ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

YOLOv7-based small object detection model

Zhu Fan
Xihua University, Chengdu, Sichuan, 610039, China

Abstract: Aiming at the problems of high model computation cost and easy feature loss during lightweighting in small target detection task, this paper proposes a compact feature migration method based on YOLOv7, which reduces the number of model parameters by combining with knowledge distillation technique and can maintain relatively good detection effect. The method reduces the number of model parameters by designing a lightweight student network through channel pruning; then dynamically assigns spatial weights to enhance the feature alignment in small target regions through the Adaptive Feature Distillation (AFD) module; and finally, further compresses the model by combining structured pruning with quantized perceptual training. The experimental results show that this paper's method decreases the model mAP by only 1.8% with 35% reduction in the amount of parameters and 53% reduction in the amount of computation. It can be seen that the method provides an effective lightweight solution for small target detection, which has certain practical significance.
Keywords: object detection; YOLOv7; knowledge distillation; lightweight model; multi-scale feature transfer

1. Introduction

Small target detection is the process of locating and identifying objects in videos or images, outputting the object category, the position of the bounding box, and the detection confidence. From general object detection to specific scenario applications, it finds use in various industries such as autonomous driving, security surveillance, and medical image analysis, offering promising development and research prospects. Below is an introduction to relevant deep learning-based object detection algorithms.

Models such as the YOLO series [1], Faster R-CNN [2], SSD [3], and RetinaNet [4] have significantly improved detection accuracy and speed through strategies including multi-stage network design, anchor box optimization, and feature pyramid fusion. YOLOv7 [5] achieves 56.8% mAP@0.5 on the COCO dataset and an inference speed of 68 FPS (Tesla V100) through its Extended Path Aggregation Network (E-ELAN) and dynamic label assignment strategy. However, these models heavily depend on computational resources—YOLOv7 has 66.3. million parameters and 165.2. GFLOPs, posing severe challenges for deployment on edge devices (e.g., drones, mobile robots, embedded cameras). In small target detection scenarios, existing models struggle to balance accuracy and efficiency due to insufficient resolution in shallow features and the loss of semantic information in deep layers. Target Detection Process as is shown **Figure 1**.

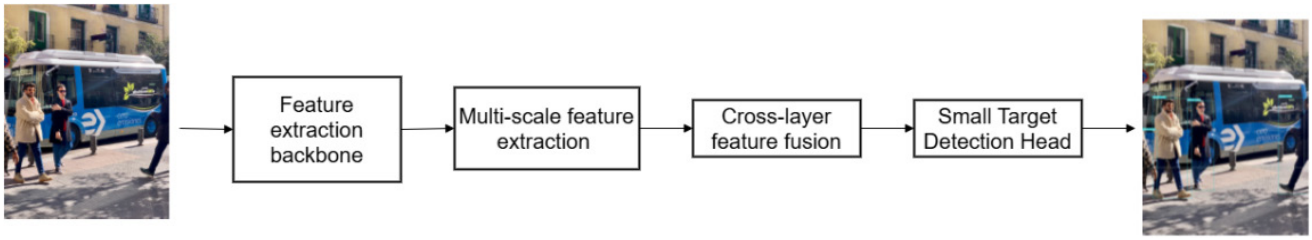

Figure 1. Target detection process.

ISSN: 2424-8460(O)
2251-2608(P)

2. Related theories and technologies

2.1. Overview of target detection technology

Target detection, being a core task in computer vision, has witnessed rapid technological iterations—Evolving from early reliance on hand-crafted features, to the application of deep learning algorithms, and now incorporating new technologies such as Transformers.

In 2001, the Viola-Jones face detection algorithm, based on Haar features and a cascade of AdaBoost classifiers, achieved real-time face detection for the first time, reaching a detection rate of 97% on the FERET face database. It became a milestone in the field of object detection. Subsequently, HOG (Histogram of Oriented Gradients) features combined with SVM classifiers achieved breakthroughs in pedestrian detection, boosting detection accuracy to 85% on the INRIA pedestrian dataset. However, these methods were limited in complex scenarios; challenges like lighting variations, occlusions, or scale differences could cause feature failure. Furthermore, traditional methods required extensive manual parameter tuning, lacked generalizability, and struggled with massive data and complex scenes.

2.2. YOLOv7 model architecture and features

The introduction of the YOLO (You Only Look Once) model in 2015 fundamentally shifted the technological landscape in object detection, primarily due to its "end-to-end"detection framework and real-time performance capability.

The neck network adopts a multi-path aggregation strategy (e.g., E-ELAN structure), achieving deep fusion of shallow detail features and deep semantic information through more flexible cross-layer connections, significantly boosting feature representation capabilities.

3. YOLOv7-based small target detection model design

3.1. Model overall architecture design

The teacher model is a complete YOLOv7 structure that is used to generate predictions for the teacher model. The student model is also a YOLOv7 structure, but its parameters are adjusted by knowledge distillation during the training process. The knowledge distillation component calculates the distillation losses of the teacher model and the student model on bounding box predictions and category predictions and combines these losses into a total distillation loss. The output outputs the final detected bounding boxes and their categories based on the predictions of the student model. The training module utilizes the total distillation loss for backpropagation to update the parameters of the student model.

3.2. Teacher-student network architecture design

where the teacher model is based on the original YOLOv7 architecture, retaining its multi-scale prediction module (PANet) and dynamic label assignment mechanism to ensure strong feature capture capability for small targets. By retaining the advantages of the core, it ensures a higher detection capability, especially for small target detection.

The student model is lightened and modified based on the YOLOv7 model by replacing the depth separable convolution, which reduces the number of parameters by replacing some of the standard convolutional layers with depth separable convolution. Channel pruning is used to remove redundant channels by sparse training to compress the model width. Cross-stage feature reuse, a cross-layer connection module is designed to reuse shallow features to reduce computation.

ISSN: 2424-8460(O)
2251-2608(P)

The design of the student-teacher architecture enables the teacher to maintain high detection performance and the student to maintain good detection efficiency, and realizes the double improvement of detection performance and detection effect through knowledge distillation.

4. Summary

In the field of target detection, the existing models generally have the problems of large parameters and high resource consumption of computation, which are difficult to meet the mobile's demand for real-time and lightweight target detection.

To address the above problems, this paper proposes a compact feature migration method for small target detection, which migrates the multi-scale feature expression capability of the YOLOv7 teacher model to the lightweight student model through the knowledge distillation technique and the hierarchical sensitivity pruning technique. The target detection model is realized to ensure the effectiveness of detection while the amount of parameters is reduced.

References

[1] Zeni L F, Jung C R. Distilling knowledge from refinement in multiple instance detection networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 768-769.

[2] Farhadi M, Ghasemi M, Vrudhula S, et al. Enabling incremental knowledge transfer for object detection at the edge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 396-397.

[3] Zheng Z, Ye R, Hou Q, et al. Localization distillation for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10070-10083.

[4] Dai X, Jiang Z, Wu Z, et al. General instance distillation for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7842-7851.

[5] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.