Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Pedestrian detection method based on improved YOLO v5

Wenhui Zhang, Zhihua Luo
Guilin University of Electronic Technology, China

Abstract: The flow and gathering of high-density people in public places are becoming more and more frequent. Using computer vision technology to analyze the video data in public places can obtain pedestrian information, and then complete the analysis of pedestrian flow in public scenes, so as to strengthen the management and safety of public scenes. YOLO v5 is a representative algorithm of target detection in the current stage. It mainly designs the network architecture based on convolution, but the network based on convolution has insufficient ability to extract and global features. Therefore, this paper improves its backbone structure combined with vision transformer. According to a large number of documents, vision transformer has redundancy in image coding in visual tasks. Therefore, this paper improves the token of vision transformer and combines the token based attention mechanism to make the backbone network better learn useful features. Experiments show that the improved method proposed in this paper has certain advantages.

Keywords: Pedestrian detection, YOLO v5, Vision Transformer.

1. Introduction

Pedestrian detection in key areas is a hot research issue in machine vision monitoring. Because pedestrians have the characteristics of both rigid and flexible objects, they can be used as a special moving target detection problem. The main feature of pedestrian detection is to use the methods of machine vision and video analysis to make the computer locate, identify and track the pedestrian target in the video, reduce the phenomenon of missed detection and false detection, and improve the accuracy and speed of detection. Target detection refers to the task of classifying and locating targets in images or videos. This task can be easily completed for us. However, it is difficult for computers to classify and locate any one of the targets. Target detection needs to identify and locate all instances contained in the object (such as cars, people, street signs, roads, obstacles, etc.) in the field of vision. In recent years, due to the wide application of target detection technology in various industries, it has been

concerned and studied by a large number of scholars. At present, target detection technology is mainly divided into two categories: one-stage target detection and two-stage target detection. The two-stage processing method first generates regional suggestion boxes of possible objects, and then further predicts these suggestion boxes. The one-stage processing method is to return the target area directly on the feature map and give the final prediction result.

Since girshick et al. [1] proposed r-cnn in 2014, pedestrian detection task has officially entered the deep learning stage. The detection methods based on deep learning are mainly divided into two categories. However, r-cnn has some disadvantages: (1) because each scheme is characterized by deep convolution network extraction (i.e. the calculation is not shared), it will lead to a large number of repeated calculations. Therefore, the training and testing of r-cnn is very time-consuming; (2) The whole detection framework cannot be optimized end-to-end. The three steps of r-cnn (suggestion generation, feature extraction and region classification) are independent components, so it is difficult to obtain the global optimal solution; (3) It is difficult to generate high-quality regional suggestions in complex environments because selective search depends on low-level visual cues. He et al. [2] proposed using spatial pyramid pool (SPP) layer to process images of any size or aspect ratio. Spp net only moves the convolution layer of CNN to the front of the area recommendation module and adds a pooling layer, so that the network does not depend on the size and aspect ratio of the input image and reduces the amount of calculation. Spp (spatial pyramid pooling) network is more efficient than r-cnn model and has considerable accuracy. In 2015, Redmon j et al. Proposed a one-stage target detection model Yolo (you only look once) [6]. Different from the previous RCNN series, Yolo uses a single convolutional neural network to change the target detection problem into a regression problem. When the picture is input into the detection network, the position coordinates and type information of the target can be obtained at the output layer of the network, which realizes end-to-end optimization and simplifies the processing flow[3,4,5]. Compared with the two-stage detection network, Yolo has fast speed and low error rate. However, Yolo still has some disadvantages, which are that each box can only predict one type of target. If multiple objects appear in one box at the same time, Yolo can only detect one type; Secondly, the positioning ability of Yolo needs to be strengthened, the recall rate is relatively low, and the detection effect of small targets and relatively dense targets is not good.

YOLO v5 is a representative algorithm of target detection in the current stage. It mainly designs the network architecture based on convolution, but the network based on convolution has insufficient ability to extract and global features. Therefore, this paper improves its backbone structure combined with vision transformer. A large

number of documents [7] show that vision transformer has redundancy in image coding in visual tasks. Therefore, this paper improves the token of vision transformer and combines the token based attention mechanism to make the backbone network better learn useful features [8, 9]. Experiments show that the improved method proposed in this paper is superior to other methods.

2. Related work

The development of target detection technology can be roughly divided into three stages. The method of the first stage is based on the traditional algorithm. The main steps are generating candidate regions, extracting features manually and classifying targets. The second stage is a two-stage target method based on deep learning, which mainly includes regional recommendation network and convolution network for extracting classification targets [10]. The third stage mainly refers to the end-to-end one-stage target detection method. The main method is Yolo (you only look once).

