Conference on Future Data and Security Engineering. Berlin. 104-116.

[5] KHAN J A, YEO D, SHIN H. (2018) New dark area sensitive tone mapping for deep learning based traffic sign recognition. Sensors, 18(11):1-13.

[6] Li Menghao, and Yuan Sanan. (2023) Improved traffic sign detection algorithm of YOLOv5s. Journal of Nanjing University of Information Science and Technology (Natural Science Edition), 13.