Fruit Recognition Based on YOLOX*

Keying Ren [1], Xiaoyan Chen[1]*, Zichen Wang[1], Xiaoning Yan[2], Dongyang Zhang[2]

[1] Tianjin University of Science and Technology, China;

[2] Shenzhen Softsz Co. Ltd., China

E-mail: renkeying@mail.tust.edu.cn
www.tust.edu.cn

Abstract

Pattern recognition is an urgent problem to be solved in the field of computer vision. In this paper, the network of fruit recognition based on YOLOX is studied. Due to the problem of slow training speed and low accuracy in the classical algorithms, the de-coupling detection head is optimized in YOLOX to overcome the above shortcomings. In terms of data enhancement, a new method combining Mosaic and MixUp is proposed. Through experimental verification, the method proposed in this paper has a great improvement over related algorithms such as YOLOv5, the accuracy is 98.6%, which is increased 5.2%.

Keywords: Pattern recognition, de-coupling detection head, YOLOX, Fruit recognition

1. Introduction

Pattern recognition is an urgent problem to be solved in the field of computer vision, aiming to detect objects of predefined categories. Accurate object detection impacts on various applications including image recognition and video surveillance. In recent years, with the development of convolutional neural network (CNN), object detection divided into one-stage and two-stage methods. A typical one-stage method is R-CNN algorithm. The main idea is to generate a series of bounding boxs through selective search or CNN network, These bounding boxes are classified and regressed[1]. The advantage of two-stage method is high accuracy. One-stage method, like You Only Look Once(YOLO) and Single Shot MultiBox Detector(SSD)[2,3].The main idea is to sample uniformly and concentratedly in different positions of the image. Different ratios can be used in sampling. CNN is used to extract features and then directly perform classification regression. The whole process only needs one step, so the advantage is fast speed. However, the disadvantage is that uniform and dense sampling is difficult to train. Due to the positive sample and the negative sample (background) unbalanced, which leads to the model accuracy decrease.

Both of them first tile a large number of preset anchors on the image, then predict the category and refine the coordinates of these anchors by one or several times, finally output these refined anchors as detection results. Currently, the leading algorithms of One-Stage are YOLO series. YOLO transforms the target detection task into regression problem, greatly speeding up the detection speed. At the same time, because the network

used global information to predict each target window, the proportion of false positive was greatly reduced.

Recent academic attention has been geared toward anchor-free detectors due to the emergence of FPN[4] and Focal Loss[5]. Anchor-free detectors directly find objects without preset anchors in different ways. One is keypoint-based methods. Another is center-based method. These anchor-free detectors are able to eliminate those hyperparameters related to anchors and have achieved similar performance with anchor-based detectors, making them more potential in terms of generalization ability.

2. YOLOX

Megvii Technology proposed a new high-performance detector — YOLOX. YOLOX is an object detector which uses the feature of deep convolutional neural network. it makes some empirical improvements to the YOLO series, such as anchor-free, a decoupled head and the leading label assignment strategy SimOTA.

2.1. *Anchor-free*

Anchor-free detectors have developed rapidly in the past two years. These work have shown that the performance of anchor-free detectors can be on par with anchor-based detectors. Anchor-free mechanism significantly reduces the number of design parameters which need heuristic tuning and many tricks involved for good performance, making the detector, especially its training and decoding phase, considerably simpler.

Keypoint-based method. This type of anchor-free method first locates several pre-defined or self-learned keypoints, and then generates bounding boxes to detect objects. CornerNet [6] detects an object bounding box as a pair of keypoints (top-left corner and bottom-right corner). The R-CNN [7] locates objects via predicting grid points with the position sensitive merits of FCN and then determining the bounding box guided by the grid. ExtremeNet detects four extreme points (top-most, leftmost, bottom-most, right-most) and one center point to generate the object bounding box.

Center-based method. This kind of anchor-free method regards the center of object as foreground to define positives, and then predicts the distances from positives to the four sides of the object bounding box for detection. DenseBox uses a filled circle located in the center of the object to define positives and then predicts the four distances from positives to the bound of the object bounding box for location.

