Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/39/20230585

A special target detection called You Only Looks Once

Liyun Li

Artificial Intelligence, School of Engineering Science, Shanghai, 200336, China

028121395@sues.edu.cn

Abstract. This paper proposes a unique target detection called You Only Look Once (YOLO for short) and a recognition method. Unlike the previous idea of many classifiers with object detection functions, the object detection box is set as a spatially separated bounding box and the regression problem of related class probabilities is realized. The neural network model can directly scan the entire image during testing, predicting bounding boxes and class probabilities from the complete picture. At the same time, because the whole detection channel relies on a single neural network, it is more straightforward and concise when upgrading and updating. The unified architecture used in this paper is fast, with smaller versions of the YOLO model processing a staggering 155 frames per second. At the same time, it has also achieved excellent results in mAP. Compared with other detection systems, YOLOv3 has optimized many past problems, including positioning errors. At the same time, the probability of predicting false detections in the absence of false detections is small. Finally, like the YOLO base model, YOLOv3 may produce significant errors when processing abstract works of art and images with a large number of small objects. However, its actual results are still better than detection methods such as Region -Convolutional Neural Networks (R-CNN).

Keywords: Convolutional Neural Network, YOLO, Target Detection.

1. Introduction

In the 1984 movie “Terminator”, the bionic mechanical warrior T800 has a lot of particular logic in terms of action, different from the conventional human way of turning the head. T800 has a large number of horizontal head cameras, this way the same as the way humans rotate the camera when playing Frames Per Second (FPS) games through the controller. Even if there was no 3D game that year, the production team still showed through this detail that T800 is not a human but a machine in human skin.

On the other hand, the human eye is much more efficient than machines. Humans can identify the contents and location of an image simply by looking at it. Just by looking at a picture, humans can know what’s in the image and where it is. And they can figure out how they interact. This demonstrates the human visual system is quick and accurate at item recognition. This paper can instinctively complete some logically complex tasks, such as driving a vehicle, with little to no consciousness. The computer does not have this particular sensor, so it needs additional auxiliary equipment to complete a similar task requiring a fast and accurate object detection algorithm. Such assistive devices can transmit real-time information about scenarios and unlock the potential of versatile, responsive robotic systems.

Among many image recognition systems, You Only Look Once (YOLO) is an exceptional category that differs from traditional picture cutting and object-by-object recognition. The design idea of YOLO is to restore the human visual mode, and the entire image can be seen entirely in training and testing,

DOI: 10.54254/2755-2721/39/20230585

just like a person opens his eyes to see the world. It's like reading the entire paragraph in its entirety, rather than breaking the paragraph into separate sentences [1].

In principle, YOLO is a very simple model. YOLO doesn't have complex construction. It only uses a single neural network. With a single neural network, YOLO can simultaneously predict multiple bounding boxes in the entire image, as well as the class probabilities corresponding to those bounding boxes. Compared with traditional object detection methods, YOLO trains on a complete image instead of cutting the image into several parts for separate detection. Meanwhile, YOLO can directly optimize and check its performance. Compared with other target detection methods, this unified model has several advantages.

The first one is efficiency, YOLO uses very little time in the test process, or it is extremely fast. Because in the testing process, this article defines detection as a regression problem, so there is no need for complex pipelines or multiple inspections, testing in this way can save a lot of time, and the efficiency is very cost-effective. This work is simply running a neural network on a new image while testing the predicted results before.

In addition, YOLO has also achieved good results in accuracy, from the data point of view, YOLO in real-time analysis and recognition of the average accuracy is almost close to more than twice other systems, it can be said that although it is not the highest, but higher than its accuracy is inferior to it in speed or far more complex, faster than its accuracy is not as good as it. Since YOLO can see the entire picture, it doesn't create some problems caused by not being able to see the context. The visual comparison is that other systems will come up with the result of cats and dogs falling from the sky when translating the sentence, it rains cats and dogs, while YOLO will come to the conclusion: it rains heavily. Fast R-CNN is an excellent image detection method, but due to the inability to see the complete context, it can mistake the background elements in the image as the detection object, and also treat a component of the detection object as the overall object. For example, when translating a sentence 'It rains cats and dogs', it will draw the conclusion of cats and dogs fall from sky which based on the segmentation of the word meaning, rather than arriving at the correct translation. However, because YOLO analyses by observing the entire image, although YOLO can quickly analyse the image, the accuracy rate when analysing small items, especially when multiple items overlap, is not very high, because the overlap of multiple items will lead to abnormal judgment weights [2].

