Research on a Pedestrian Detection Algorithm Based on Improved SSD Network

Cui Dong*, Xiaoshu Luo*

College of Electronic Engineering Guangxi Normal University Guiling, China

*dong877437617@163.com,
*Corresponding author e-mail: lxs@mailbox.gxnu.edu.cn

Abstract. When the SSD network uses an independent feature layer to detect objects, there is no connection between the layers, which leads to the problem of insufficient expression of contextual features. This paper proposes a pedestrian detection algorithm based on improved SSD network. The algorithm uses cross-layer feature adaptive fusion, and combines residual channel attention modules with different convolution rates of holes. While increasing the receptive field, this algorithm enhances important features and weakens unimportant features. Make the extracted features more directional, thereby improving the accuracy of pedestrian detection. Experiment on the improved network and algorithm on the INRIA pedestrian detection dataset, and the mixed pedestrian dataset extracted from the COCO dataset and the Crowd human dataset. The experimental results show that the average precision of pedestrian detection in the two datasets is improved by 1.7% and 4.0% respectively compared with the original network.

Keywords: SSD network; feature fusion; channel attention; residual connection; cavity convolution; receptive field.

1. Introduction (Heading 1)
The research of pedestrian detection system started in the mid-1990s, and is a branch of target detection. In recent years, with the rapid development of computer vision, target detection tasks have received more and more attention due to their wide application. Pedestrian detection with extremely high application value has become a hot research topic. Pedestrian target detection refers to the identification and detection of pedestrians in a specific scene to determine whether the object in the picture or video is a pedestrian; If it is a pedestrian, it is marked with a rectangular frame [1]. Pedestrian detection plays a vital role in human behavior analysis [2], image analysis, intelligent transportation, vehicle assisted driving [3], handheld PTZ control, intelligent monitoring systems and intelligent service robots [4] and other fields. Therefore, pedestrian detection has very important research value. The traditional pedestrian detection methods based on computer vision mainly include: using the gradient histogram of the global feature method as the description operator, for extracting features to perform pedestrian detection [5]; Using the method based on human body parts, each part is detected separately by splitting the body parts, and finally the detection results are integrated in a specific method and sent to the detector to determine whether it is a pedestrian. If it is recognized as a pedestrian, mark out pedestrians. Using a method based on stereo vision, first collect images with multiple cameras, then analyze the images, and
finally identify pedestrians. These three traditional pedestrian detection methods require researchers to design specific and adaptable feature extraction methods based on different detection scenarios. They rely on the researchers’ experience and have poor generalization capabilities.

2. SSD network and algorithm

2.1. SSD network composition structure
In the literature [15], the backbone network of the SSD network uses the convolutional layer of the VGG16 network as the basic network part, and at the same time it replaces the FC6 fully connected layer and FC7 fully connected layer of the original VGG16 with conv6 convolutional layer and conv7 convolutional layer. The feature extraction part uses conv4_3 convolutional layer for small target detection. The features extracted from the conv7 convolutional layer are used for large target detection. Then successively add new convolutional layers: conv8_2, conv9_2, conv10_2, conv11_2, as additional feature extraction levels to enrich feature extraction. A total of 6 feature maps of different sizes are formed, the size of the feature maps is [38,19,10,5,3,1], which forms a pyramid-like feature map structure, which greatly improves the detection accuracy. The SSD network structure is shown in Fig 1 below.

