Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230905

Research of target detection method YOLO

Rongyu Nie

School of Computer Science and Technology, XiDian University, Xifeng Road, Xi 'an, China

631401110118@mails.cqjtu.edu.cn

Abstract. Regression theory is used by YOLO technology to build the one stage detection technique. It merely employs a trunk CNN to predict various targets using the "feature extraction-direct regression" method. In comparison to previous algorithms, it detects things much faster and with far higher accuracy. Researchers have become interested in the YOLO model because it is widely employed in sectors such as autonomous driving, camera display, video surveillance, vehicle identification, face recognition, remote sensing satellite, infrared detection, and others. In this study, the model structural properties of the yolo model and the yolov1-v7 models are primarily analyzed and contrasted. According to the model design perspective, it compares and summarizes the structural enhancement and corresponding performance optimization of each model in the model structure, assesses the benefits and drawbacks of each model, and assesses their performance in light of the actual application impact of each model and the primary application fields, serving as a reference for the study of related topics. The article's summary and future direction are provided at the end.

Key words: Artificial Intelligence, Target Detection, Image Processing, YOLO.

1. Introduction

Target detection is the process of looking for, identifying, categorizing, and discovering one or more different categories of targets in an image. Target detection has long been a highly difficult research topic in the field of depth learning computer vision due to the interference of object shape, imaging time line, complex background, occlusion, and other factors. Target detection algorithms have advanced quickly in recent years. There are two categories that can be applied to the common algorithms today [1]. The R-CNN series algorithm, a two step approach, falls within the first group. Candidate boxes, feature extraction, and classification judgment are suggested through the RPN network as an alternative to the conventional method of feature extraction by sliding windows. The candidate boxes will then be precisely positioned to categorize and regression the target position after being obtained. As a result of the vast number of a priori frames that are generated and could potentially include target objects, identification accuracy is great but speed is slow. The processing boundary should then be modified, and the initial frames with no target to be measured, poor confidence, and high overlap should be filtered.

The second type is YOLO series, which is a one-stage procedure. The "feature extraction+direct regression" method is used to forecast various targets using a single trunk convolution neural network CNN. Although the method is quick, recognition accuracy is poor. YOLO is a new framework

DOI: 10.54254/2755-2721/6/20230905

suggested by Ross Girshick to speed up deep learning object identification, after RCNN, FAST-RCNN, and TASTER-RCNN [2]. He approaches the object detection problem as a regression issue in a novel way. The fundamental concept is to use the entire image as the network's input and a method of predetermined candidate areas to directly estimate the location and category of the Bounding Box at the output layer. This reduces the size of the map but results in a significant speed boost. The idea of proposal+classifier is still used by the fast RCNN, even though it also directly uses the entire graph as input. However, CNN implements the process of extracting proposal [3].

Yolo model is a cutting-edge target detection framework with broad application needs and potential in surveillance, traffic, robots, and even the conjunction with the most recent advancement in the electromagnetic brain radiation sector. Therefore, it is crucial to thoroughly investigate and analyze each model in the YOLO series. The pertinent research content is currently not thorough enough and requires additional analysis and debate. In order for readers to gain knowledge in this area, this document primarily compares and examines the yolo model's characteristics and model concept from versions 1 through 7.

2. Main Body

2.1. Model design of the YOLO

The fundamental principle of YOLOv1 is to directly regress the position and category of the Bounding Box on the output layer using the entire graph as the network's input (Figure 1). Fast, low background false detection rate, and strong universality are the benefits. The target detection effect is poor when the input size is fixed and the proportion is tiny, which is a drawback. Yolo training model only supports the same input resolution as the training image in detection; alternative resolutions must be scaled to this fixed resolution because the whole connection layer is used as the output layer. There are many factors and a great deal of information is lost. Despite the fact that each grid is capable of predicting B bounding boxes, only the bounding box with the highest IOU is chosen for each grid's object detection output. In other words, even though each grid may contain several things, only one object can ultimately be anticipated [4]. Additionally, the detection capacity of small targets is restricted, and the positioning accuracy of object detection is poor due to the IOU error of large objects and small objects in the loss function being close to the LOSS contribution value in network training.