In addition, YOLO is much more usable than other models because it analyses the entire image at the same time, and it is better than top-of-the-line inspection methods such as Deformable Part Model (DPM) and R-CNN when processing artworks, especially paintings. There will be no misunderstandings due to incorrect segmentation or separation. At the same time, during the test, YOLO showed a high degree of plasticity, in other words, it has strong compatibility and versatility. Therefore, when applied to a new domain, or accidentally entering the wrong content, it is unlikely that it will directly crash and cause a serious error.

2. Main Construction

Yolo3 has a total of 106 layers, drawing on the Deconvolutional Single Shot Detector (DSSD)'s deconvolution expansion and multi-layer Feature Map predictive structure design, that is, each image is detected on 3 different scale layers, which are the 82nd, 94th, and 106th layers, so as to realize the detection of large, medium and small targets. In order to facilitate the following explanation and explanation, this article prepares images to visually show the operation logic of YOLO, Figure 1 below shows its construction:

DOI: 10.54254/2755-2721/39/20230585

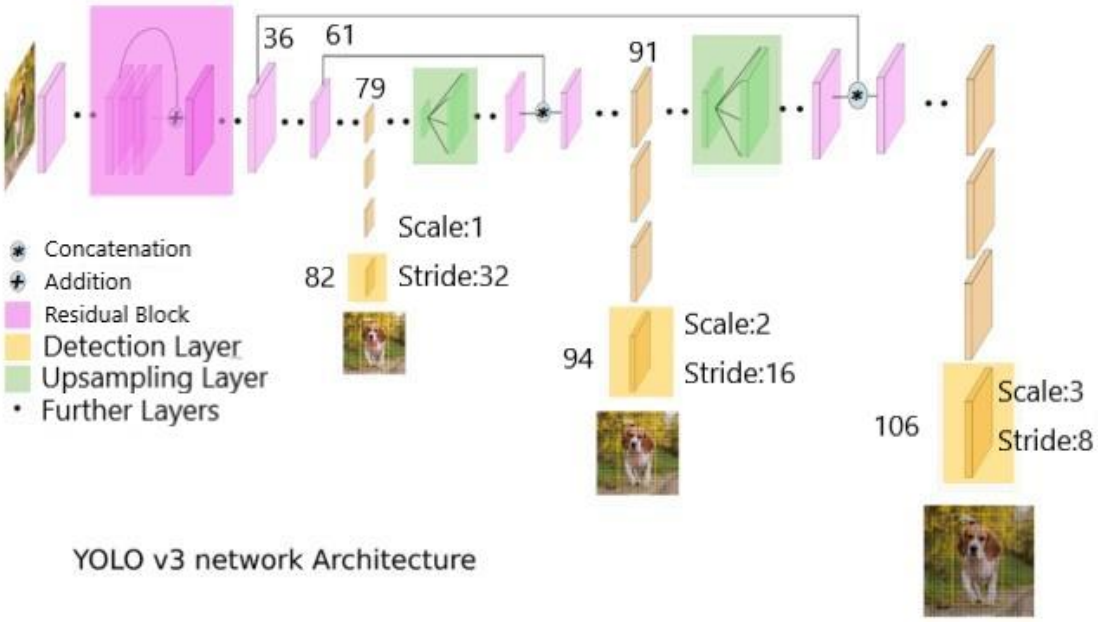

Figure 1. YOLO v3 Main Construction.

2.1. Residual Block (RB)

Residual Block is responsible for solving the problems of gradient vanishing and gradient explosion throughout the entire structure, ensuring the good performance of the model as a whole, and can also train deeper neural networks. For ease of understanding and presentation, this paper draws a structural diagram, the Figure 2 below shows the structure of Residual Block [3].

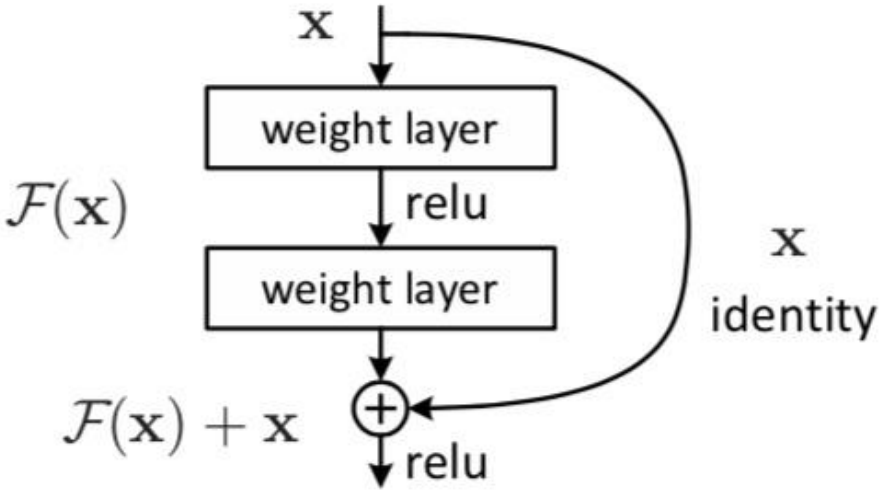

Figure 2. Residual Block.