2.2. *Decoupled head*

In object detection, the conflict between classification and regression is a well-known problem. Thus the decoupled head for classification and localization is widely used in the detectors. However, as YOLO series' backbones and feature pyramids continuously evolving, their detection heads remain coupled. This structure is shown in Fig.1. The coupled detection head may harm the performance. Replacing YOLO's head with a decoupled one greatly improves theconverging speed. The decoupled head is essential to the end-to-end version of YOLO.

YOLOX replaced the YOLO head with a decoupled YOLO head, the convergence speed has been greatly improved. Concretely, it contains a 1 × 1 conv layer to reduce the channel dimension, followed by two parallel branches with two 3 × 3 conv layers respectively. This structure is shown in Fig.2.

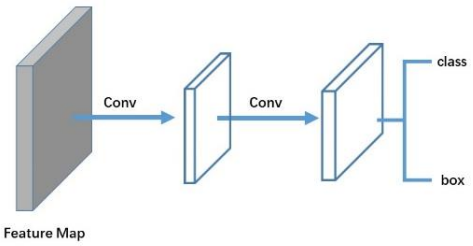

Fig.1. The original YOLO series coupling head structure, a convolution directly carried out classification and anchor regression.

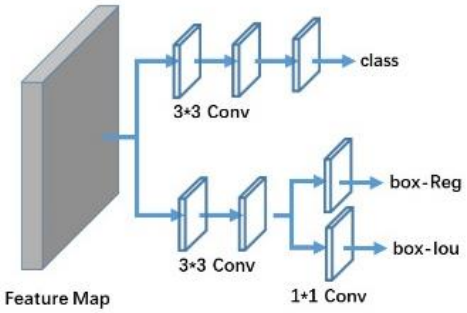

Fig.2. replace the YOLO head with a decoupled YOLO head, the convergence speed has been greatly improved.

Concretely, it contains a 1 × 1 conv layer to reduce the channel dimension, followed by two parallel branches with two 3 × 3 conv layers respectively.

3. Related work

3.1. *Images dataset*

Our data set consists of 30,000 photos of 60 fruits, labeled by VOC. Images were divided into training and test datasets with ratio of 9:1. The hardware configuration include one NVIDIA GeForce GTX1080ti SUPER graphic card, Intel(R) Xeon(R) CPU E5-2630 - Core Processor, and. The CSPDarknet framework and Python 3.6 were used.

3.2. *Image Preprocess*

Mosaic and MixUp were added to the enhancement strategy to improve YOLOX performance. Mosaic is an effective enhancement strategy proposed in Yolov3. and widely used in YOLOv4. MixUp is originally designed for image classification task but then modifie for object detection training. MixUp and Mosaic are uesd during training and hand off for the last 15 epochs.

Fig.3. First, the dataset was expanded using mirroring and rotation. Then, Mosaic method was used to merge images.

Fig.4. The MixUp method

4. Experiments and Results

Object detection performance is evaluated with Average Precision (AP) and mean Average Precision (mAP) between ground truth and predicted bounding box (IOU).

To verify the feasibility of YOLOX in fruit recognition, two ex-periments are performed in this study. The experiment results are shown in Table 1.

Table 1. Comparison of mAP and IOU on YOLOv4 and SO-YOLO.

Method	Average IOU%	mAP%
YOLOv5-s	54.09	93.4
YOLOX-s	53.27	**98.6**

Experiment results reported in Table 1 is the average of multiple experiments. It can be seen that YOLOX has higher accuracy than YOLOv5-s. As shown in Table 1, mAP is increased from 93.4% to 98.%. Comparing to YOLOv5-s the proposed method has better performance and also well-balanced accuracy and processing time, experimental results are shown as Table 2.

Table 2. Comparison of network model parameters of each model.