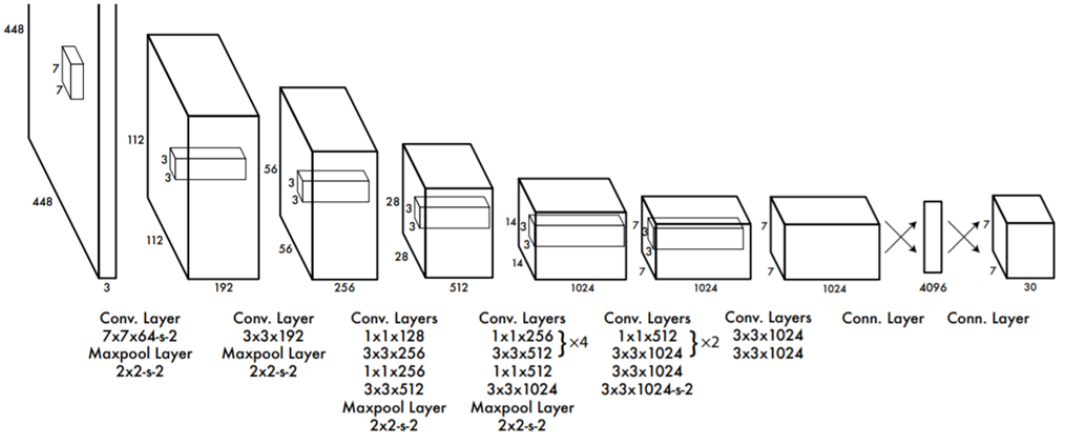

Figure 1. YOLOv1 pictures are divided into 7 × 7's network (gird cell) [4].

When compared to v1, YOLOV2 can improve the model's nonlinear expression capability while simultaneously acting as a regularizer thanks to BatchNorm. The resolution of the training image is improved, and this modification causes the model to increase by 4%. The anchor based approach is added. Only the addition of BatchNorm increases YOLO's performance by nearly 2%. Unlike the R-CNN series, YOLO v1 directly predicts and constructs anchors without an a priori box. However, if

DOI: 10.54254/2755-2721/6/20230905

the two can be integrated, going from an initial anchor-free model to an anchor-based model, we can benefit from integrating two types of target detection, obviously increasing forecast accuracy while preserving the speed advantage. When employing anchors with RPN networks, there are certain changes. The model's convergent rate may be slowed down by a big offset value in RPN. As much as feasible, YOLO V2 has upgraded it to benchmark YOLO V1, which means that the prior box will always be forecasted on the same grid and won't be moved to another grid. Each anchor in YOLO V2 now has its own classification probability, which is a significant advance over the previous two techniques and one that boosts model performance by more than 5%. Since two bboxes share a set of classification probabilities in V1, we know that a grid can predict just one object. One innovation in V2 is that each anchor has its own set of classification probabilities.

User can use the clustering approach to determine the right anchor size [5]. The generic model in R-CNN will pre-set a number of anchor sizes that are super parameters and can be changed independently. The process of manually choosing an a priori box is also used in Faster R-CNN. According to experience, the anchor box's size and proportion are determined, and throughout training, the network modifies these. However, in YOLO, the author uses a clustering-like technique called Dimension Clusters, clustering every bbox in the training set and choosing the best bboxes as a priori boxes. The priori box is the subject of the next challenge. Based on this, YOLOv2 has improved. To create a suitable priori box, its clusters on the training set bbox using k-means [6]. The larger bbos will yield greater mistakes than the smaller bbox because the conventional k-means, or the European distance, is employed to measure the difference, and the IOU is independent of the bbox size. So that good IOU scores can be gained through these anchor boxes, the IOU is employed to participate in the distance computation(Figure 2)。

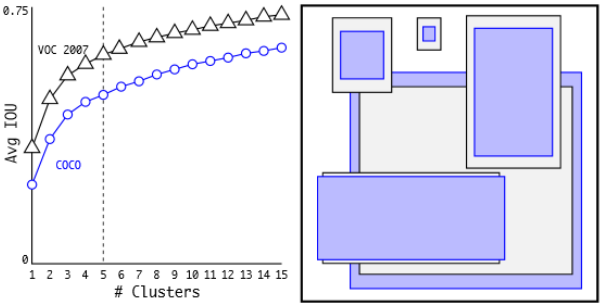

Figure 2. Distance formula in K-mean clustering: d (box, centroids)+1-IOU (box, centroids) [6].

Including a passthrough layer This layer is meant to address the issue that YOLO V1 cannot detect small targets. Instead of introducing the RPN layer for mitigation, the next to last size is 26 × 26 × 512 feature map is also used, a 2 × 2 ×512 Pixels in the local area are extracted and become 1 × 1 × 2048 pixels, so the penultimate size of the feature map which is 26 ×26 × 512 becomes 13 ×13× 2048, and then used together with the feature map of the last layer; Multi scale training. A popular rising point technique in the realm of object detection is multi scale training. The model can be more resilient and it is simpler to learn more features if the image is randomly scaled to the designated region rather than to the designated size. For instance, scaling up small things helps with detection.