![SSD network framework](image)

**Fig.1 SSD network framework**

2.2. SSD network workflow and shortcomings
The SSD network workflow is: Input a 300 × 300 image, the basic SSD network contains a total of 13 convolutional layers. First, after a series of convolution operations, a large number of feature maps are obtained, and then 2 feature maps are extracted from the conv4_3 convolution layer and the conv7 convolution layer; Then extract the remaining 4 feature maps from the added convolutional layer, and finally form a total of 6 feature maps as the basis of the original bounding box and convolution prediction, and then send them to the detector for prediction. In the SSD network, feature extraction is abstracted layer by layer, shallow features contain more detailed information, which is conducive to positioning, and deep features contain rich abstract semantic information to facilitate classification. The expression of deep level feature information and the expression of shallow level information have a mutual guiding role. However, the SSD network sends the feature map to the detector independently, and predicts the detection result on this basis. This detection method makes there is no correlation between the feature layer levels and ignores the interaction relationship between the shallow features and the deep features, resulting in insufficient use of context information and insufficient feature extraction.
2.3. SSD network anchor box settings

The feature map structure in the SSD network is similar to the pyramid structure. As the convolution operation progresses, the number of channels of the feature map gradually increases, while the size of the feature map gradually decreases. The network sets anchor point boxes on the convolutional layer to correspond to different positions on the feature map. The size of the anchor point boxes corresponding to each feature map is different. That is, the anchor point boxes will change according to the scale of the feature map. Assuming that $m$ feature maps are used for prediction and judgment, the size of the anchor box of each feature map is:

$$ s_k = s_{\text{min}} + \frac{s_{\text{max}} - s_{\text{min}}}{m-1} (k-1), k \in [1, m] $$

(1)

In the formula, $m$ represents the number of feature maps. Set in reference [15], $s_{\text{min}} = 0.2$, $s_{\text{max}} = 0.9$, which means that the minimum scale is 0.2 and the maximum scale is 0.9. The multi-scale default boxes on each layer of feature maps are generated according to different aspect ratios. There are 5 aspect ratios in the SSD network, which are $a_r = \{1, 2, 3, 1/2, 1/3\}$. In practical applications, the value of the aspect ratio can be modified according to the characteristics of the dataset to better adapt to the dataset. The width and height of each default box are calculated as:

$$ w_k^a = s_k \sqrt{a_r} $$

(2)

$$ h_k^a = s_k / \sqrt{a_r} $$

(3)

For the case where the aspect ratio is 1 in the SSD network, an additional anchor box is added

$$ s_k = \sqrt{s_k s_{k+1}} $$

(4)

The number of default anchor boxes in the SSD network settings are $\{4, 6, 6, 6, 4, 4\}$. In practical applications, different anchor points are set according to different datasets to make them closer to the detection target. This article sets different anchor boxes for different datasets. On the INRIA pedestrian detection dataset [16], because pedestrians are all standing and the pixels are greater than 100. Therefore, according to the dataset and the characteristics of pedestrians (there is no 3:1 aspect ratio for pedestrians), the anchor box with an aspect ratio of 3:1 is removed in the experiment. The number of anchor boxes for each feature layer used in this dataset is changed from $\{4, 6, 6, 6, 4, 4\}$ to $\{4, 5, 5, 5, 4, 4\}$, the number of anchor boxes, The number has changed from 8732 to 8247. For mixed datasets with different poses, the experiment in this paper uses the anchor box set by the SSD network in [15].

3. Improved SSD network structure

In order to solve the problem of the lack of correlation between the feature layers of the SSD network, the interaction between the shallow features and the deep features is ignored, which leads to insufficient use of context information and insufficient feature extraction. This paper adopts a cross-adaptive feature fusion method to cross-utilize shallow and deep information to strengthen the fusion and expression of information; at the same time, the residual channel attention [11] module with different convolution rates is used to enhance the detection ability of the network and improve the detection ability of pedestrians.