Version	parameters	Latency
YOLOX-s	9.0 M	9.8 ms
YOLOv5-s	**7.3 M**	**8.7 ms**

Comparing with YOLOv5-s, the model parameters re-duced from 7.3M to 9.0M. Although the parameters and reasoning speed increased, it was acceptable for our project.

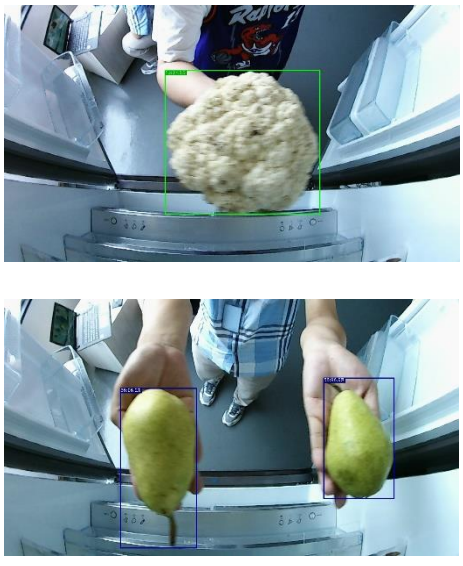

Fig.4. The result of recognition by YOLOX detector, it can be seen that the category confidence is very high, and box regression is very good.

5. Discussion

In order to improve the classification and detection of fruits, YOLOX method was applied and verified in this paper. We can see that YOLOX is good at fruit recognition tasks. YOLOX integrates the advantages of other target detection algorithms. Mosaic and MixUp were added to the enhancement strategy, which improved YOLOX-s performance. Replacing YOLO's head with a decoupled one greatly improves the converging speed. the performance of anchor-free detectors can be on par with anchor-based detectors. Anchor-free mechanism significantly reduces the number of design parameters.

References

1. Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(6):pp.1137-1149.
2. Redmon J , Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection, *IEEE conference on computer vision and pattern recognition*, 2016: pp. 779-788.
3. Liu W , Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector, *European Conference on Computer Vision. Springer, Cham*, 2016: pp.21-37.
4. Lin T Y , Dollar P , Girshick R , et al. Feature Pyramid Networks for Object Detection, *Proceedings of the IEEE conference on computer vision and pattern recognition. 2017* : pp.2117-2125.
5. Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, PP(99):2999-3007.
6. Duan K , Bai S , Xie L , et al. CenterNet: Keypoint Triplets for Object Detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2019: pp.6569-6578.
7. Lu X , Li B , Y Yue, et al. Grid R-CNN, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2019: pp.7363-7372.

Authors Introduction

Mr. Keying Ren

Keying Ren is an Master of Control Science and Engineering, Tianjin University of Science and Technolog. The research topic is target detection and tracking based on deep learning.

Prof. Xiaoyan Chen

Xiaoyan Chen professor of Tianjin University of Science and Technology, graduated from Tianjin University with PH.D (2009), worked as a Post-doctor at Tianjin University (2009.5-2015.5). She had been in RPI, USA with Dr. Johnathon from Sep.2009 to Feb.2010 and in Kent, UK with Yong Yan from Sep-Dec.2012. She has researched electrical impedance tomography technology in monitoring lung ventilation for many years.

Mr. ZiChen Wang

ZiChen Wang was born in Tianjin, China in 1997. He received the B.E. degrees from Tianjin University of Science and Technology, Tianjin, China, in 2019. He is currently pursuing the M.S. degree in Information and Automatic College at Tianjin University of Science and Technology, Tianjin, China.

Mr. Xiaoning Yan

Xaoning Yan graduated from South China University of Technology with a bachelor's degree in software engineering. 2013.8-2015.5 studied for a master's degree in computer science at the University of Texas at Dallas.

Mr. Dongyang Zhang

Dongyang Zhang, a master's degree in control science and engineering from Tianjin University of Science and Technology. is currently an algorithm engineer of Ansoft Technology, engaged in deep learning and computer vision algorithm research, It mainly involves the research direction of target detection and pedestrian re-recognition.