DOI: 10.54254/2755-2721/6/20230905

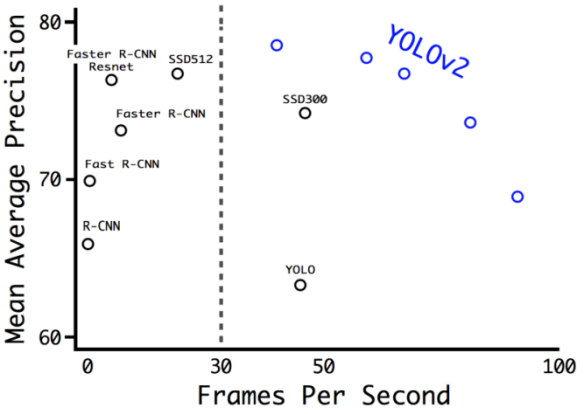

Figure 3. Comparison between YOLOv2 and other models in VOC2007 dataset [7].

Three new feature extraction networks are proposed from v1-v3 and yolo series. The model in v1 uses a backbone network similar to GoogLeNet, with a fixed input and output; V2 builds a new DarkNet-19 on the basis of v1 and VGG network (Figure 3). The precision is equivalent to VGG, but the floating point operation amount is only 1/5 of VGG, so the speed is very fast; In v3, DarkNet-53 is proposed, which combines the residuals module of ResNet. The output feature maps are 13x13, 26x26, and 52x52 with different sizes. It prevents the network convergence issue brought on by the network gradient explosion while simultaneously strengthening the network structure. As a result, it can adapt better to physical examinations of various sizes. The network model has been improved, making small target recognition easier. In order to extract more information, YOLOv3 also uses tensor splicing to expand tensor dimensions [5, 7]. The characteristics are more particular, and information from many persistent feature maps is combined to anticipate objects with various requirements. Yolov3's priori boxes are richer than Yolov2's. YOLOV3 has enhanced YOLOv2's single label classification to multi label classification in order to detect objects of various sizes. Head has changed the Softmax classifier used for single label classification to multiple independent Logistic classifiers used for multi label classification, eliminating the mutual exclusion between categories, which makes the network more flexible. Three scales are designed, each with three specifications, and a total of nine, as shown in Figure 4 below, there are five types of yolo2. The overall accuracy and the accuracy of small object detection are milestone improvements.

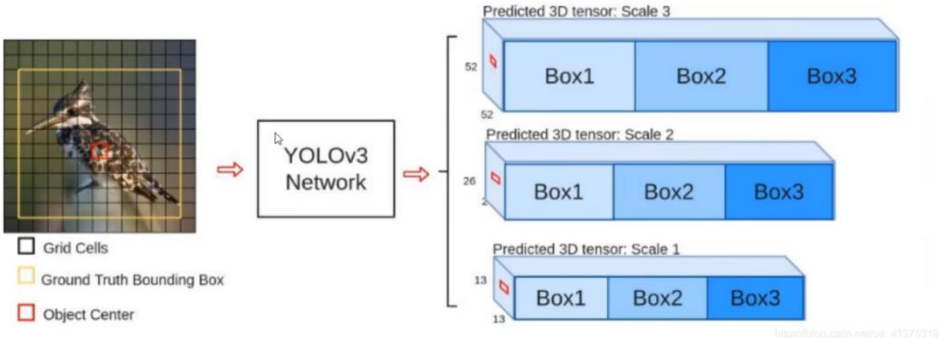

Figure 4. YOLOv3 designs three networks of different specifications to divide the original image [7].

DOI: 10.54254/2755-2721/6/20230905

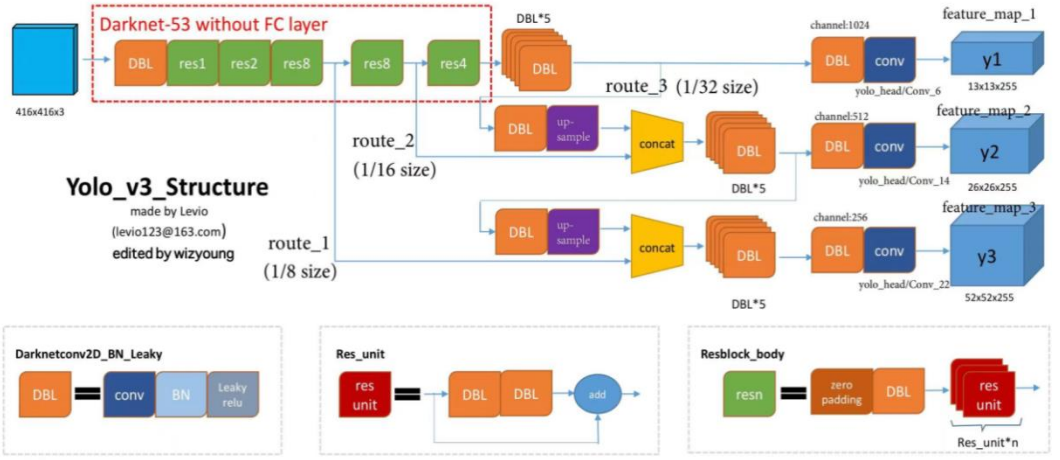

Figure 5. YOLOv3 Network Structure [8].

