Research on Pedestrian Detection Algorithm Based on Deep Learning

Zhengqin Xu*, Wendong Zhao*, Laixian Peng* and Juan Chen*

Institute of Communications Engineering, Army Engineering University of PLA, Nanjing, China, 210007
Email: 155267201@qq.com
Email: nj_mouse@163.com
Email: lxpeng@hotmail.com
Email:pumpkin0910@163.com
*Corresponding Author

Abstract. Pedestrian detection is a kind of computer vision technology which is used to judge whether there is a pedestrian in a given image or video and estimate its position accurately. Pedestrian detection technology is widely used in intelligent security monitoring, driverless and robot fields. Pedestrian detection technology has nearly 30 years of development. Although it has greatly improved in accuracy and speed, there are problems such as lighting, pedestrian individual differences, occlusion and multi-scale pedestrian targets in the actual application scenarios, which lead to more missed and false detection in pedestrian detection. The algorithm is not robust enough, which makes the actual deployment difficult.

1. Overview of Improved Algorithm
Aiming at the task of pedestrian detection, this paper proposes the yolo-crd algorithm to meet the detection accuracy and speed requirements of pedestrian detection. The feature extraction network is improved to improve the ability of the main network for feature extraction[1], so that it can make full use of the shallow image features, and has better robustness for the multi pose and multi-scale problems of pedestrians, so as to better complete the following pedestrian detection tasks. In order to solve the problem that the proportion of positive and negative samples is seriously unbalanced in one stage target detection, we modify the classification loss function to reduce the weight of a large number of simple negative samples in the loss function. We can understand it as an online hard example mining (ohem) method [2]. In addition, in order to solve the multi-scale problem of the target, the loss function of the bounding box regression will be modified to make it more sensitive to the scale of the box, so that the network can complete the prediction of the bounding box more accurately. In this paper, when training the network, the image data will be enhanced, mainly by scaling, adjusting the contrast, rotating, flipping and other ways. Without increasing the image category and number, the amount of data will be increased and the generalization ability of the network will be improved. In the complex scene, there are many problems of pedestrian occlusion, which is a huge challenge for the detection algorithm, and it is easy to miss detection and false detection [3]. In order to solve this problem, this paper introduces attention into the yolo-crd algorithm, aiming to improve the feature extraction ability of the network for pedestrian targets, especially the extraction of local features of pedestrians, alleviate the occlusion problem, and improve the detection of the algorithm Measurement accuracy [4].
2. Improvement of Backbone Network

2.1. Inceptionmodular

In the task of pedestrian detection, real-time is one of the most important requirements. Although deep and complex backbone network can extract deeper and more features in theory, it is difficult to train, and it is prone to gradient disappearance and other problems, and it will slow down the forward speed of the network, and it is difficult to meet the real-time requirements in the real scene. Therefore, the improved backbone architecture of yolo-crd network still follows the idea of residual network to avoid the gradient disappearance caused by deep network and make the network easier to optimize. Another problem brought by deep network is a large number of parameters, so this paper uses the concept of smaller parameters and stronger feature extraction ability for reference In the module, perception can perform multiple convolution and pooling operations on the input in parallel[5]. Finally, the output results are spliced into a very deep feature map, and then the dimension is reduced by 1 * 1 convolution to promote the fusion of multi-channel information and reduce the amount of parameters.

Large scale convolutions are replaced in the yolo-crd network. Large size filtering convolution (such as 5 * 5, 7 * 7) has a large amount of calculation and parameters. A 5 * 5 convolution filter has 25 / 9 = 2.78 times more computation than a 3 * 3 convolution filter, but a 5 * 5 filter can learn more information and have a larger receptive field than a 3 * 3 convolution filter. The main idea of the improvement is shown in Figure 1. Under the premise of keeping the receptive field unchanged, reduce the parameter amount.