Compared with the previous yolov1 and yolov2, yolov3 [11] has a strong speed advantage. It uses a logical classifier to replace the softmax of the classifier layer for multi label tasks. Redmon J and others have greatly improved the accuracy of the model with the improved darknet-53 model. At the same time, they detect targets on multiple scales, enhance the recognition of small targets and overlapping occluded targets, and have a balanced performance in speed and accuracy. Therefore, they are widely used in industry. The performance of Yolo algorithm has not been brought into play, but yolov3 has made a breakthrough, but the author has not written a paper so far and believes that there is no substantive change.

Yolo V4 [12] proposed an efficient and powerful target detection model, which can train fast and accurate target detectors. A high-precision and real-time network is designed, and the network can be trained quickly with only one GPU. It is verified that bag of freebies and bag of specialties have a great impact on target detection, and it is used in Yolo V4 network. Two months later, ultralytics LLC launched yolov5 after yolov4 to transmit each batch of training data through the data loader and enhance the training data at the same time. The data loader performs three kinds of data enhancement: scaling, color space adjustment and mosaic enhancement. Mosaic zYolov5 uses the pytoch framework, which is very user-friendly and easier to put into production. It can effectively infer the input of single image, batch image, video and even webcam port directly. Transformer [13] is a common model in the field of natural language processing (NLP), which has been successfully applied in all research directions in the field of NLP. Recently, transformer has been applied in the field of computer vision, visoin transformer (VIT). The network structure of visoin transformer is mainly composed of two parts [14], namely encoder and decoder. First convert the

input image into path form, and then flatten each path into a fixed length vector, which is combined with position coding and input into the encoder. The features extracted by the encoder can be used to perform computer vision tasks. After vision transfosmer is applied to computer vision tasks, thanks to its advantages in large-scale database training, it performs well in various tasks, even surpassing convolution network. Unlike convolution networks, vision transformer trains network parameters using global image features.

3. Method

Pedestrian detection is the premise of analyzing pedestrian trajectory and estimating pedestrian density. In public scenes, due to the complex background and high pedestrian density, there are problems such as target occlusion and light change; In pedestrian detection, because pedestrians are non rigid objects, the change scale of walking posture is large, and the scale of pedestrians in the image or video is different. The target close to the camera has a larger size, while the target far away has a smaller size in the image or video; In addition, because the state of pedestrians may be static or moving, and the moving target has some problems such as imaging blur to a certain extent, which brings difficulties to pedestrian detection.

3.1 Improved YOLO v5 based on ViT

Although vision transformer has achieved better performance in computer vision tasks by using the global information of images, there are differences between images and natural languages. If it is directly applied to the field of computer vision, it will ineViTably learn a lot of redundant information, resulting in large model size and slow reasoning speed. A feature of ViT token in [15, 21] and ViT token in the literature is analyzed. From the visualization results of resnet50, it can be seen that the convolution network only learns some low-level features of the image, such as edge features[16, 22]; In ViT feature visualization, it can be seen that useful image features are not learned in the red box, which leads to redundancy in the network structure; In T2T ViT, it can be observed that among the shallow features (green box), T2T ViT can also learn the low-level features of the image. These low-level features are beneficial in computer vision tasks. With the deepening of the number of network layers, it pays more attention to the global features.