3.1. Receptive field enhancement module

In the convolutional neural network, the definition of the receptive field is the size of the area mapped on the input image by the pixel points on the feature map output by each layer of the convolutional neural network, that is, a point on the feature map corresponds to the area on the input map. Effectively increasing the receptive field can improve the detection effect of the network on specific targets. Therefore, this paper uses residual blocks with different convolution rates of holes to combine with channel attention to effectively increase the receptive field. RCAN [17] proposed that the combination of residual block and channel attention can improve the super-resolution of the image. Tridentnet [18]
proposed that different hole convolution rates have different effects on detection objects of different sizes. The larger the hole convolution rate, the worse the detection effect on small objects. Therefore, in the pedestrian detection experiment, this paper uses the combination of the residual block with hole convolution with different hole rates and the channel attention. Channel attention re-weights the distribution of channel features, thereby enhancing the expression of useful information. The residual convolution blocks with different void rates are used to expand the receptive fields of different feature layers. This experiment uses a combination of the two to improve the ability of pedestrian detection.

Channel attention (CA) has the same principle as the SeNet [19] network. First, the input feature dimension is $H \times W \times C$ (height $\times$ width $\times$ number of channels), and the channel is compressed through global pooling and fully connected convolution. This paper takes the compression ratio $r=16$, which means that the channel is compressed to $1/16$ of the original, so that the two-dimensional channel becomes a real number with a certain degree of global receptive field. Then pass a fully connected layer to change the channel back to its original dimension to ensure the matching of the input and output channel dimensions; Then use a sigmoid activation function to get the weighting coefficient of the channel; finally, multiply the original input feature map to get a new feature map with weighting properties. This improves the expression of useful features and suppresses the expression of features that are not very useful for the current detection task. The CA module structure is shown in Fig 2.

![CA Module](image)

According to the characteristics of the output feature map of the SSD network, this paper sets the conv4_3 layer, conv7 layer, and conv8_2 layer to use the hole convolution rate to 1, 2, 3. Assuming the expansion rate is $ds$ and the size of the convolution kernel used is $3 \times 3$, the size of the receptive field using dilated convolution is $3+2\times2\times(ds-1)$, and the corresponding receptive field increases to $3 \times 3$, $7 \times 7$, $11 \times 11$. Because the current mainstream CNN-based networks treat each channel feature equally, they lack the ability to recognize and learn across feature channels, and ignore the rich low-frequency information expression of low-resolution input features. Therefore, this paper uses the residual attention module with hole convolution to enhance the ability of information expression. The cavity convolution is used to increase the receptive field, and the residual is used to enhance the feature expression of the upper layer. Channel attention is modeled based on the interdependence between feature channels [19], and adaptively learns more useful channel features, thereby strengthening the feature expression of the original feature map and improving learning and recognition capabilities. This enhances the pedestrian detection capabilities of the network. The residual attention module with hole convolution (DRCA for short) is shown in Fig 3.
3.2. Feature Fusion Module

This paper selects the basic conv4_3 layer, conv7 layer and the additional conv8_2 layer of the SSD network for feature cross fusion. Since the three feature maps extracted by other convolutional layers are relatively small and contain relatively rich semantic information, feature fusion processing is not performed. Mark the feature maps of the selected 3 feature layers as x1, x2, x3. The cross-feature fusion process is as follows: For the fusion of conv4_3 layer, first, use the upsampling method to adjust the size of the other two feature maps to make them the same size as the conv4_3 layer feature map; Then the new feature map X1 is obtained after the feature fusion operation. Then the new feature map undergoes a 1×1 convolution operation to obtain the weight parameters α1, β1, γ1. In the same way, use the up-sampling or down-sampling operation on the other two feature maps, re-adjust the feature maps to the same size and then perform the feature fusion operation to obtain new feature maps X2, X3. In the same way, the feature map undergoes a 1×1 convolution operation to obtain the corresponding weight parameters α2, β2, γ2, and α3, β3, γ3. The weight parameters obtained after concat operation, and then through the softmax activation function, the range is between [0,1] and the sum is 1. The obtained weight parameters are respectively multiplied with the corresponding feature layer to obtain new features block1, block2, and block3. The fusion process can be expressed as:

\[ y^l_{ij} = \alpha^l_{ij} x^l_{ij} + \beta^l_{ij} x^{2^l}_{ij} + \gamma^l_{ij} x^{3^l}_{ij} \]  