For example, a 5 * 5 convolution is decomposed into two 3 * 3 convolutions in series, and the receptive field of the two-level 3 * 3 convolutions is the same as that of 5 * 5.Suppose that the characteristic numbers of 5 * 5 and 2-stage 3 * 3 convolution output are the same, and the calculation amount of 2-stage 3 * 3 convolution is (3 * 3 + 3 * 3) / 5 * 5 = 18 / 25 of the former, about 0.72. The convolution of 3 * 3 is further decomposed. As shown in Figure 2, the convolution kernel of 3 * 3 is decomposed into 3 * 1 + 1 * 3 to replace the convolution of 3 * 3. The amount of calculation after improvement is reduced by 33%. By using this kind of asymmetric convolution instead of the original convolution in the network, the parameters in the network can be greatly reduced without sacrificing the performance of the network, and the calculation cost can be reduced. Especially in the pedestrian detection task, the improved convolution module can maintain the ability of extracting features of the network, effectively reduce the amount of network parameters, accelerate the forward speed of the network, and embedThe FPS (frames per second) on the computer without GPU can reach about 18.

2.2. The Improvement of Pyramid of Feature Map

The structure of the feature graph pyramid (FPN) is shown in figure 4-2-4, through which low-level to high-level feature information can be combined. From the bottom to the top is the forward process of the normal convolution neural network, forming a group of hierarchical structure, each layer of the size of the feature map is half of the size of the previous layer. In the pyramid structure of feature map, the output feature map size of each layer is the same. For the feature map pyramid, each level defines it as a pyramid level, and the final input of each level is recorded as the reference feature map of each level.
The deepest layer of each stage should have the most robust and representative features. Corresponding to the bottom-up structure, the feature pyramid also has top-down branches. By connecting the two branches horizontally, the features of each layer of feature map can be enhanced and the semantic information of the upper and lower layers can be fused. Figure 3 is the schematic diagram of the pyramid structure of the feature map. In the right branch, the feature map with lower resolution but higher semantic features is sampled 2 times from top to bottom, so that its size is the same as that of the corresponding left branch feature map in the lower layer. The corresponding left branch feature map is reduced by 1 * 1 convolution, and then the two feature maps are fused one by one, which is the fused feature map in this layer. This process is iterated until the most refined resolution feature map is generated. Before the prediction operation on each level, these characteristic images need to be convoluted through a convolution layer of 3 * 3 size to reduce the aliasing effect caused by the up sampling. The final feature set is called {P2, P3, P4}, corresponding to {C2, C3, C4} of the same space size.

![Figure 3. Pyramid structure of feature map](image)

In the task of pedestrian detection, the multi-scale problem of pedestrian is always one of the most difficult problems for detection accuracy. Although the pyramid structure of feature map can alleviate this problem to a certain extent, the robustness for small targets is still not strong. Therefore, in the yolo-crdd network, the feature enhancement module (FEM) is added to the pyramid output feature map of each layer. J + 1. Different dimension information of anchor frame a (I, J, L + 1) and a (I, J, L + 1) in the upper layer. FEM

The operation of the pyramid structure of the feature map in yolo-crdd is shown in Figure 4. Firstly, the 1 * 1 convolution kernel is used to denormalize the feature map; then, the upper layer feature map is sampled and added with the current layer element by element; finally, the feature map is divided into three parts on the channel and connected to the corresponding three subnetworks, each of which contains different number of 3 * 3 expansion convolutions with expansion rate of 3.

![Figure 4. Improved pyramid operation of feature map](image)
2.3. Attention Mechanism
In the scene of pedestrian detection task, pedestrian occlusion is a difficult problem that affects the detection accuracy of the algorithm. Pedestrians are occluded with each other or by objects, resulting in only the characteristics of pedestrians in the image, such as legs and hands. When the conventional algorithm detects the occluded pedestrians, there are a lot of missed and false detection. In order to alleviate the problem of occlusion and improve the accuracy of pedestrian target detection, attention module is added to the yolo-crd in this study.

Attention mechanism is inspired by human visual attention, which is a unique processing mechanism of human neurons. When people observe the surrounding things, they will quickly scan the current global image, then find the areas that need special attention, and then focus on the areas that need special attention, and no longer pay attention to other irrelevant areas around, which is the result of human evolution. The attention mechanism helps people to screen out the important information for themselves from a large number of external information.