In Yolov3(Figure 5), the author lists the data results of training using the coco dataset. It shows that when the size of input image is 416x416, the mAP size of the model is 31%. When the input image's size is 608x608, the mAP of the model reaches 33%. The strength of Yolov3 can be seen from the results shown in the table 1 [8].

Table 1. Data Results of YOLOv3 Training with Coco Dataset [8].

	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_{50}	AP_L
Two-stage methods							
Faster R-CNN+++ [5]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [8]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [6]	Inception-ResNet-v2	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [7]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
One-stage methods							
YOLOv2 [9]	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD [3]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD [3]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [9]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [9]	ResNeXt-101-FPN	40.8	61.1	44.1	24.1	44.2	51.2
YOLOv3	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Yolov4 has improved its substructure on the basis of v3, deleting the final pooling layer, full connection layer, and softmax layer (Figure 6). Its trunk has five CSP modules. The introduction of SSP and PANet modules significantly increased the receptive field.

DOI: 10.54254/2755-2721/6/20230905

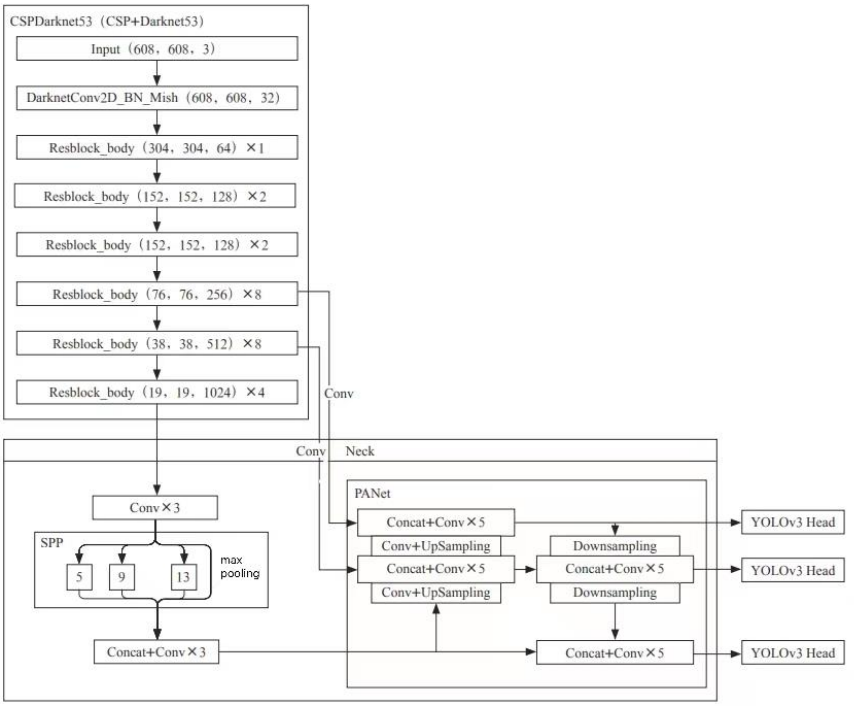

Figure 6. YOLOv4 Network Structure [9].

YOLO V4 basically aims to choose a few target detection models that have been applied to a variety of detectors since YOLO V3 was released and can increase detection accuracy. YOLO V4 can guarantee speed while significantly enhancing the model's detection accuracy [9]. Although the detection accuracy is not as good as Efficient Det, the speed is well ahead, as seen in the image below. Yolov4's AP is 41 and Yolov3's is 32, which is a straight gain of 9 percentage points, when the FPS is equal to or around 100 in the coco data set (Figure 7).

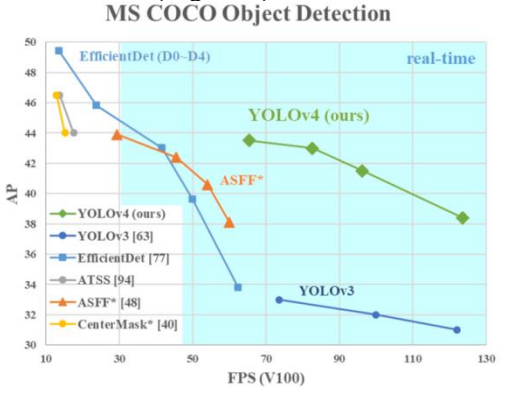

Figure 7. Comparison between YOLOv4 and other target detection models [9].