(5)

Where: \( l \in \{1,2,3\} \) is the mapping relationship from each level to another level, \( i \in \{1,2,3\} \); \( j \in \{1,2,3\} \) is the coordinate mapping of each level; \( y^l_{ij} \) Represents output characteristics; \( x^l_{ij} \) Represents input characteristics; \( \alpha^l_{ij}, \beta^l_{ij}, \gamma^l_{ij} \). Refers to the spatial importance weights of the three different levels of features obtained by network adaptive learning that are mapped to another level. The process of block1 obtained after feature fusion is: \( X^{1^l} \otimes \alpha^l + X^{2^l} \otimes \beta^l + X^{3^l} \otimes \gamma^l \). The overall structural framework is shown in Fig 4.

![Fig.4 Feature Fusion Module](image-url)
3.3. Improved SSD feature map structure

In this paper, the conv4_3, conv7, and conv8_2 layers of the SSD network in the literature [15] are cross-featured to meet the needs of the shallow feature layer for semantic information and the deep feature layer for detailed information. This enriches the extraction of features and enhances the detection effect of pedestrians. On the fused feature map, the residual structure with different cavity convolution rates and channel attention modules are used to further expand the receptive field; strengthen the expression of pedestrian detection information and further improve the accuracy of pedestrian detection. Fig 5 and Fig 6 show the extraction of network feature maps before and after the improvement.

![Feature map extraction structure diagram of SSD network](image1)

![Improved feature map extraction structure diagram of SSD network](image2)

4. Experimental results and analysis

4.1. Experimental environment

Operating system: Ubuntu16.04; GPU: GTX 1080Ti; memory size is 128 GiB; CPU: E5-2637 v4.

Acceleration library: CUDA9.0; programming language: Python3.5; deep learning framework: PyTorch0.4.

In the literature [15], the SSD network uses a batch size of 32 and an initial learning rate of 0.001. This paper uses a batch size of 16, with an initial learning rate of 0.0001 and momentum of 0.9. The loss function is the same as that in literature [15], using SmoothL1 and cross-entropy function. The sum as the total loss function.

4.2. Dataset and evaluation indicators

For the binary classification problem of pedestrian and background (non-pedestrian), the average detection accuracy is used as the evaluation index.

The accuracy rate and recall rate are respectively

\[
\eta_{\text{precision}} = \frac{\eta_p}{\eta_p + \eta_{fp}} = \frac{\eta_p}{n} \quad (6)
\]

\[
\phi_{\text{recall}} = \frac{\eta_p}{\eta_p + \eta_{fn}} \quad (7)
\]
In formulas 6 and 7 above: $\eta_p$ is the number of samples with correct predictions, that is, the number of samples with pedestrians both predicted and actual. Is the number of samples that are falsely detected, that is, the number of samples that are actually non-pedestrians and predicted to be pedestrians; $\eta_f$ is the number of samples that are falsely detected, that is, the number of samples that are actually non-pedestrians and predicted to be pedestrians; $\eta_{fn}$ is the number of samples that were missed, that is, the number of samples that are actually pedestrians and predicted to be non-pedestrians; $\eta_{precision}$ indicates the accuracy of detection in the test set; $\eta_{recall}$ indicates the missed detection rate in the test set. Change the recognition threshold so that the model can recognize the first K pictures in turn. The change of the threshold will cause the value of $\eta_{precision}$ and $\eta_{recall}$ to change at the same time, so that different precision-recall curves can be obtained. The area under the curve is the value of the average detection accuracy. The better the performance of a classifier, the higher the average detection accuracy. Target detection not only requires high accuracy, but also meets the speed requirements. Only a certain detection speed can meet the application requirements of the actual scene. For speed detection, the frame rate per second is generally used, and the number of pictures processed in one second. Another evaluation indicator of the detection model is the intersection and union ratio (IOU). This index is used to evaluate the similarity between two rectangular boxes, and can judge whether the detection of the detection box is correct or not. If the IOU threshold is too large, it will result in missed selection, and if it is too small, it will cause misjudgment. For a general image dataset, the average detection accuracy corresponding to the IOU threshold value of 0.5 is selected as the evaluation index.

