Application Research of YOLOv3 Incorporating Self-attention Mechanism

Jiaxin Zhang, Yinshan Jia(✉), and Hongfei Yu

School of Computer and Communication Engineering, Liaoning Petrochemical University,
Fushun 113001, Liaoning, China
yinshanjia@163.com

Abstract. Improving the accuracy of target recognition is always the focus of machine learning research. YOLOv3 multi-scale detection accelerates the speed while ensuring accuracy, and the self-attention mechanism takes into account the attention weight of each pixel feature to enhance the ability of information extraction. In view of the large size difference between different targets in the target detection task, which makes it difficult to effectively detect multi-size targets, and the publication of the latest self-attention mechanism COT.NET, a combination of the YOLOv3 network Darknet-53 and the self-attention network COT.NET is proposed. The idea of YOLOv3 is improved by adding a self-attention mechanism to the residual structure of YOLOv3.Through verification on the VOC image set, the improved YOLOv3 is 1.34% higher than the original YOLOv3 map. Experimental results show that YOLOv3 integrated into the self-attention mechanism can improve the accuracy of image recognition.

Keywords: Yolov3 · Self attention mechanism · Residual structure · COT.NET

1 Introduction

Deep learning is currently one of the most important technologies in the field of artificial intelligence, and YOLO is a popular deep learning neural network. As a classic of the YOLO series, YOLOv3 has excellent detection results in traffic road markings [1], airport runway foreign body detection [2], cancer tumor detection [3], face recognition and gesture recognition [4]. YOLOv3 is a multi-scale fusion prediction framework, each scale contains 3 a priori boxes to be predicted, and each box will get 9 cluster centers, which are equally divided into 3 different scales according to their size [5–9]. The backbone network is the Darknet 53 structure. The key features are extracted by increasing the number of channels through convolution to obtain a better hierarchical structure. The three feature layers finally obtained are used for upsampling and fusion respectively. The three feature layers after fusion are for large objects, Medium and small objects are detected and predicted respectively, and then a priori frame decoding adjustment and non-maximum suppression are performed to determine the best prediction frame [10, 11].

Y. S. Shmaliy and A. Abdelnaby Zekry (Eds.): CCIE 2021, LNEE 920, pp. 454–461, 2022.
https://doi.org/10.1007/978-981-19-3927-3_45

Self-Attention (Self-Attention) mechanism was first used in natural language processing. In recent years, it has shown its head and feet in image recognition due to its excellent ability [12]. When considering the local features of pixels, pay attention to those pixels that have a greater impact on it. This feature gives image recognition a new research direction. COT.NET is the newly proposed self-attention mechanism structure network, which enhances the self-attention mechanism through the guidance of local context modeling [13]. YOLOv3 relies on a multi-scale fusion mechanism to ensure the accuracy of recognizing objects. The attention mechanism considers the weight of each pixel feature and strengthens the extraction of information. Integrating the self-attention mechanism into YOLOv3 may improve the accuracy of recognition, which is a problem worthy of study.

2 YOLOv3 Network

2.1 The Basic Idea of YOLOv3 Backbone Network

The basic network structure of YOLOv3 is Darknet-53, which consists of 5 residual blocks, used to compress the size and increase the number of channels. Taking the input image 416 × 416 × 3 as an example, the image width and height are compressed through convolution, and the number of channels is pulled to the specified dimension. With reference to the FPN pyramid structure, the input image 416 × 416 × 3 is convolved through the backbone network to get the last three outputs. Combining the key information of the setting frame, three characteristic layers of y1, y2, and y3 are obtained, the sizes of which are 13 × 13 × 1024, 26 × 26 × 512, 52 × 52 × 256, respectively. Fuse these three feature layers to judge and recognize objects separately. The backbone network structure is shown in Fig. 1.

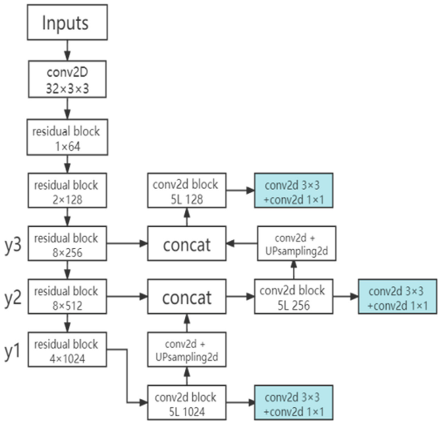


Fig. 1. Yolov3 backbone network structure.