The attention mechanism in deep learning refers to the special mechanism of human visual attention, which is used to screen and focus information. It can make more effective use of computing resources and improve the ability of feature extraction of network.

Another reason for using attention mechanism is that in pedestrian detection scene, there are many problems of pedestrian occlusion, which is easy to cause missed judgment and misjudgment. It is helpful to improve the detection accuracy of the network to detect pedestrian targets from the local characteristics of the human body that has never been occluded. Using attention mechanism can make the network easier to extract and learn the local characteristics of pedestrian targets, which is an effective solution to the problem of pedestrian occlusion One of the solutions to the problem of human occlusion.

In the deep learning method, the attention module usually builds a set of networks to generate the same size filter graph as the feature graph, filters the input according to the rules, or assigns different weights to each value on the input feature graph. The attention mechanism in this study mainly refers to the attention mechanism in neural network. Attention mechanism is used to select the channel and spatial area of convolution neural network.

3. Analysis of Experimental Results
As shown in Figure 5, it is the schematic diagram of the attention mechanism module (CNAM) used in this study, which is a simple and effective attention module for feedforward convolution neural network. An intermediate feature map is given. The module infers the attention map along two independent dimensions, channel and space, and then multiplies the attention map to the input feature map for adaptive feature refinement.

![Convolutional Block Attention Module](image)

**Figure 5.** Schematic diagram of attention mechanism module

In the channel branch, the average pooling and maximum pooling operations are used to aggregate the spatial information of the feature map, which is sent to a shared network, compress the spatial dimension of the input feature map, and sum and merge each element, In order to generate channel attention map MC, in the spatial branch, the average pooling and maximum pooling are used to compress the input characteristic map. The compression here becomes the compression at the channel
level, which is connected to generate the spatial attention map with 7 * 7 convolution. The CBAM integrated network can learn the information in the target area and aggregate the features from it. It is found that the effect of two attention modules in series is better than that in parallel. Channel attention module is better than spatial attention module in front.

By introducing the attention mechanism, the network can focus its attention on the pedestrian, which is the target of this detection task. It can improve the recognition ability of the network and the robustness of the network. In the complex street scene and other environments, the network can better find pedestrian targets and improve the detection ability of the algorithm.

Use the same training set to train the original yolov3, record the loss value of each training round, and after visualizing the relevant data, you can get two curves as shown in figure 6. The connection line of the positive line below is the loss curve of the improved yolov3, and the other is the loss curve of the original yolov3. Through comparison, you can see the loss of the improved network. The value is always lower than the original version, which shows that the improvement strategy of the Yo network in this study improves the accuracy of pedestrian detection, and it can converge faster than the original version of the yo ov3.

![Image of the loss curve in network training](image)

**Figure 6.** Comparison of loss curve in network training

After testing, the average accuracy (AP) of the improved yolov3 network in pedestrian detection task can reach 95.6, while the original version of yolov3 is 80.3, which is far better than the original version of yolov3. In the real-time aspect of pedestrian detection task, the original yolov3 can achieve 33 FPS per second on nvidia Titan x card, while the improved network can achieve 25 FPS per second. Although the speed is reduced, considering the improvement of accuracy, the loss of speed can be ignored.

4. Conclusion

Compared with the original yolov3 network, the improved network is more complex, but through the improvement of the backbone network, the overall parameters are reduced, the training speed of the network is accelerated, while the accuracy is improved, the real-time performance is taken into account. Compared with the original version, the real-time performance is slightly reduced, but the accuracy is greatly improved. In conclusion, the improved yolov3-crd algorithm in this study can be better applied to pedestrian detection tasks such as the actual scene, and achieved the expected results.

5. References

[1] Sun QJ, Wu EH, Huamn detection based on rectangle fitting[J]. Journal of Software, 2003, 14 (8): 13881393.

[2] Ma Chun. Pedestrian detection and tracking in dynamic context [D]. Nanjing University of Posts and telecommunications, 2016

[3] Li Jing. Research and application of intelligent video monitoring system [D]. Shanghai Jiaotong University, 2014
[4] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] Peng Zhiyong. Pedestrian recognition in intelligent video monitoring system [D]. 2016