International Core Journal of Engineering Volume 6 Issue 7, 2020
ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

Target Detection Algorithm Introducing Attention Mechanism: Attention_SSD

Hongtai Zhu, Chaoyan Gu

School of Shanghai Maritime University, Shanghai, China.

Abstract

The real-time target detection algorithm SSD has a faster detection speed and a higher accuracy rate, but there is a lack of accurate positioning of the bounding box, and it is not robust to the samples that are erroneously marked in the data set. An Attention_SSD target detection algorithm is proposed. This algorithm adds the channel attention mechanism and the spatial attention mechanism to the feature extraction network, replaces the original feature vector with the weighted feature vector, and uses tempered softmax and bi-tempered logistic loss Replacing the original softmax function and cross-entropy loss function makes the noise and error labels in the data set less impact on the model and accelerates model convergence. Experiments on the PASCAL VOC2007 and PASCAL VOC2012 data sets show that the algorithm reduces the positioning error of the bounding box and improves the detection accuracy and model convergence speed.

Keywords

Target detection; SSD target detection algorithm; Bi-tempered logistic loss; Attention mechanism.

1. Introduction

In recent years, deep learning algorithms [1] have been applied in many fields due to their unique advantages. In the field of computer vision , the combined products include R-CNN [3], fast R-CNN [4], faster R-CNN [5], YOLO [6] algorithm, and SSD [7] representative. Among them, the YOLO algorithm and the SSD algorithm belong to the one-stage detection method, which can achieve the effect of real-time detection, but it treats the target detection process as a regression problem and cannot distinguish the foreground and background well, and the false detection rate is high. The faster R-CNN target detection algorithm belongs to the two-stage detection method, in which the RPN layer detection algorithm can roughly determine the area containing the object to be detected in advance, which greatly improves the detection accuracy, but the speed is faster than the one-stage method Lowered.

In recent years of research, the combination of deep learning and computer vision attention mechanism has improved the target detection effect and classification effect. Jie Hu et al. Applied the channel attention module to the ResNet feature extraction network in the SENet proposed by ImageNet 2017, reducing the error rate by about 1%, and won the image classification champion that year. In addition, Woo et al found that modeling and weighting the channels and spaces in the convolution operation at the same time can better filter out the required features. In the traditional SSD target detection algorithm, the feature map extracted by the feature extraction network does not weight the different positions in the convolution kernel, that is, each part of the feature map is equally important, which is the same as in real life. The reaction is obviously different. When people look at the target items, they often automatically ignore their surrounding items, which is equivalent to automatically adding weight to the target. Similarly, we also need to add different weights to different positions of the feature map, so that the network can be better positioned on the feature to be detected.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

Based on the SSD algorithm, this paper proposes an attention-SSD target detection algorithm based on the attention mechanism. The attention mechanism is added to the feature extraction network VGG16 so that the attention effect can be reflected on the features. In addition, the softmax Replace the original softmax function and cross-entropy loss function with the bidirectional harmonic logic loss function, which reduces the impact of noise and error labels on the model in the data set and accelerates model convergence. Experiments show that attention-SSD has a certain performance improvement compared to the original SSD algorithm.

2. Related work

2.1 Attention mechanism

Early attention research was based on the analysis of the brain imaging mechanism, using a winner-take-all [2] mechanism to study how to model attention. With the rise of deep learning, the attention mechanism is mainly through Use a mask to form. The principle of the mask is to identify the key points in the picture data by forming a new layer of weights, and through learning and training, let the convolutional layer network learn the areas that need attention in each new picture.

According to the classification of attention mechanism, it can be divided into soft attention mechanism and strong attention mechanism. The key point of the soft attention mechanism is that this attention pays more attention to the area or channel. After the neural network training is completed, it can be directly passed The network is generated, and soft attention is differentiable, so that the gradient can be calculated by the neural network and the weights of attention can be learned by forward propagation and backward feedback. The difference between strong attention and soft attention is that first, strong attention is a more focused point, that is, each point in the image may extend attention, and at the same time, strong attention is a random prediction process, which emphasizes more Dynamic changes. Strong attention is a non-differentiable attention, and the training process is often done through reinforcement learning.