2.2. Up sampling Layer

YOLO3 up sampling uses bilinear interpolation, see Bilinear Interpolation in Faster Region - Convolutional Neural Networks (Faster R-CNN) 2.3.2 Rol Alignment.

DOI: 10.54254/2755-2721/39/20230585

2.3. Detection Layer

It is equivalent to the layer formed after prediction of the Feature Map layer in Faster R-CNN and Single Shot Detector (SSD).

2.4. FPN (Feature Pyramid Network)

The image pyramid is shown in Figure 3, which is to resize the image to different sizes, and then obtain the features of the corresponding size respectively, and then make predictions [4].

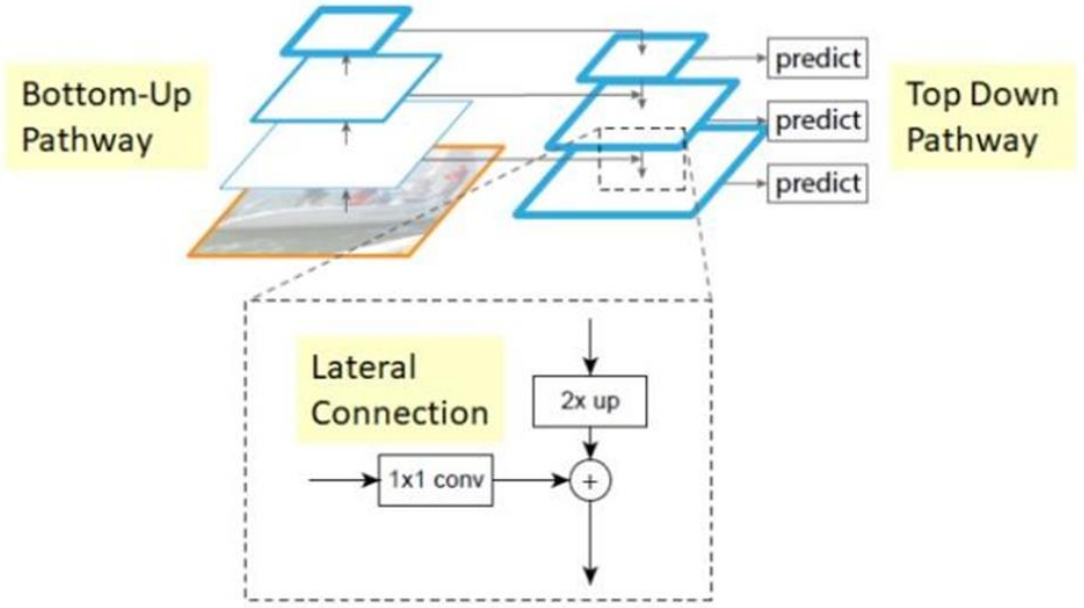

Figure 3. Feature Pyramid Network.

3. Target Detection

Taking the first detection layer (i.e., layer 82) as an example, the size of the original image is 416 X 416, and the cumulative Stride reaching the layer is 32, the size of the detection layer is 13 X 13 (that is, 416/32=13), which is also equivalent to dividing the original image into a 13 X 13 grid. The original image, whose shape = [416, 416, 3], is formed by a deep learning convolutional neural network to form the first detection layer, whose shape = [13, 13, 255]. The red grid in the right figure (i.e., the detection layer) in the figure below has 255 values, indicating the values predicted by the three rectangles of yellow, purple and blue in the left figure (i.e., the original image) (all three are red boxes in the centre). For ease of understanding and presentation, this paper draws a structural diagram, the Figure 4 below shows the structure of the first detection layer.

DOI: 10.54254/2755-2721/39/20230585

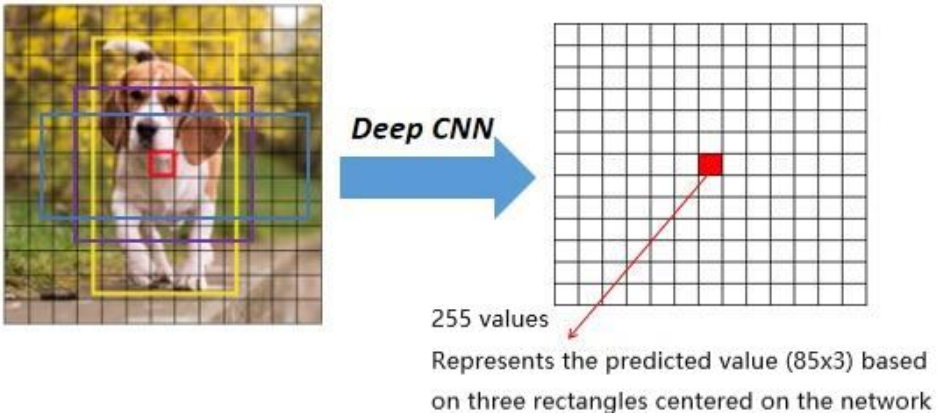

Figure 4. The first detection layer.