2.2 YOLOv3 Decoding and a Priori Frame Adjustment

The three feature layers of YOLOv3 are 13 × 13 × 1024, 26 × 26 × 512, 52 × 52 × 256, and each grid point is responsible for the detection of a region. YOLOv3 uses the Anchor Based structure. The anchor points are selected in the picture with a certain step length, and multiple frames with fixed height and width are designed with each anchor point as the center, and each frame is regenerated in the image Three a priori boxes of different sizes and different aspect ratios. Three kinds of frames with different aspect ratios are decoded and calculated to obtain the final bounding box. The bounding box coordinates are bx and by, and the width and height are bw and bh. The calculation process is as follows.

$$\begin{aligned} b_y &= \sigma\left(t_y\right) + c_y, \quad b_h = p_h e^{t_h} \\ b_x &= \sigma\left(t_x\right) + c_x, \quad b_w = p_w e^{t_w} \end{aligned} \tag{1}$$

where c_x and c_y are the number of horizontal and vertical grids from the upper left corner of the grid where the point is located from the origin of the upper left corner, respectively. p_w and p_h are the side lengths of the a priori box, and t_x and t_y are the horizontal and vertical offsets of the upper left corner of the grid where the point is located relative to the target center. t_w and t_h are the width and height of the predicted frame respectively, and σ is the *Sigmoid* activation function. In the past, the activation function used by the YOLO series was *Softmax*, and it was improved afterwards. The *Sigmoid* function was used to prevent the output value from jumping. The decoding process is shown in Fig. 2.

2.3 IOU Non-maximum Suppression

IOU (Intersection over Union) non-maximum suppression is the best prediction frame obtained by calculation. Take out the box with a score greater than a given threshold, and determine the most suitable box by comparing the size of the IOU. IOU is the ratio of the intersection of the predicted frame and the real frame to the union. The principle of IOU calculation is shown in Fig. 2.

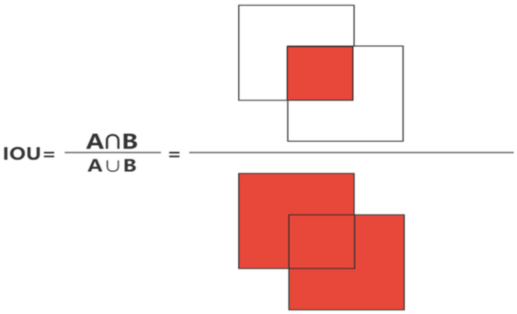


Fig. 2. Principle of IOU calculation.

3 COT.NET Self-attention Mechanism

3.1 COT.NET Principle

The self-attention mechanism was first used in natural language processing and achieved extraordinary results, and then applied to image processing. In order to improve accuracy, people continue to improve the existing self-attention mechanism. Most existing self-attention mechanisms such as Transformer directly focus on the two-dimensional feature map, perform Self-Attention operations, and obtain the attention matrix based on the query and key of each spatial position, but there is no contextual information between adjacent keys. Is fully utilized. COT.NET combines the existing advantages to design a new attention structure COT block, which makes full use of the key context information to guide the learning of the dynamic attention matrix, which improves the visual expression ability. The COT block structure is shown in Fig. 3.

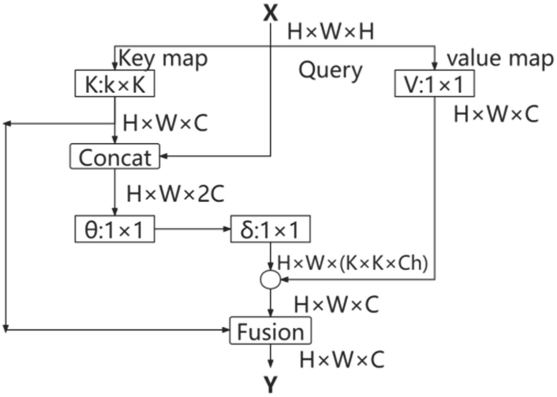


Fig. 3. COT.net structure.