When using the attention mechanism to process pictures in computer vision, it is assumed that the feature sequence to be processed by the attention mechanism is:

$a=\{a_1,a_2,\cdots a_n\}$, $a_i \in R^L$ Where n represents the number of feature vectors and L represents the spatial dimension.The weight a_i of the current time t of each feature vector $\alpha_{t,i}$ is calculated as follows:

$$\mathrm{e}_{\mathrm{ti}} = f_{att}(a_i, h_{t-1}) \tag{1}$$

$$\alpha_{\mathrm{t,i}} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})} \tag{2}$$

Among them, $f_{att}(.)$ represents the multilayer perceptron, e_{ti} represents the intermediate variable, h_{t-1} represents the hidden state at the last moment. Through weights, the model filters the input sequence a to obtain the filtered feature sequence o_i, where $o_t=\sigma(\{\alpha_i\},\{a_i\})$,

Among them, σ represents a function, the strength or softness of attention depends on it, let s_t denote the position of attention points, $s_{t,i}$ represents the one-hot encoding vector,think of $a_{t,i}$ as a probability, and the polynomial distribution composed of it gives the final probability selection o_t, the formula is as follows:

$$p(s_{t,i}=1 \mid h_{j<i},\alpha)=a_{t,i} \tag{3}$$

$$\mathrm{o}_t=\sum_i s_{t,i}\alpha_i \tag{4}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

In order to visually illustrate the effect of adding attention to the image classification results, we used the grad-cam method to visualize the classification results. The grad-cam is improved on the basis of the cam, and the cam is a fully connected layer by replacing the network architecture For the global tie pooling layer, the weights are obtained from the new training.The grad-cam uses the global average pooling of the gradient to calculate the weight, so that the weight of the feature map corresponding to each category can be obtained without modifying the original model structure or retraining the model, and then the weight and the corresponding The feature maps are weighted and summed. Therefore, a visual heat map can be obtained for each category. As shown in Figure 1:

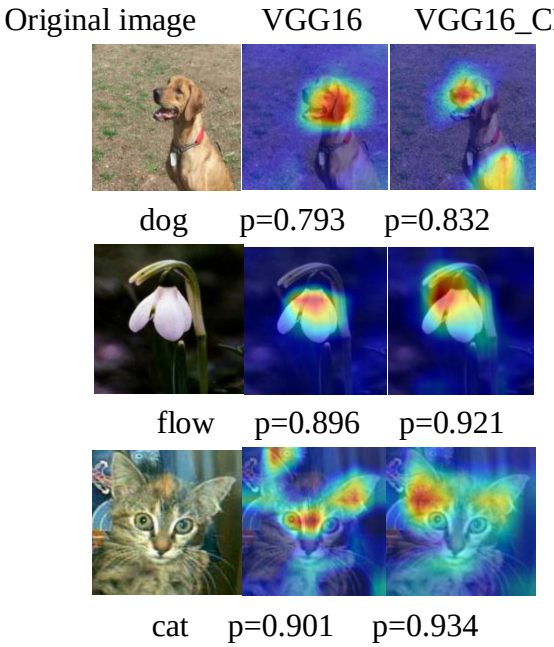

Figure 1 Contrast heat map with attention mechanism

In the above figure, the leftmost column corresponds to the category of objects in the picture are dogs, flowers, and cats. The second column corresponds to the feature map extracted by the VGG16 network model. The grad-cam method is used to visualize the classification results. The red part is where the confidence obtained in the classification process is higher.