The 255 values are grouped according to their corresponding rectangles as figure5 shows:

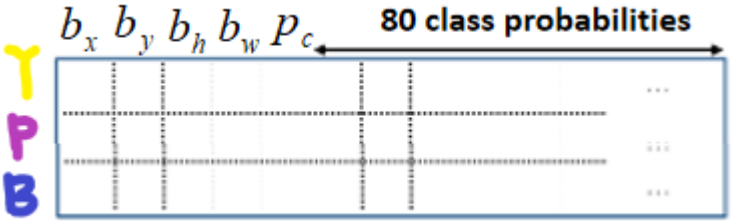

Figure 5. Corresponding rectangles.

Where b_x, b_y, b_h, b_w: the relative position coordinate information of the target; p_c: probability of having a target; The last 80 values are the predicted probabilities of each of the 80 targets.

By merging the above, the needy data about the final prediction will be easy to get, and in order to visualize its operation process, this paper chooses to show it in the form of picture, as Figure 6 shows:

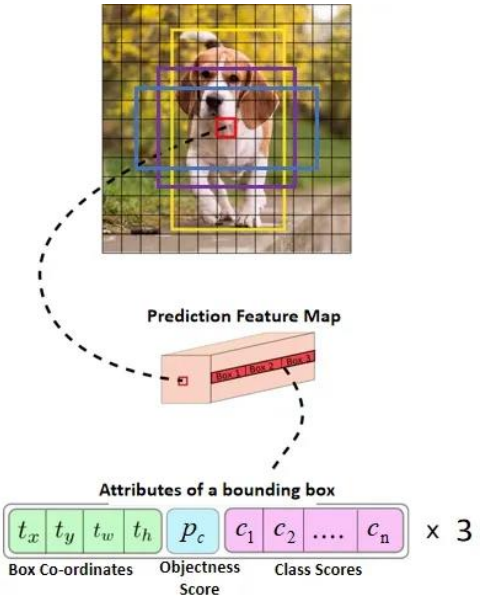

Figure 6. Predicted probabilities.

DOI: 10.54254/2755-2721/39/20230585

In this paper, the yellow rectangle and its corresponding 85 values are selected as an example, and the final prediction result is calculated as shown in Figure 7:

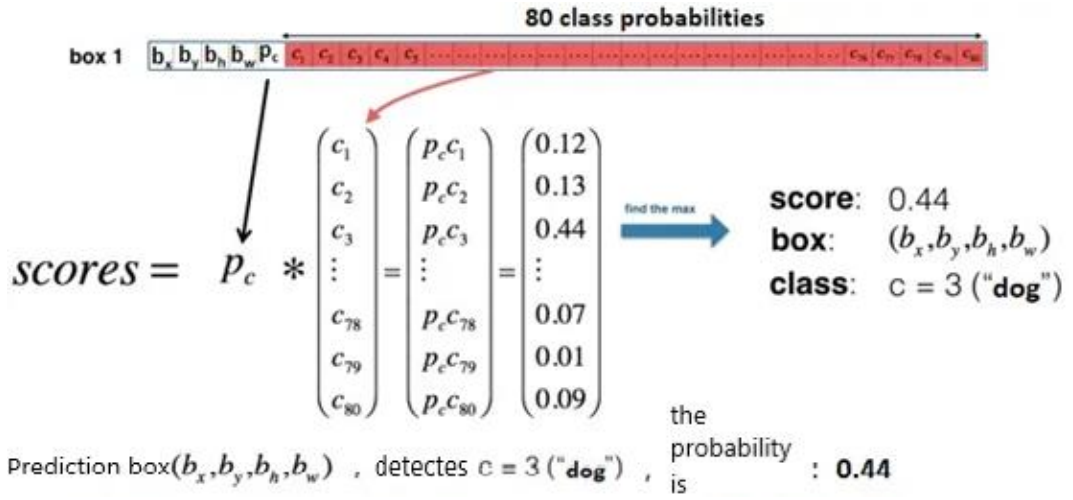

Figure 7. Final prediction.