The advantages of YOLOv5 include taking into account the neighborhood's positive sample anchor matching method and adding positive samples(Figure 8); Flexible configuration options can be used to create models of varying complexity; enhance overall performance using several super parameter optimization techniques built-in; It makes advantage of Mosaic enhancement like yolov4. To increase the model's capacity for generalization and boost the detection effect, YOLOv5 makes advantage of Mosaic data improvement [10]. On the basis of the CutMix data augmentation method, this algorithm is improved. Images are spliced using random scaling, random clipping, and random layout. The benefit is that it enhances the detected object's backdrop and small target and enhances the

DOI: 10.54254/2755-2721/6/20230905

effectiveness of small object detection [11]. Yolov5s network is the network in the Yolov5 series with the smallest depth and smallest breadth of the characteristic graph. On this premise, the last three are all deepened and broadened. Of course, the feature extraction capability of the network is stronger the more convolution kernels there are and the bigger the feature map is. The smallest network, slowest speed, and least accurate AP are all on Yolov5s. It works best for spotting big targets and pursuing quickly. Based on this, the three other types of networks—YOLOv5m, YOLOv5x, and YOLOv5l—continue to deepen and extend their networks, improve their AP accuracy, and use more speed.

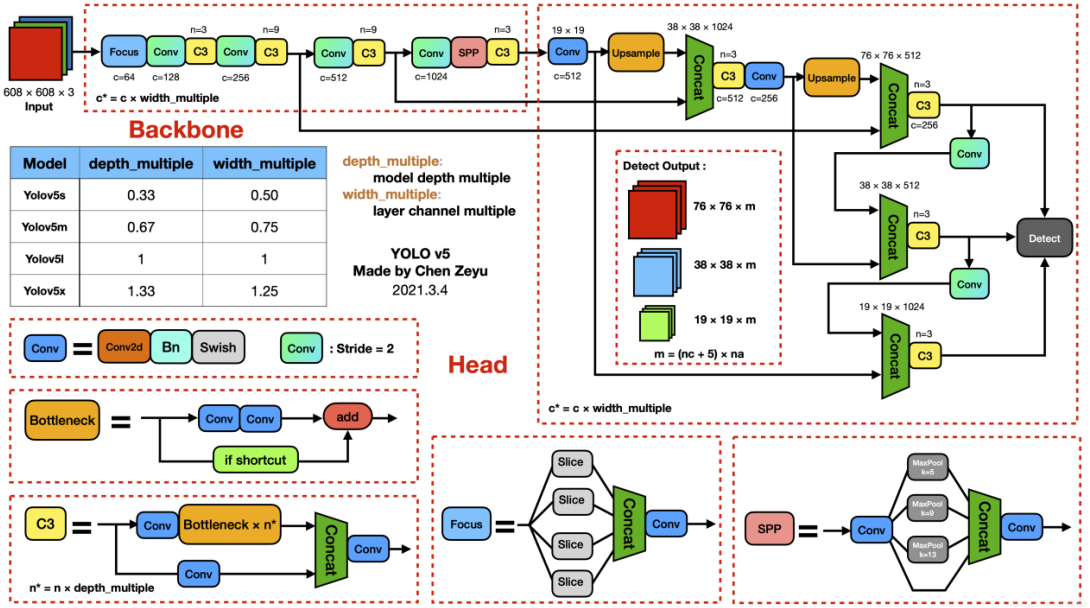

Figure 8. YOLOv5 Network Structure [10].

YOLOx uses three Decoupled Heads, focusing on cls (classification information), reg (detection frame information) and IOU (confidence level information) (Figure 9). YOLOx also uses the anchor free idea. Compared with the conventional anchor based in the YOLO series, the parameters on the head side can be reduced by about 2/3. Compared with anchor based method, which uses prior knowledge to design anchor size, anchor free thought takes receptive field as "anchor" information. In addition, Yolox has designed sample preliminary screening+SimOTA logic, which enables the algorithm to dynamically allocate positive samples to further improve the detection accuracy. At the same time, SimOTA algorithm uses Top-k approximation strategy to obtain the best matching of samples, which greatly speeds up the training speed [12].