The third column corresponds to the visualization of the classification results through the grad-cam method after adding the channel attention mechanism and spatial attention mechanism to the VGG16 network model. The red part corresponds to the part with high confidence in the corresponding category of the feature map. Through comparison, it can be found that after the attention mechanism is introduced in VGG16, the red area in the figure becomes significantly larger, the category softmax score increases, and the classification result is better. Therefore, in the SSD target detection algorithm, the fusion of channel attention is used. The feature extraction network of force mechanism and spatial attention mechanism will make the regression prediction and classification training get better results later.

2.2 SSD target detection model

The SSD target detection algorithm is a representative algorithm in one-stage. It uses a direct regression method to obtain the position and category of the target. Compared with the two-stage algorithm, it does not require the process of candidate box extraction, so the detection speed is faster. It performs training prediction on feature maps of different scales, so the detection accuracy can also be guaranteed for pictures with lower resolution. It uses the anchor mechanism in the R-CNN series of algorithms to classify and distinguish each pixel on the feature map. Through such a process, better accuracy can be achieved

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

2.2.1 Backbone network

The network architecture of the SSD target detection algorithm is mainly composed of the basic feature extraction network VGG16 and the detection and prediction layer. The feature extraction network is not limited to VGG16, and other feature extraction networks such as ResNet can also be used, but it will increase the calculation amount and time cost. This article still uses the classic VGG16 network. Among them, the final fully connected layer of the basic feature extraction network VGG16 is changed to a convolutional layer, and four additional convolutional layers are added. The full connection is removed because the traditional VGG16 is used for classification tasks. For the classification task, the final output will use a fully connected layer to map the feature map to a vector, which corresponds to the probability distribution of different objects. While for target detection, we only need VGG16 for feature extraction, so we need to turn the fully connected layer into a convolutional layer. The additional four convolutional layers are on the one hand to extract deeper semantic features, and on the other hand, feature maps of different scales can be added as input to the final detection layer to improve model performance.

2.2.2 Multi-scale feature prediction

For SSD, he will use six different scale feature maps as the input to the prediction network, including 38 * 38, 19 * 19, 10 * 10, 5 * 5, 3 * 3, 1 * 1, different scales The feature map is usually down-sampled using the Pooling method, and the feature map for each layer is then input into the prediction network. The prediction network will include a priority box layer, which is similar to the anchor mechanism in faster r-cnn. It will use each pixel in the feature map as a cell, and use this cell as the center to perform proportional scaling to find him in the original. The position in the image, and use this point as the center to extract the bounding boxes of different scales, define these bounding boxes of different scales as the priority box, and compare this priority box with the true value to get its label. For each priority box We will predict its category value and coordinate offset separately.

2.2.3 Prior box screening

During training, it is not necessary to calculate the loss of the 8732 bounding boxes generated, but to first filter them for positive and negative samples. The positive sample construction method is as follows:

Starting from the Prior box set, find the largest Prior box that satisfies the IOU greater than 0.5 with the GT box and put it into the candidate positive sample set, and not less than 0.5 is put into the negative sample, because this will cause too many negative samples and make the model difficult convergence. Generally, the ratio of positive and negative samples is set to 1: 3, and the intermediate samples can be selected and discarded. During training, make sure that the priority box classification is accurate and return to the GT box as much as possible.

2.3 Loss function

For each priority box, the category probability and coordinate offset will be predicted, so its loss function contains two parts, namely classification loss and regression loss.As in formula 5. For classification loss, because each priority box predicts c category probabilities, softmax loss is used for regression loss, smooth l1 loss is used.

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g) \quad (5)$$

Among them, α is used to balance the two models to optimize the scale factor.

$L_{conf}(x,y)$ is the confidence loss, $L_{loc}(x,l,g)$ is the positioning loss, c is the confidence, l is the prediction box, and g is the true box.

3. Network based on channel and spatial attention mechanism

This article introduces the attention mechanism from the two dimensions of channel and spatial based on the original SSD target detection algorithm to improve the detection result. In order to improve

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

the detection accuracy on the basis of not increasing the amount of calculation and running time as much as possible, this paper only improves the convolutional layer of the feature extraction network VGG16, and filters the incoming feature maps by channel direction and spatial direction respectively , Making the filtered features more conducive to classification and positioning.