3.2 Combination of YOLOv3 Residual Network and Self-attention Mechanism

Aiming at the insufficient ability to extract information in YOLOv3 multi-target size, resulting in poor target detection accuracy, the newly proposed self-attention mechanism COT.NET function is added to the residual network structure of YOLOv3, and the self-attention mechanism is used to improve information extraction Ability to enhance the semantic strength of features and improve the accuracy of target positioning.

The residual network uses a Shortcut to open up a highway between non-adjacent network layers [14]. That is to say, the original network output result is F(x), The current output must be added to the previous input x. Such a structure is a residual network, which has both convolution to deepen the network part and constant input part to prevent overfitting. The residual network avoids the vanishing gradient and degradation problems caused by the excessive depth of the network. The residual network structure is shown in Fig. 4.

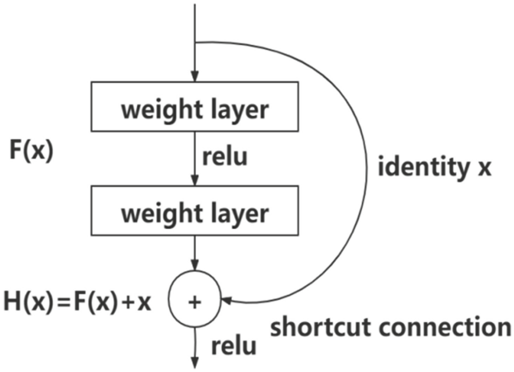


Fig. 4. Residual network structure.

In order to solve the lack of long-distance modeling ability of CNN structure. 3 × 3 convolution is used to model static context information, and then query and context information modeling are fused, and then two consecutive 1 × 1 convolutions are used to generate dynamic context, and the static and dynamic context information is finally merged To the output. That is, the local information is extracted through convolution first, so that the static context information inside the key is fully explored. This structure aggregates the mining of contextual information and the learning of Self-Attention into one structure, making the self-attention mechanism more effective.

By integrating the COT.NET network structure into the residual network part, the self-attention mechanism is added to YOLOv3. Incorporating static and dynamic context information, enhancing the ability to extract local important information, making YOLOv3 more complete and accurate. On the one hand, the input x uses the self-attention mechanism COT.NET network to strengthen the extraction of information, on the other hand, the output is added to the previous input x to prevent over-fitting, and this network structure is used to replace the previous residual network structure. The combined structure of the residual network and the self-attention mechanism is shown in Fig. 5.

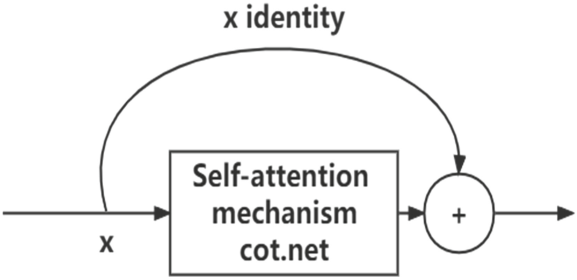


Fig. 5. Combination of residual network and self-attention mechanism.

4 Experiment

4.1 Experimental Data

The experiment uses the pascal VOC data set [15]. The experimental data is divided into a training data set and a test data set. The training data set includes the training part of VOC2007 and all the data of VOC2012. The test data set uses the test part of VOC2007.

4.2 Experimental Process and Experimental Results

In the experiment, the training efficiency is improved by freezing and thawing training. During the training process, the number of images (Batch Size) of the frozen part of the network is 4, and the number of images of the training part after thawing is 8.

The experimental results are shown in Table 1.

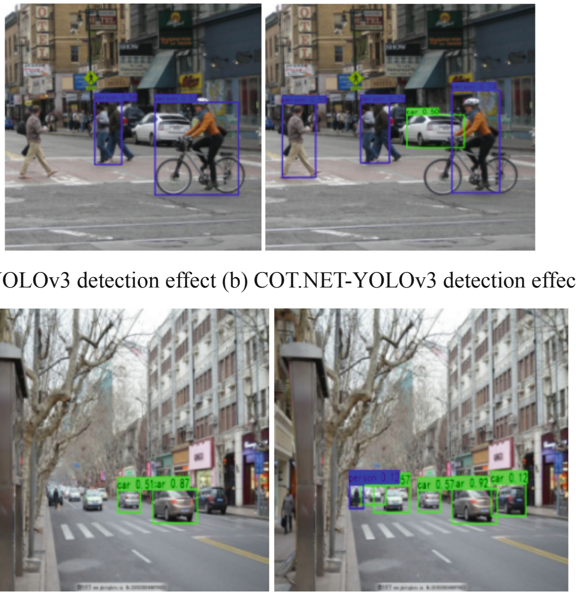
Table 1. Comparison of YOLOv3 and improved YOLOv3.