DOI: 10.54254/2755-2721/6/20230905

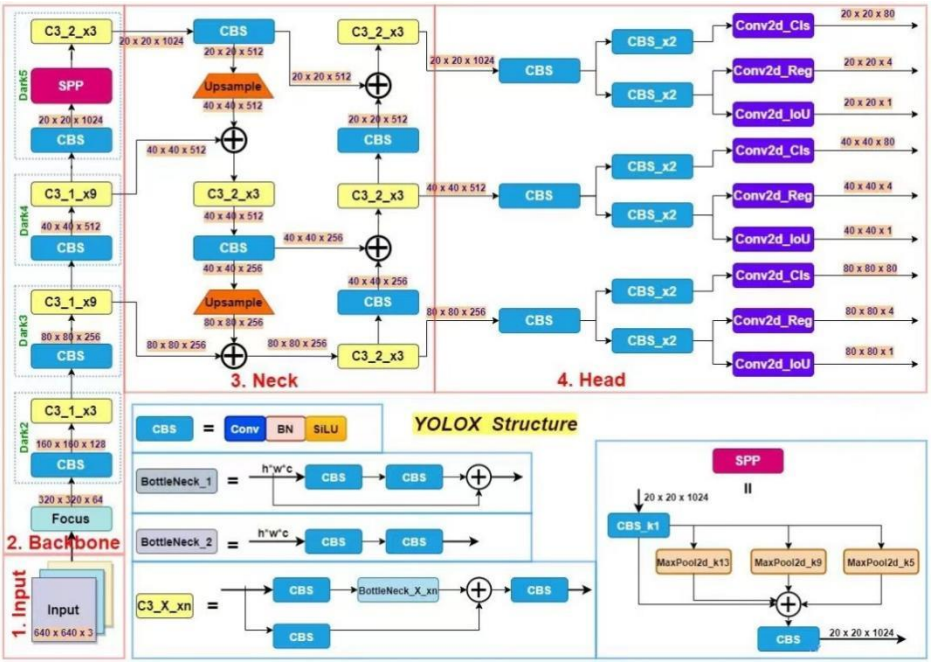

Figure 9. YOLOx Network Model [12].

YOLOv6 has made a new design for YOLO series backbones and necks, and designed RepBlock to replace CSPParknet53 module with reference to RepVGG network (Figure 10). The experimental results based on COCO dataset show that the backbones designed based on RepBlock have stronger representation ability and more efficient GPU utilization. In addition, all activation functions in YOLOv6 are ReLU, so as to improve the speed of network training. Transpose deconvolution is used for upsampling, and Rep-PAN is also used to replace the CSP module in the next, but the FPN-PAN structure is still retained. This modification makes its training convergence speed faster and its reasoning speed faster. To further improve the regression accuracy, YOLOv6 uses the SIoU detection box regression loss function to optimize the network learning process.

DOI: 10.54254/2755-2721/6/20230905

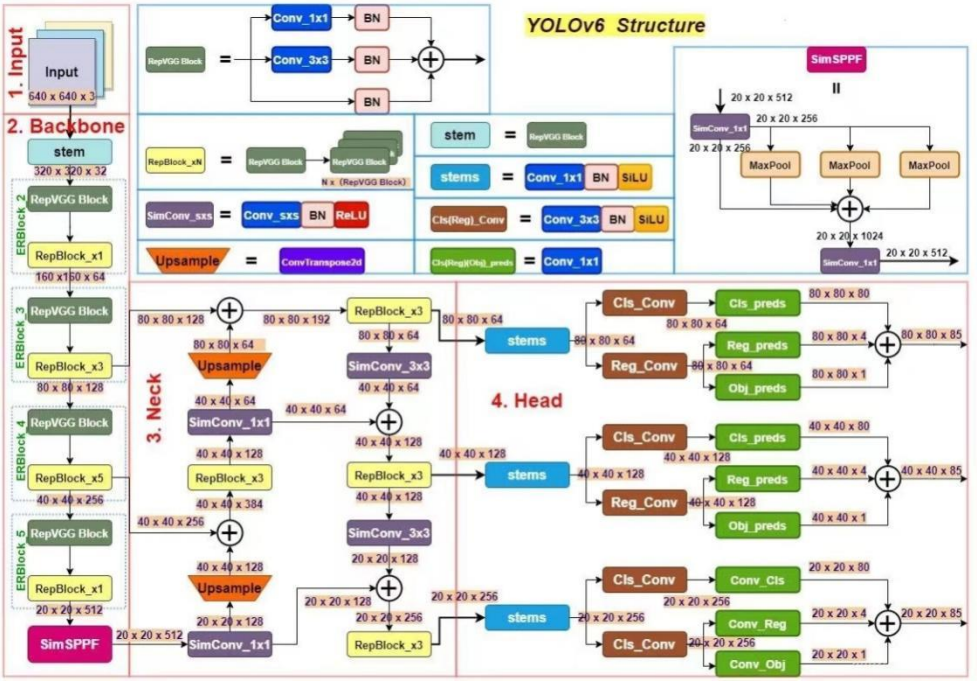

Figure 10. YOLOv6 Network Model [13].