3.1 Join the channel attention mechanism

The first is channel attention. We know that a picture will get a feature matrix after several convolutional layers[8]. The number of channels in this matrix is the number of convolutional layer kernels. Well, there are dozens to hundreds of common convolution kernels, not every channel is very useful for information transmission. Therefore, by filtering these channels, that is, paying attention to increase the weight of effective channels and reduce the weight of invalid channels, the optimized features are obtained. The formula is as follows:

$$M_c(F) = \delta(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (6)$$

Among them, $F \in R^{C*H*W}$, F is the feature map after convolution, AvgPool () stands for global average pooling,Maxpool stands for global maximum pooling, and δ stands for activation function. The structure diagram is shown in Figure 2.

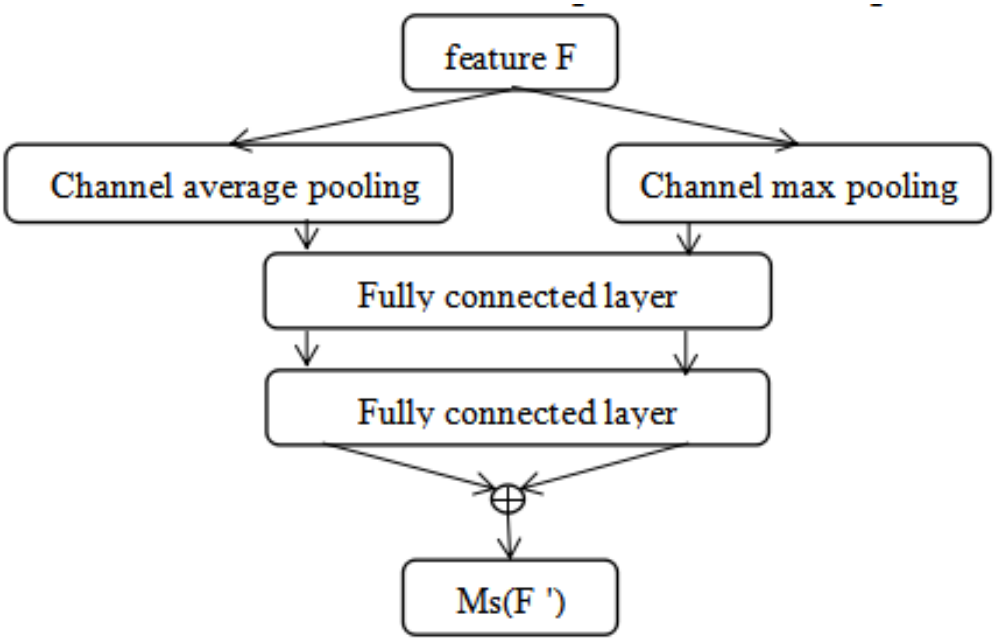

Figure 2 Channel attention mechanism

In Fig. 2, it is assumed that the result of the convolution of a certain layer is F, and Favg and Fmax are obtained by performing global average pooling and global maximum pooling, respectively. The one-dimensional weight sequence Favg can play a role in screening the global background information of the object, and Fmax can well highlight the salient features of the target object. Then, input Favg and Fmax to two fully connected layers. The two fully connected layers share parameters during training. The outputs of the fully connected layer are outputavg and outputmax, and the two values are added and then obtained after the activation function sigmoid Channel attention Mc[9].