Experiment number	Model network	Accuracy (%)
A	YOLOv3	75.72
B	COT.NET-YOLOv3	77.06

The experiment numbered A in Table 1 is the original YOLOv3 experimental result on the pascal VOC data set, and the experiment numbered B is the experimental result after YOLOv3 is integrated into the self-attention mechanism COT.NET. It can be seen from the experimental results that the experimental accuracy of the improved YOLOv3 is 77.06%, which is 1.34% higher than the original Yolov3. It can be concluded that integrating the self-attention mechanism helps to improve the accuracy of target detection.

4.3 Multi-size Target Detection Test

Figure 6(a) and (c) are the detection results of the YOLOv3 instance, (b) and (d) are the detection results of the improved instance integrated with the self-attention mechanism. Through the comparison, it can be found that the Fig. 6(a) does not detect the object blocked behind, and the Fig. 6(b) successfully detects the car and the person on the left. Figure 6(c) compares with Fig. 6(d), Fig. 6(d) has more detected the person on the left and the two cars in the middle.



(a) YOLOv3 detection effect (b) COT.NET-YOLOv3 detection effect

(b) YOLOv3 detection effect (d) COT.NET-YOLOv3 detection effect

Fig. 6. Comparison of yolov3 and improved Yolov3.

5 Conclusion

This article takes the network structure of YOLOv3 as the backbone, analyzes the relationship and principle between the backbone network and COT.NET, modifies the residual network of YOLOv3 and incorporates the self-attention mechanism, through freezing and thawing training, multi-scale target detection and comparison, it is effective It verifies the role of COT.NET structure in YOLOv3, and proves that the self-attention mechanism can improve the accuracy of image target detection.

References

1. Jin, L.S., Guo, B.C., Wang, F.R., Shi, J.: Dynamic multi-target detection algorithm in front of vehicles based on improved YOLOv3. J. Jilin Univ. (Eng. Technol. Edn.) **51**(04), 1427–1436 (2021)
2. Luo, S.J.: Research and application of YOLOv3 algorithm in the field of intelligent transportation. Lanzhou University, Lanzhou (2020)
3. Xu, L.F., Fu, Z.J., Mo, H.W.: Recognition of breast ultrasound tumor based on improved YOLOv3 algorithm. J. Intell. Syst. **16**(01), 21–29 (2021)
4. Diao, R., Hu, Y.L., Jiang, Y.Z., Lu, W.: Human body detection technology based on YOLOv3 improved algorithm. Inf. Technol. Inf. Technol. (08), 249–252 (2021)
5. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

6. Lin, T., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 936–944. IEEE Press, Washington D.C., USA (2017)
7. Joseph, R., Ali, F.: YOLOv3: an incremental improvement. Comput. Vis. Pattern Recogn. **2021**(01), 48–50 (2021)
8. Ge, Z., Liu, S.T., Wang, F., Li, Z.M., Sun, J.: YOLOX: exceeding YOLO series in 2021. ArXiv abs/2107.08430 (2021)
9. Fang, Y.X., et al.: You only look at one sequence: rethinking transformer in vision through object detection. ArXiv abs/2106.00666 (2021)
10. Tong, N., Lu, H.C., Zhang, L.H., Ruan, X.: Saliency detection with multi-scale superpixels. IEEE Signal Process. Lett. **21**(9), 1035–1039 (2014)
11. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. ArXiv abs/1804.02767 (2018)
12. Chefer, H.L., Shir, G., Lior, W.: Transformer interpretability beyond attention visualization. In: Proceeding of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, pp. 782–791 (2021)
13. Li, Y.H., Yao, T., Pan, Y.W., Mei, T.: Contextual transformer networks for visual recognition. ArXiv abs/2107.12292 (2021)
14. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Press, Washington D.C., USA (2016)
15. Pascal VOC Dataset Mirror. pjreddie.com