This article experiment uses 2 datasets. The first dataset is the INRIA public pedestrian detection dataset, which is currently the most used static pedestrian detection dataset. The training set of this dataset has 614 positive samples, 1218 negative samples; and 288 positive samples in the test set. In this dataset, only upright pedestrians are marked in each image, and the clarity of the image is high, which meets the needs of the actual scene. The second dataset is to extract pedestrian photos from the COCO dataset and Crowd Human dataset. A total of more than 3,300 mixed pedestrian pictures of different poses and sizes (including upright and jumping poses) are selected, which are closer to daily life and are more practical. Among them, 3000 pieces are used as the training dataset, and 300 pieces are used as the test dataset, the above pedestrian pictures are not repeated.

4.3. Experimental results and analysis
It can be seen from Table 1 that after the feature fusion module is added, the average detection accuracy of the improved network in the INRIA dataset and the mixed dataset is 1.2% and 3.1% higher than that of the SSD network in the literature [15]. After adding the DRCA module at the same time, the average detection accuracy of the improved network's pedestrian dataset INRIA and mixed dataset has increased by 1.7% and 4.0% respectively compared with the original SSD network. It can be seen that the feature fusion module has a greater impact on the performance of the network, but the effect of the network model after the fusion of the two modules is better. Experimental results show that in different datasets, the improved SSD network significantly improves the average detection accuracy.

| dataset | Feature Fusion Module | DRCA module | Average detection accuracy /% |
|---------|----------------------|--------------|-------------------------------|
| INRIA   | ×                    | ×            | 88.5                          |
|         | ✓                    | ✓            | 89.7                          |
|         | ✓                    | ✓            | 90.2                          |
| Mixing  | ×                    | ×            | 61.4                          |
|         | ✓                    | ×            | 64.5                          |
|         | ✓                    | ✓            | 65.4                          |
Tab.2 Comparison of Average detection accuracy of different networks on the INR

| Network model | Average detection accuracy /% |
|---------------|-------------------------------|
| Fast-RCNN     | 85.0                          |
| Faster-RCNN   | 87.6                          |
| SSD           | 88.5                          |
| This article  | 90.2                          |

Tab.3 Comparison of Average detection accuracy in mixed datasets

| Network model | Average detection accuracy /% | FPS |
|---------------|-------------------------------|-----|
| SSD           | 61.4                          | 37  |
| This article  | 65.4                          | 30  |

The detection effect of pedestrian images before and after the improvement is shown in Fig 7 and Fig 8 below.

Fig.7 SSD network

Fig.8 Improved model of this article
On the INRIA dataset and mixed dataset, the improved network proposed in this paper has significantly improved the average detection accuracy of pedestrian detection. The speed reached 30FPS on the mixed dataset, meeting the demand of 25FPS for video streaming. It can be seen from Fig 7 and Fig 8 that the improved SSD network has a better detection effect on pedestrians in different poses and small target pedestrians. In summary, it can be seen that the improved SSD network has better detection performance.