Location information values b_x, b_y, b_h, b_w values: indicate the position of the prediction box, and its value is not directly given by the convolutional network, but converted as follows:

$$b_x = \sigma(t_x) + c_x \quad (1)$$

$$b_y = \sigma(t_y) + c_y \quad (2)$$

$$b_w = p_w * e^{t_w} \quad (3)$$

$$b_h = p_h * e^{t_h} \quad (4)$$

Where b_x, b_y, b_h, b_w represent the center coordinate values x,y of the predicted target position rectangle box x,y and the length and width values of the rectangle; t_x, t_y, t_h, t_w is the value given directly by the convolutional network, while c_x, c_y,refers to the coordinates of the point in the upper left corner of the small red box in the detection layer, and the c_x and c_y of the example small red box are (6,6) [5].

Target probability value p_c: indicates the probability that the prediction box contains the target, and the activation function used is sigmoid;

Target category probability$c_1 \sim c_{80}$: indicates the probability of each class, in versions before yolo3, the activation function used is soft max, and yolo3 modifies it to sigmoid function; The reason is that soft max is exclusive, that is, the detected object can only belong to a certain target class, if the target class has an inclusion and inclusion relationship class, such as the target class contains the Person class, Man class and Woman class, soft max is not applicable, and sigmoid is more suitable for this situation.

The above is the prediction of a rectangular box of a grid, if the input picture scene is more complex, all the rectangular box prediction results of all meshes of the entire detection layer are visualized as figure8:

DOI: 10.54254/2755-2721/39/20230585

Figure 8. Detection layer visualization.

It can be seen directly that YOLO can completely judge cars, roads, traffic lights, billboards and street lights during the recognition process. This effectively avoids situations where the car is recognized as a tire, piece of metal, a metal case, and a street lamp as a light bulb and metal due to the position of the segmented image.

This is a prediction of a detection layer, and yolo3 uses a total of three detection layers, as the figure9 shows:

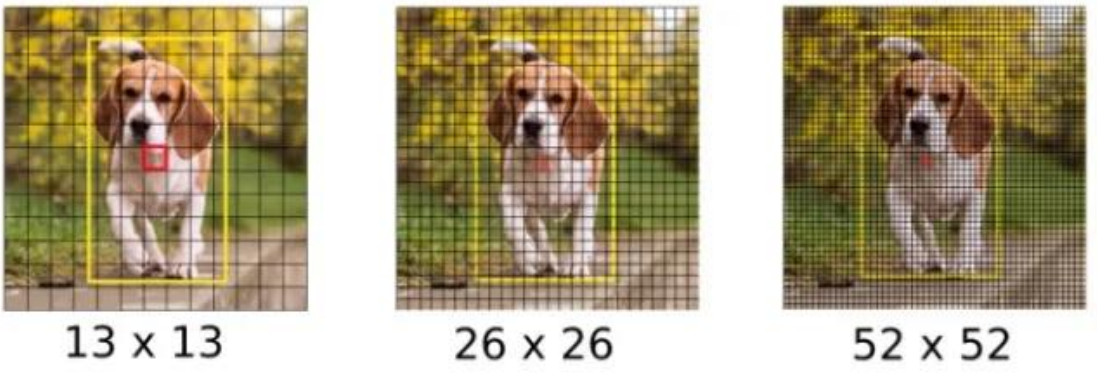

Figure 9. Three detection layers.

Each detection layer independently makes the above predictions; The difference lies in the different target sizes, and the target sizes of the three detection layers from left to right are large, medium, and small, that is, 3 large rectangular boxes, 3 medium rectangular boxes and 3 small rectangular boxes; These 9 rectangular boxes are K-means clustering acquisition for all sizes of objects on a specific dataset, such as on the dataset COCO, the final size of these 9 rectangular boxes is as follows:

DOI: 10.54254/2755-2721/39/20230585

(373, 326), (156, 198), (116, 90): applied to the first detection layer 13 X 13;
(59, 119), (62, 45), (30, 61): applied to the second detection layer 26 X 26;
(33, 23), (16, 30), (10, 13): applied to the third detection layer 52 X 52;
In this way, 9 rectangular boxes of different sizes can be used on the three detection layers to better identify large and small targets on the image. In addition, NMS is also used in the same detection layer.