3.2 Add spatial attention mechanism

If channel attention focuses on: what, then spatial attention focuses on: Where, the formula is as follows:

$$M_S(F') = \delta(f^{7*7}([AvgPool(F'); MaxPool(F')])) \quad (7)$$

Where δ represents the activation function, AvgPool represents the average pooling of the channel, and Maxpool represents the maximum pooling of the channel. f^{7*7} stands for 7 * 7 convolution operation.Its structure is shown in Figure 3:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

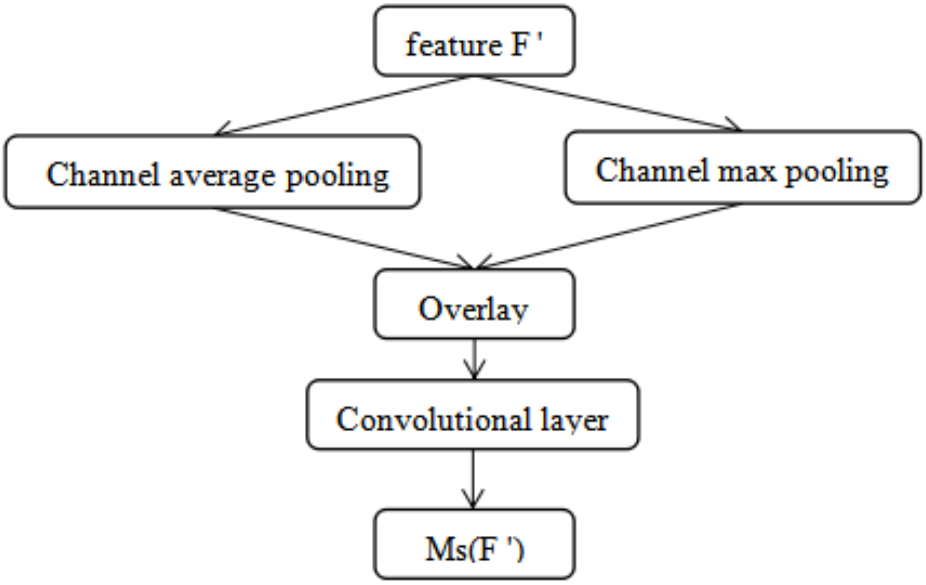

Figure 3 Spatial attention mechanism

As shown in the figure above, the feature F 'input to the spatial attention mechanism is channel average pooling and channel maximum pooling, respectively, and then the output results are superimposed along the channel dimension. In order to obtain two-dimensional weight information, you need to next Dimensionality reduction of the convolution,final output spatial attention Ms(F ').

3.3 Add reconciled softmax and bi-tempered logistic loss

It is well known that the logarithmic loss is also called softmax loss[10]. The softmax loss function has become a standard choice for training classification neural networks. The network loss is to input the final output activation value of the network into the softmax function to form a class probability, and then use real labels and prediction The probability is obtained by relative entropy[11]. The logical loss function for the activation value of the final output of the classification neural network is a convex function. The convex function is very convenient for solving the global optimal solution. When the loss of each sample is unlimitedly increased as the activation function, the convex loss function is easy to throw out abnormal values. Such outliers may be the extreme gradients caused by individual extreme samples, or misclassified training samples located far from the classification boundary. In this case, it is somewhat arbitrary to place a convex loss function on the output layer. Another problem is caused by the exponentially decay tail of the softmax function. On the performance of the error label training set near the classification boundary, the short tail of the softmax probability forces the classifier to closely follow the noise training sample. On the contrary, the softmax probability heavy tail alternative can significantly improve the robustness.The logarithmic loss function is essentially taking the logarithm of the predicted category probability, which is calculated by the input normalized index. In order to deal with its convexity and light tail, the harmonic version is used here:

$$\log_t(x) = \frac{1}{1-t}(x^{1-t} - 1) \tag{7}$$

$$\log_t : R_+ \to R \tag{8}$$

This $\log_t$ is monotonically increasing and concave, and the standard logarithmic function is the limit for t approaching 1. Unlike standard logarithmic functions, for $0 \le t < 1$,

The lower bound of this $\log_t$ function is -1 / (1-t). This attribute is more robust to defining outliers for bounded loss functions,Also use the harmonic index function to realize the heavy-tail replacement of the softmax function, $\exp_t : R \to R_+, t \in R$

$$\exp_t(x) = [1 + (1-t)x]_+^{1/(1-t)} \tag{9}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

And $[\bullet]_+ = \max\{\bullet, 0\}$ This standard exponential function is taken at the limit when t approaches 1. Compared with the standard exponential function, for t> 1 to achieve a heavier tail, this attribute is used to define the heavy-tail analogue of the softmax probability of the output layer.