5. Conclusion

In view of the lack of connection between the various layers of the SSD network when detecting objects, resulting in insufficient contextual feature extraction, this paper proposes a cross-layer feature adaptive fusion, combined with residual channel attention with different convolution rates of holes. The algorithm realizes pedestrian detection. While increasing the receptive field, the algorithm enhances important features and weakens unimportant features, making the extracted features more directional, thereby improving the average detection accuracy of pedestrian detection. In the experiment of the INRIA pedestrian dataset, the average detection accuracy of this article reached 90.2%, which is 5.2% higher than the average detection accuracy of the Fast-RCNN algorithm; For the mixed dataset extracted from the COCO dataset and the Crowd Human dataset, the average detection accuracy of this method is 4.0% higher than that of the original SSD network model. The experimental results show that the average detection accuracy of the model in this paper has been significantly improved for datasets with different characteristics, indicating that the generalization of the model is good, and it can meet the needs of pedestrian detection in different scenarios. Although the scale of the improved network proposed in this paper is large, the pedestrian detection speed has reached 30FPS, which can meet the needs of the actual video stream detection frame rate.

Acknowledgments

Fund Project: Guangxi Science and Technology Major Project (No. Guike AA18118004).

References

[1] Li Hai long. Research on Pedestrian Detection Problem Based on Regional Convolutional Neural Network [D]. Hangzhou Dianzi University, 2017.
[2] Hu Wei. Research on Human Behavior Recognition Method Based on Fourier Hidden Markov Model [J]. Electronic Design Engineering. 2018, 26(7): 185-188. DOI: 10.14022/j.cnki.dzsjgc.2018.07.041.
[3] Luo Xu, Hao Wangpei, Li Shaohua, etc. Dynamic target detection method for vehicle assisted driving system [J]. Computer Engineering, 2018, 44(1): 311-316. DOI: 10.3969/j.issn.1000-3428.2018.01.052.
[4] Zhou Wei. Research on pedestrian detection algorithm based on convolutional neural network [D]. Hangzhou Dianzi University, 2019.
[5] Lei Ke. Research on Recognition Technology of Unmanned Vehicle Traffic Signs and Markings [D]. Chang'an University of Technology. 2018.
[6] Yue Xin. Research on pedestrian detection based on HOG features and SVM classifier[J]. Science and Technology Innovation and Application, 2016(05):62-63.
[7] GIRSHICK R, DONAHHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 580-587. DOI: 10.1109/CVPR.2014.81.
[8] GIRSHICK R. Fast R-CNN [J]. Computer Science, 2015: 1440-1448. DOI: 10.1109/ICCIV.2015.169.
[9] HE K, HANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-16. DOI: 10.1007/978-3-319-10578-9_23.
[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region
proposal networks [J]. IEEE Trans Pattern Analysis & Machine Intelligence, 2017, 36(6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031.

[11] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C]// Computer Vision & Pattern Recognition. IEEE, 2016. DOI: 10.1109/CVPR.2016.91.

[12] REDMON J, FARHADI A. Yolo9000: better, faster, stronger [C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017: 6517-6525. DOI: 10.1109/CVPR.2017.690.

[13] REDMON J, FARHADI A. Yolov3: an incremental improvement [J]. arXiv e-prints, 2018.

[14] BOCHKOVSKIY A, WANG C Y, LIAL H Y M. Yolov4: optimal speed and accuracy of object detection [J]. 2020.

[15] LIU W, ANGUELOV D, ERHAN D et al. SSD: single shot multi-box detector [J]. Lecture Notes in Computer Science, 2016, 9905: 21-37. DOI: 10.1007/978-3-319-46448-0_2.

[16] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [J]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - San Diego, CA, USA (20-26 June 2005) DOI: 10.1109/CVPR.2005.177.

[17] ZHANG Y, LI K, LI K, et al. Image super-resolution using very deep residual channel attention networks [J]. 2018. DOI: 10.1007/978-3-030-01234-2_18.

[18] Li Y, Chen Y, Wang N, et al. Scale-aware trident networks for object detection [J]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019: 6053-6062, DOI: 10.1109/ICCV.2019.00615.

[19] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99). 10.1109/TPAMI.2019.2913372. DOI: 10.1109/TPAMI.2019.2913372.

[20] LIU S, HUANG D, WANR Y. Learning spatial fusion for single-shot object detection [J]. arXiv preprint arXiv: 1911.09516, 2019.