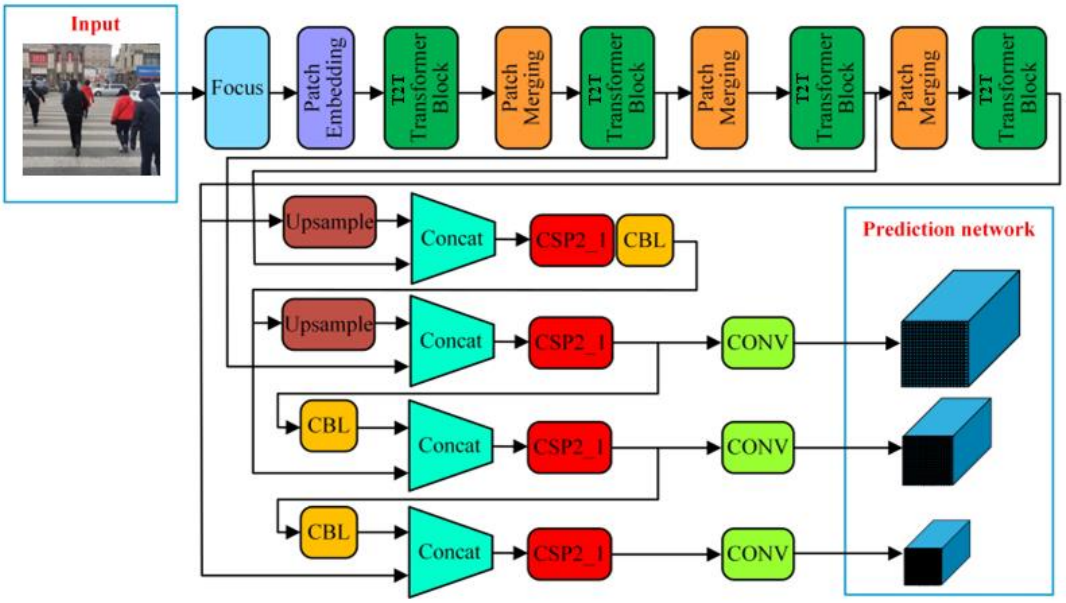

Figure 1 The network of improved YOLO v5 based on ViT

In T2T ViT, in order to improve the limitations of ViT's simple tokenization and inefficient backbone network, a tokens to token module [21] is proposed, which can gradually mark images into tokens and has an efficient backbone network. Therefore, T2T ViT consists of two main components The first component is the hierarchical T2T module, which is mainly used to model the local structure information of the image and continuously reduce the length of the token; The second component, T2T ViT backbone network, is used to learn the global attention relationship on tonken of T2T module. Based on the architecture design of CNN, the backbone network with deep and narrow structure is adopted to reduce redundancy and improve the richness of features. The network structure proposed in this paper is shown in Figure 1.

3.2 Improved T2T module

In the pedestrian detection task, due to the high similarity of the targets, that is, the targets all belong to the category of pedestrians, after the input image is patched and token, this a priori knowledge can be used to improve the network [17]. In vision transformer, after the image is converted into a patch, it is converted into a token, that is, each token represents a patch. Based on the above analysis, the T2T module based on token attention mechanism proposed in this paper aims to guide the network to further learn the characteristics of pedestrians. The improved T2T module network structure proposed in this paper is shown in Figure 2.

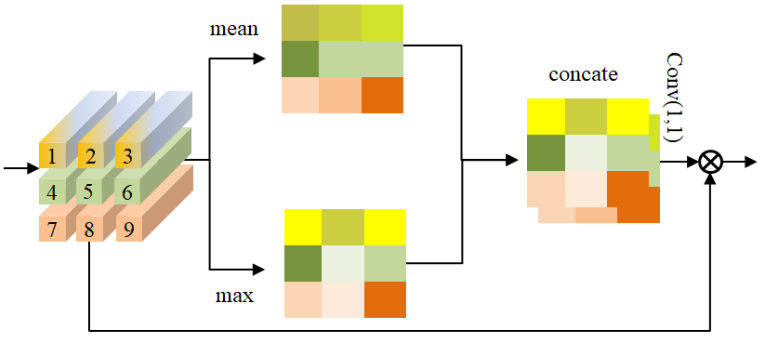

Figure 2 The network of improved T2T module

The token attention module proposed in this paper is shown in Figure 2. Take the mean and maximum values of the reconstructed tokens respectively. After connecting through concate, the number of channels will be 1 by convolution. After multiplying the obtained weight with the reconstructed token, we can get the token with attention, which makes the network pay more attention to the important tokens in the learning process, that is, the more important areas in the image. Set the input reorganized token as, and its calculation process is as follows:

$$I_{mean} = Mean(I_i) \tag{1}$$

$$I_{max} = Max(I_i) \tag{2}$$

$$\text{attention} = \text{Conv}_{1\times1}(\text{Cat}(I_{mean}, I_{max})) \tag{3}$$

$$I_i^a = attention * I_i \tag{4}$$

Where, the $\text{Conv}_{1\times1}$ represented convolution operation by is used for dimensionality reduction; Cat indicates the connect operation, which is about to be connected with, and indicates the reorganization token with attention.

4. Experiment

4.1 Experimental details

The experimental equipment and test environment of this paper are Intel Core i9-9400k CPU, rtx3090, 24GB GPU, 32GB memory, and the operating system is windons10. Configuration software environment: Python 3 6,Pytorch=1.7. Pedestrian detection belongs to single class target detection. In this chapter, the model is trained from the beginning of the data set, the initial learning rate is set to 0.001, the network input is 640 * 640, the batchsize is set to 12, and 200 epochs are proposed to be trained.

The data set used in this experiment is a public pedestrian detection data set, which is trained and tested on caletech [16] and CUHK occupation Pedstrian (COP) data sets [11] respectively. The images in caletech and CUHK occupation Pedstrian datasets are derived from real public scenes. Evaluation index [18, 19], select and compare the average accuracy (personap) and log average miss rate (lamr) of pedestrian detection to estimate.

$$Percision_C = \frac{N(TP)_C}{N(Total\ Object)_C} \tag{5}$$

$$Person_{AP} = \frac{\sum percision_C}{N(Total\ images)_C} \tag{6}$$

$$\mathrm{LAMR} = \exp\left(\frac{1}{9}\sum_{i=1}^{9}\ln(missrate(10^{-2.25+0.25t}))\right) \tag{7}$$

Where, C represents the category, and specifically in pedestrian detection, it represents the category of pedestrian; $N(TP)_C$ indicates the number of pedestrians correctly detected in the image and the number of pedestrians $N(Total\ Object)_C$ actually existing in the image. $\sum percision_C$ represents the sum of the successful detection rates of all images, and $N(Total\ images)_C$ represents the number of pedestrians contained in all test images.