4. Target detection

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{i,j}^{obj}\left[\left(b_x-\widehat{b_x}\right)^2+\left(b_y-\widehat{b_y}\right)^2+\left(b_w-\widehat{b_w}\right)^2+\left(b_h-\widehat{b_h}\right)^2\right]+$$

$$\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{i,j}^{obj}[-log(p_c)+\sum_{i=1}^{n}BCE\left(\widehat{c_i},c_i\right)]+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{i,j}^{noobj}[-log(1-p_c)] \quad (5)$$

Thereinto:
S: the number of grids, which means S^2 is 13*13,26*26,52*52:
B: box;
$1_{i,j}^{obj}$: If Box has a target, its value is 1, otherwise it is 0;
BCE(binary cross entropy):

$$BCE(\widehat{c_i},c_i)=-\widehat{c_i},*log(c_i)-(1-\widehat{c_i})*log\ (1-c_i) \quad (6)$$

$1_{i,j}^{noobj}$: If Box has no target, its value is 1, otherwise it is 0 [6].

5. DC-SPP-YOLO (Dense Connection and Spatial Pyramid Pooling Based YOLO for Object Detection)

Convolutional neural networks are used in this study's model implementation, and the PASCAL VOC detection dataset is used to test and assess the model's performance. The Google Net model, which is also used for picture categorization, served as the basis for the network architecture that was used this time. The convolutional neural network comprises two fully linked layers and a total of 24 convolutional layers. The convolutional layers in a convolutional neural network each have different roles to play. For example, the first convolutional layer in a network is in charge of extracting the features of the image's objects from the entire image, while the second fully connected layer is in charge of producing the recognition probability and its corresponding coordinates [7]. This is compatible with the discussion from earlier.

This is consistent with the previous discussion. Unlike the previous YOLO, this paper also adds a DC (Dense Connection) layer and an SPP (Improved Spatial Pyramid Pooling) layer before the detection layer [8]. Each component of the arrangement can be retrieved with ease by separating it. In order to better understand, let's start by noting that the network's predicted value is a two-dimensional tensor P with the shape [batch,7x7x30]. Slice-wise, P ([:,0:7*7*20]) represents the category probability, P([:,7*7*20:7*7*(20+2)] represents the confidence component, and P([:,7*7*(20+2):] represents the prediction outcome of the bounding box. This is congruent with what was said earlier. The DC (Dense Connection) layer and the SPP (Improved Spatial Pyramid Pooling) layer are also added in this study, in contrast to the earlier YOLO [9], before the detection layer.

Dense Net is a deep convolutional neural network that enhances feature reuse and gradient flow by introducing dense connections in the network, thereby improving the performance and generalization ability of the model. In Dense Net, each layer takes the output of all previous layers as its input, forming a dense connection structure. Compared to Res Net, Dense Net places more emphasis on feature reuse and information sharing, which may result in a slight loss in computational efficiency, but typically performs well in model accuracy and generalization ability. In computer vision applications including picture classification, object recognition, and semantic segmentation, dense networks are frequently

DOI: 10.54254/2755-2721/39/20230585

utilized. For ease of understanding and presentation, this paper draws a structural diagram, the Figure 10 below shows the structure of DC layer:

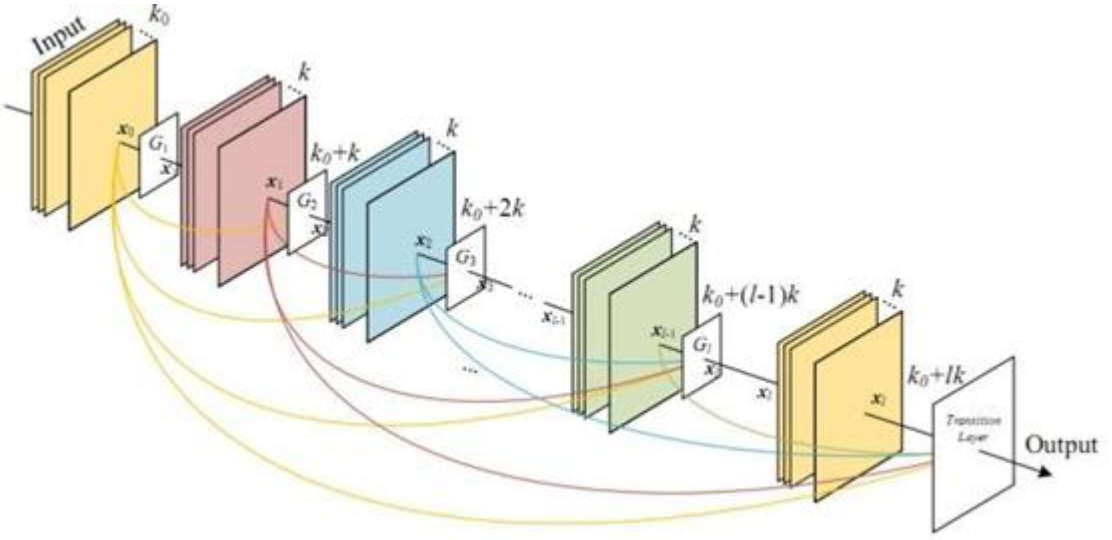

Figure 10. DC layer.