In Figure 4, we draw two loss functions when t takes different values, so that we can understand them more clearly

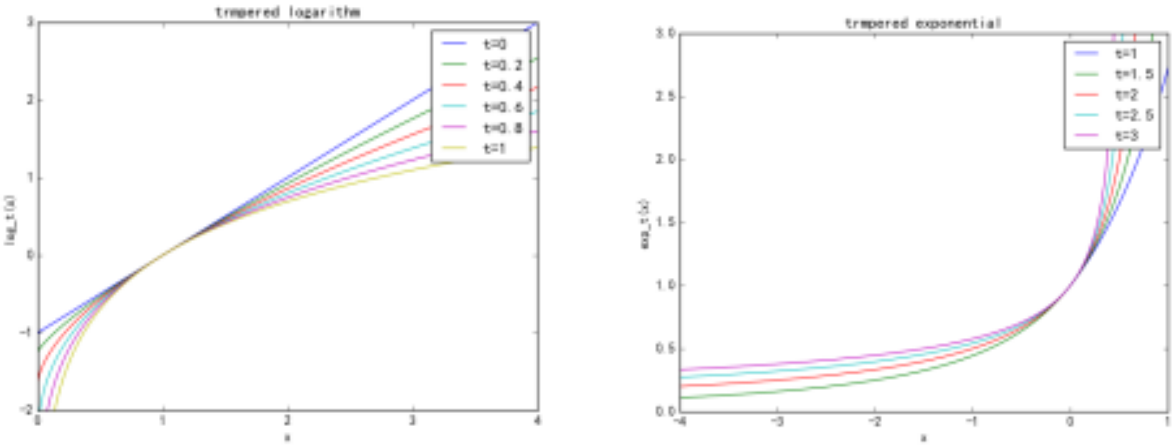

Figure 4 Logarithmic harmonic function and exponential harmonic function

4. Training and testing

4.1 Data preparation

The data set used in this article is PASCALVOC07+12,Its picture collection contains more than 21,000 pictures, of which 16,000 are training pictures and more than 5,000 are test pictures. There are 20 categories: humans; animals (birds, cats, cattle, dogs, horses, Sheep); transportation (airplane, bicycle, boat, bus, car, motorcycle, train); indoor (bottle, chair, dining table, potted plant, sofa, TV). It is a benchmark data set for the classification, recognition and detection of visual objects provided by the PASCAL VOC Challenge. Its data set has good image quality and complete annotation, which is very suitable for training and testing algorithm performance.

4.2 Data enhancement

In the preprocessing stage, the image of the input model was rotated, left-right inverted, and HSV saturation and intensity transformation, the parameter settings are as follows:

Table 1 Image preprocessing parameters

Data enhancement type	parameter
Spin	+/-5degrees
Flip left and right	40%probability
HSV saturation	+/-30%
HSV strength	+/-30%

4.3 Network training

Due to the addition of channel and spatial attention mechanism and harmonic loss function, the algorithm is more accurate in determining whether the grid contains objects to be detected. It can be seen from Figure (6) and Figure (7) that the feature extraction grid can better position in the target pixel grid where there may be objects during training, because the positioning is more accurate, and the frame regression loss is also reduced accordingly. It can be seen from the experiment that the added attention mechanism has a certain training acceleration effect on each loss of the original

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

algorithm, and the introduction of the harmonic loss function allows the algorithm model to converge better, so it has better detection performance.