4.2 Experimental analysis

In this paper, experiments and analysis are carried out on YOLO v5, T2T YOLO v5 (i.e. the framework of T2T ViT as the backbone network without using the improved T2T module proposed in this paper), and aT2T YOLO v5 (i.e. the target detection framework using the improved T2T module proposed in this paper). The detailed experimental indexes are shown in Table 1.

Table 1 Experimental results

Dataset	Method	personAP(%)	LAMR	times(s)	Avg times(ms)
Caletech	YOLO v5	88.36	0.19	48.61	57.75
	T2T YOLO v5	89.27	0.18	47.92	56.67
	AT2T YOLO v5	91.11	0.16	48.55	57.61
COP	YOLO v5	83.91	0.23	45.28	55.70
	T2T YOLO v5	85.52	0.21	44.83	54.85
	AT2T YOLO v5	86.77	0.20	45.69	54.91

As shown in Table 1, the experimental results of YOLO v5 [20] and two other target detection methods based on YOLO v5 are displayed in caletech and data set cop. It can be seen from the table that the pedestrian detection based on YOLO v5 has good effect and efficiency. The detection method T2T YOLO v5 based on YOLO v5 and T2T ViT makes use of the global information of the image. Thanks to the powerful performance of vision transformer in feature extraction, the performance of T2T YOLO v5 method is further improved compared with YOLO v5. At the same time, due to the efficient coding of patch in T2T ViT, the detection efficiency has been slightly improved. In aT2T YOLO v5, in order to pay more attention to useful patches, a patch based attention mechanism is introduced to assign greater weight to important patches, so

as to further improve the detection accuracy. At the same time, due to the expansion of network structure, the detection efficiency is slightly reduced.

4.3 Visualization of pedestrian detection results based on YOLO v5

(a)YOLO v5 (b)T2T YOLO v5 (c)AT2T YOLO v5

Figure3 Visualization of pedestrian detection results

Figure 3 shows the visualization of experimental effects of YOLO v5, T2T YOLO v5 and aT2T YOLO v5. The first column (a) in the figure is the experimental effect of YOLO v5, the middle column (b) in the figure is the experimental effect of T2T YOLO v5, and the last column (c) is the experimental effect of aT2T YOLO v5. It can be seen from the figure that the YOLO v5 method has missed picking up small-scale targets in the image, while it has been improved in T2T YOLO v5. The visualization with the third line of experimental effect shows that aT2T YOLO v5 can also detect the blocked target, which further improves the detection performance. However, due to the increase of network capacity, the detection efficiency also decreases slightly.

5. Conclusion

For pedestrian detection, this paper improves the backbone structure of YOLO v5 combined with vision transformer, so that the improved framework has better feature extraction ability. Vision transformer has excellent performance in image global feature extraction. At the same time, this paper further excavates the relationship between tokens and assigns weights to important tokens through the attention mechanism, so that the proposed framework has stronger feature extraction ability.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Regionbased convolutional networks for accurate object detection and segmentation. TPAMI, 2015. 5

[2] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[5] Qiao S, Chen L C, Yuille A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10213-10224.

[6] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[8] Wareechol T, Chiracharit W. Recognition of Similar Gait Pattern Using Transfer Learning DarkNet[C]//2021 9th International Electrical Engineering Congress (iEECON). IEEE, 2021: 381-384.

[9] Ballester P, Araujo R M. On the performance of GoogLeNet and AlexNet applied to sketches[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[10] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[11] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[12] BOLME D S,BEVERIDGE J R,DRAPER B A,et al.Visual object tracking using adaptive correlation filters[C]//Pro-ceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition .San Francisco,CA,USA,June 13-18 2010.New York:IEEE,2010:2544-2550.

[13] Krizhevshy A,Sutskever I,Hinton G E.ImageNet classification with deep convolution neural networks[J].Communications of the ACM,2017,60(6):84-90.

[14] Tang Y.Deep Learning using linear support vector machines[J].2015,arXiv:1306,0239v4.

[15] NAM H,HAN B.Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV,USA,June 27-30,2016,New York :IEEE,2016:4293-4302.

[16] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[17] Sultana F, Sufian A, Dutta P. A review of object detection models based on convolutional neural network[J]. Intelligent computing: image processing based applications, 2020: 1-

16.
[18] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[20] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
[21] Yuan L, Chen Y, Wang T, et al. Tokens-to-token ViT: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 558-567.
[22] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. 2011.