In the model, the SPP layer enables CNN Pres of different sizes to output vectors of fixed size. For ease of understanding and presentation, this paper draws a structural diagram, the Figure 11 below shows the structure of SPP layer:

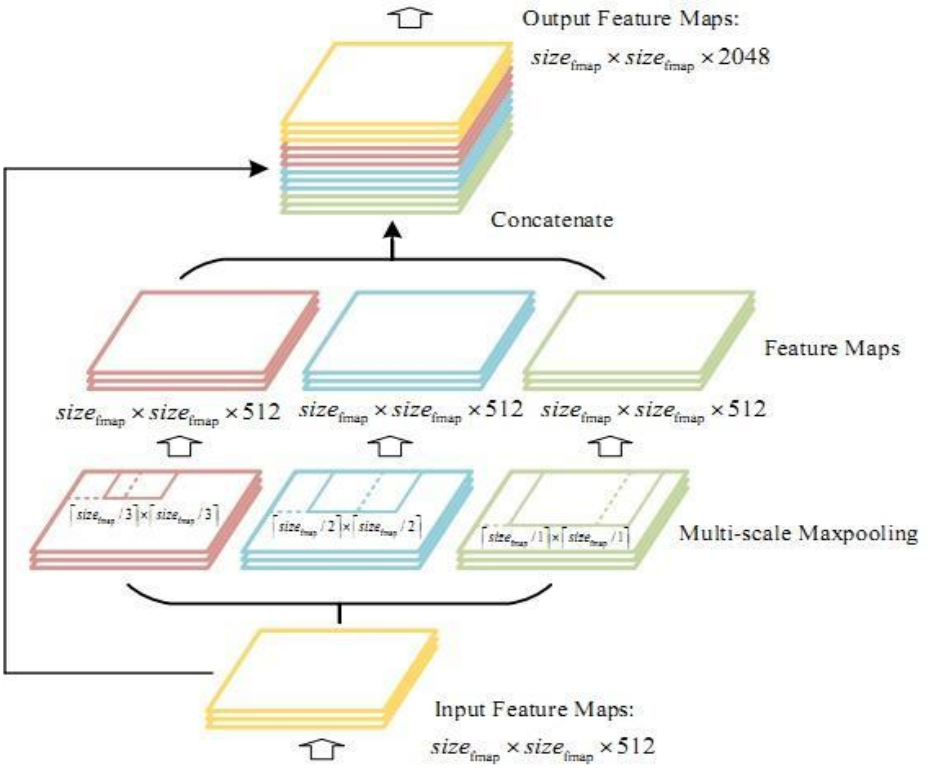

Figure 11. SPP layer.

DOI: 10.54254/2755-2721/39/20230585

After explaining the functions of each layer in the entire model, it's time to fully show the composition of the entire model. In order to facilitate understanding and visual display, this article chooses to present it through pictures. The Figure 12 below shows the structure of the DC-SPP-YOLO Model:

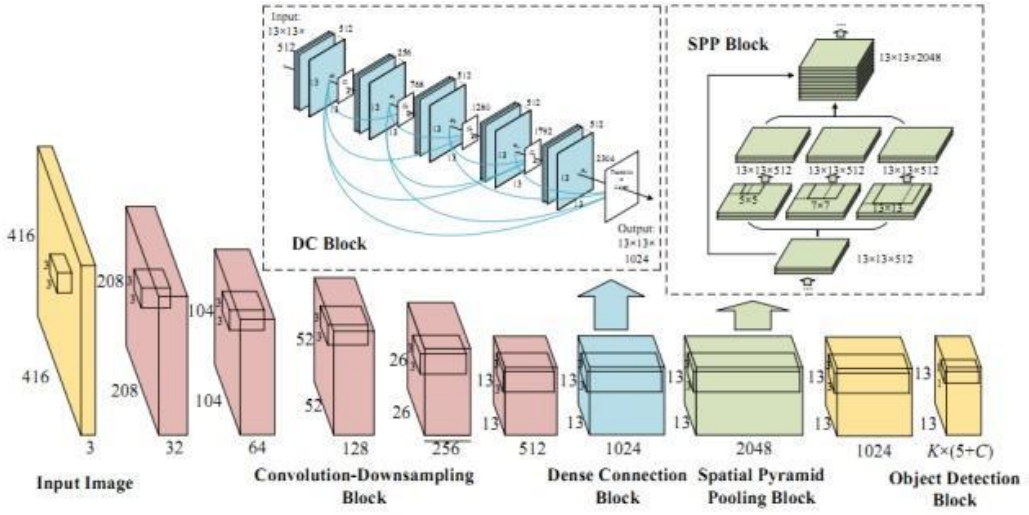

Figure 12. The DC-SPP-YOLO Model.