YOLOv7 is the sequel of YOLOv4 team, which is mainly optimized for model structure reengineering and dynamic label allocation (Figure 11). The idea of YOLOv7 detection algorithm is similar to YOLOv4 and v5. The model structure re parameterization of the plan is proposed. The YOLOv5, Scale YOLOv4, YOLOX, "extension" and "composite scaling" methods are used for reference, so as to efficiently use parameters and computation. Compared with yolov5, YOLOv7's positive and negative sample allocation strategy adds loss aware, which makes use of the performance of the current model to enable real-time fine screening; Compared with using only the SimOTA algorithm in YOLOX, it can provide more accurate prior knowledge.

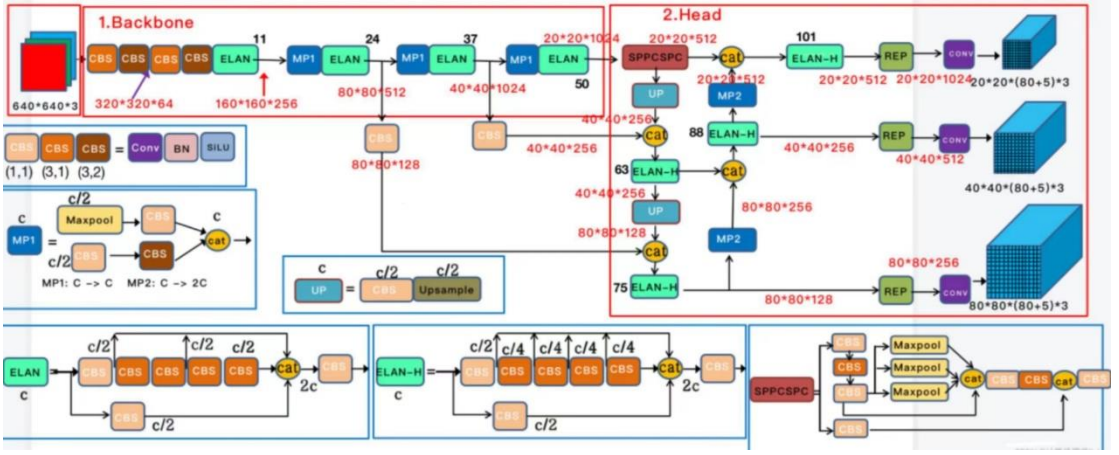

Figure 11. YOLOV7 Network Model [14].

2.2. Performance comparison of different YOLOs

The table 2 summarizes the improvements, advantages, disadvantages, and performance applications of each model.

DOI: 10.54254/2755-2721/6/20230905

Table 2. The performance comparison of different YOLOs.

Yolo model	Improvement	Advantage	Disadvantage	Performance
YOLOv1	The position and category of BBox are directly regressed on the output layer using the entire image as the network's input.	Fast, low background false detection rate, and strong universality	Each cell can only predict one category. If the overlap cannot be resolved. The aspect ratio is optional but single, and the detecting effect of small objects is average.	YOLO is also useful to finding objects in works of art. Compared to DPM and RCNN series detection approaches, it has a substantially greater detection rate for anomalous picture objects.However, its detection effect for small objects is poor, and it can only predict the maximum number of objects in the image.
YOLOv2	Use BatchNormalization; Improve the resolution of training image,; The anchor box is introduced. Based on YOLOv1, the final full connection layer is removed, and convolution and anchor boxes are used to predict the detection box.v	Compared with v1, the prediction is more accurate, faster, and more objects are recognized in terms of keeping the processing speed. It can adapt to various image sizes and provide balance between accuracy and speed.	The detection accuracy is not enough, which is slightly worse than SSD; Not good at detecting small objects; Low accuracy for close objects.	Especially for animals, the recognition effect is very good, but for clothing or equipment and other categories, the recognition effect is not very good.

DOI: 10.54254/2755-2721/6/20230905

Table 2. (continue).

Yolo model	Improvement	Advantage	Disadvantage	Performance
YOLOv3	Adjust the network structure; FPN is introduced to detect objects using multi-scale features; Put forward the idea of multi label classification; The loss function is optimized; No softmax is used by the classifier, and two classification cross loss entropies are used for classifying loss.	The problem of small object detection is mainly solved. The generalization ability is stronger on the day specific dataset, and multi-scale object detection ensures the diversity of detected objects.	Compared with RCNN, it is still lack of accuracy. The loss function is constructed by using the strategy of two classifications, which makes the result inevitably biased. When 50% area of the object image is taken as the recognition standard, the accuracy is the best, but when the standard is higher, the accuracy gradually decreases.	It is excellent in small object detection accuracy, but it is a little poor in large object positioning accuracy.
YOLOv4	CIOU Loss and DIOU NMS are introduced to improve the overall performance of the head side, Mosaic data is used for enhancement, and dropBlack is introduced.	Expand the database and balance the number of large, medium and small targets.	Slightly less flexible.	Yolov4 has better detection performance for occluded objects than yolov3.
YOLOv5	Auto Learning Bounding Box Anchors and neighborhood positive and negative sample allocation strategy are introduced. Because yolov4 lacks an adaptive anchor box, the anchor box is automatically learned from the training dataset.	Positive sample anchor matching strategy of the neighborhood is taken into account, and positive samples are added. Small size, light weight, quick speed, and positive sample addition; flexible configuration parameters allow for models of various complexity.	The vertical and horizontal comparison describes the relative value, which is somewhat vague; The balance problem of difficult and easy samples is not considered.	Yolov5 is slightly weaker than yolov4 in performance, but much faster and more flexible than v4, which enriches the background and small objects detection and improves the performance of small object detection.