6. Comparison to Other Real-Time Systems

Since the entire image can be seen in its entirety, YOLO can avoid some misjudgements caused by miss mentation. At the same time, for images in the background, segmentation can lead to misunderstanding, such as listing the wheels of the car separately while ignoring the entire vehicle itself. Therefore, YOLO is better than models such as Fast R-CNN in terms of background testing. Its performance has been significantly improved by eliminating Fast R-CNN's background detection by using YOLO. For each bounding box predicted by R-CNN, this article checks whether YOLO predicts similar boxes. This paper improves the prediction based on the probability of the YOLO prediction and the overlap between the two boxes [10].

Table 1. The Quantitative results on the VOC 2012 test.

VOC 2012 test	mAP	human	car	AP
YOLO	68.9	73.9	75	69.2
Fast R-CNN + YOLO	80.7	81.2	83	80.2
Fast R-CNN	68.4	72.3	71	69.6

After testing in this paper, the Fast R-CNN model achieved an optimal value of 71% for mAP on the VOC 2007 test set. After combining with YOLO, mAP achieved an increase of almost 3.2%, reaching 75%. At the same time, this article also attempted to combine R-CNN with several other versions of the model, and the mAP of these models increased slightly between 3% and 6%.

7. The result of Real-Time Detection Test

As an efficient and accurate object detector, YOLO is connected to a webcam to verify its ability to maintain real-time performance. This includes the images it obtains from the camera, the results of its recognition and detection, and the time it takes for detection. According to the test results, YOLO is currently the ideal choice for practical applications of computer vision.

In the test process, this paper selected VOC 2012 test as the test dataset, and the results of different recognizers are compared as the Table 1 shows.

DOI: 10.54254/2755-2721/39/20230585

8. Conclusion

This paper introduces YOLO, a unified object detection model. The model structure shown in this paper is not complex, and object recognition does not take up too much resources and time. YOLO can directly perform object recognition and detection on the complete image, which restores how human vision is generated as much as possible. YOLO, unlike typical classifier-based detection systems, is trained on a loss function, allowing it to directly correlate to detection performance.

In the course of this experiment, YOLO was not effective at identifying abstract images because it could not accurately analyse the true meaning behind abstract elements, such as the colour blocks of abstract paintings may refer to people. At the same time, since YOLO analyses by identifying the entire image, the accuracy rate will be seriously reduced when processing many small objects in the image, because multiple objects appear in the same box at the same time.

Future related research will mainly include the following aspects:

In future research, this paper will further increase the accuracy and speed of YOLO in recognizing objects by optimizing the algorithm and optimizing the neural network structure.

References

[1] Zhang, S., Chai, L., and Jin, L., 2020. Vehicle detection in UAV aerial images based on improved yolov3. 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC).

[2] Zhang, S., Wu, Q., Su, P., and Ma, J., 2022. Design and simulation investigation of 3*9 led matrix headlamp to realize anti-glare function. Advances in Transdisciplinary Engineering.

[3] Chen, S., Zhao, S., and Lan, Q., 2022. Residual block based nested U-type architecture for Multi-modal brain tumor image segmentation. Frontiers in Neuroscience, 16.

[4] Shakil A.*, S., and Kureshi, Dr.A., 2020. Object detection and tracking using YOLO V3 framework for increased Resolution Video. International Journal of Innovative Technology and Exploring Engineering, 9(6), pp.118–125.

[5] Zhang, S., Wu, Q., Su, P., and Ma, J., 2022. Design and simulation investigation of 3*9 led matrix headlamp to realize anti-glare function. Advances in Transdisciplinary Engineering.

[6] Huang, Z., Zhang, P., Liu, R., and Li, D., 2021. Immature apple detection method based on improved yolov3. ASP Transactions on Internet of Things, 1(1), pp.9–13.

[7] Auliya, A., Pradani, W., and Haryanto, T., 2022. Kenaf flower detection using yolov3. 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS).

[8] Li, S., Li, Y., and Lu, Y., 2020. Comparative optimization and application of rectification schemes for shear wall structure. IOP Conference Series: Earth and Environmental Science, 455(1), p.012023.

[9] Alisetti, S.N., Purushotham, S., and Mahtani, L., 2019. Guiding and navigation for the blind using deep convolutional neural network based predictive object tracking. International Journal of Engineering and Advanced Technology, 9(1s3), pp.306–313.

[10] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 2016. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).