DOI: 10.54254/2755-2721/6/20230905

Table 2. (continue).

Yolo model	Improvement	Advantage	Disadvantage	Performance
YOLOx	Decoupled head is proposed, Decoupled head is introduced into the network structure based on YOLOv5, and the anchor free idea and SimOTA positive and negative sample allocation strategy are used to calculate and optimize the loss function.	Network computing speed is improved; The receptive field of the network is expanded and more features are extracted when SPPneck is added to the backbone network; Reduced parameters [15] .	Quit perfect.	Good performance for small objects.
YOLOv6	On the basis of yolovx, the structure of Decoupled Head is improved, and the SloU bounding box regression loss is introduced into the loss function. A new design has been made for backbone and neck. All activation functions are ReLU.	The time consumption is further optimized to further improve the performance of yolo detection algorithm, and the recognition speed and accuracy are faster and higher.	It is still possible to increase the speed and accuracy.	Single level target detection framework for industrial applications.
YOLOv7	Convolution reparameterization is proposed and improved; Adopt efficient aggregation network and scale the model; Introduced the auxiliary training module (coarse to fine) guidance label allocation strategy	The amount of parameters and computation is greatly reduced, but the performance is still improved slightly	When dealing with complex backgrounds, errors and missed inspections are still easy to occur	For GPU devices supporting mobile GPUs and from the edge to the cloud, it can be used for the deployment of efficient detectors.

3. Conclusion

This paper combs and summarizes the yolo series widely used in target detection at present, from yolov1 to yolov7, in terms of the main features and improvements of model design, combined with the comparative analysis of network structure and training results, from point to surface, and then analyzes their performance according to the application status and effectiveness of each model. In target detection, the conflict between classification and regression tasks is a well-known problem. In the continuous evolution and development of the model, we can feel the continuous improvement and rapid performance improvement of the model. It has developed greatly in various practical application fields, and solved many problems that were difficult to overcome by the previous generation of models,

DOI: 10.54254/2755-2721/6/20230905

such as small object detection, low detection accuracy, complex background problems, and so on. With the development of network models in the direction of lightweight, to solve the problem of large computation and too many parameters, so that some embedded devices can also be used in mobile devices with poor computing performance. The yolo series has begun to explore the improvement of tiny. How to keep its detection accuracy while lightweight remains to be explored.

References

[1] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection. IEEE Conf. Com. Vis. Pat. Rec. 2016: 779-788.
[2] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. Euro. Conf. Com. Vis., 2016: 21-37.
[3] Redmon J, Farhadi A. Yolo 9000: Better, faster, stronger. IEEE Conf. Com. Vis. Pat. Rec. 2017. 6517-6525.
[4] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pat. Anal. Mac. Intel., 2016, 38(1): 142-158.
[5] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv: 1804.02767, 2018.
[6] Chen XL, Fang H, Lin TY, et al. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv: 1504.00325, 2015.
[7] Ren SQ, He KM, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. Inter. Conf. Neu. Inf. Proc. Sys. Montréal, Canada. 2015. 91-99.
[8] Girshick R. Fast R-CNN. IEEE Inter. Conf. Com. Vis. Santiago, Chile. 2015. 1440-1448
[9] Bochkovskiy A, Wang C-Y, Mark-Liao H-Y. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934, 2020.
[10] Liu S, Qi L, Qin H, et al. Path Aggregation Network for Instance Segmentation IEEE Conf. Com. Vis. Pat. Rec., 2018: 8759-8768.
[11] Arthur D,Vassilvitskii S. k-means 1+:the advantages of careful seeding. Symp. Disc. Algo., 2007, 1027-1035.
[12] Everingham M, Gool LV, Williams C K I, et al. The Pascal Visual Object Classes (VOC) Challenge. Inter. J. Com. Vis., 2010, 88(2):303-338.
[13] Jocher Glenn. YOLOv5 release v6.1. https://github.com/ultralytics/yolov5 /releases/tag/v6.1, 2022. 2,7,10.
[14] 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. arXiv preprint arXiv: 2207.02696, 2022.
[15] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.