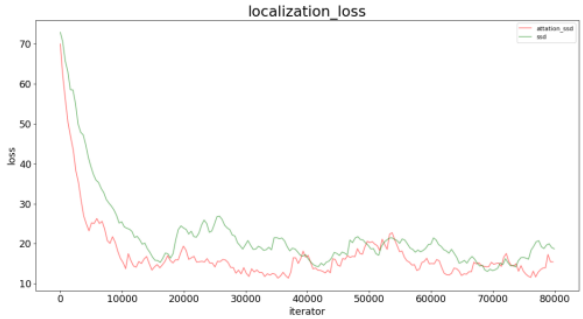

Figure 6 Comparison of border regression loss training process

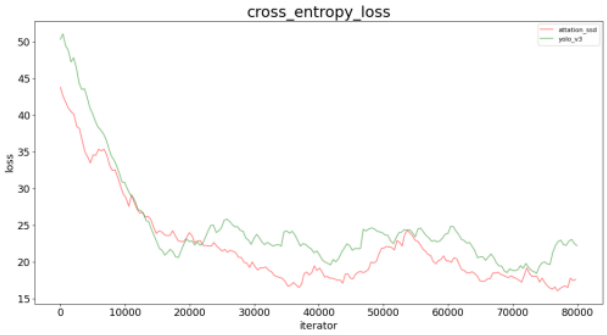

Figure 7 Comparison of classification loss training process

5. Analysis of experimental results

During the training process, observe the entire training process by observing tensorflow's tensorboard. When the loss of the model during training does not change significantly, stop training. In this paper, the data of various training processes in tensorboard is derived, and after smoothing, the smoothing factor is 0.80. Draw as shown in figures 6, 7, 8 etc. for easy observation and research.

5.1 Vertical comparison

In order to compare with the SSD algorithm that combines the attention mechanism and the harmonic loss function, we used the PASCAL VOC 2007 + 2012 data set to train the original SSD target detection algorithm and the SSD that introduced the attention mechanism. IOU selected [0.5: 0.9] to To comprehensively evaluate the detection performance of the model, The test results are shown in Table 3

Table 3 mAP test results of the two models on the voc2007 test set

IOU	0.5	0.6	0.7	0.8	0.9
SSD	74.3	72.6	64.2	53.2	38.1
Attention SSD	81.7	80.5	78.6	65.4	59.2

5.2 Horizontal comparison

In order to take the improved SSD target detection and other similar detection algorithms, this paper still uses the PASCAL_VOC07 + 12 data set for training and testing, and the evaluation standard uses mAP and rate. In the training process, two classic target detection algorithms are selected to draw the loss function convergence results (8) for observation and comparison. The results show that, depending on the excellent design of the SSD algorithm, the attention mechanism and the harmonic loss function, the Attention_SSD algorithm performs optimally on the total loss convergence speed

ISSN: 2414-1895 DOI: 10.6919/ICJE.202007_6(7).0038

and results.Achieved 81.7mAP performance on the test set.Compared with the regional suggestion detection algorithm, it has better application prospects.

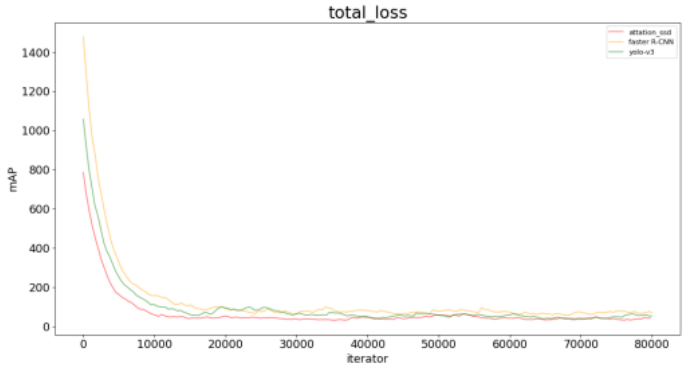

Figure 8 Comparison of training loss of three target detection algorithms

6. Conclusions and prospects

This paper proposes an SSD target detection algorithm that introduces channel attention and spatial attention mechanisms. The main idea is to model and weight the channel and space directions of the feature map, so that the filtered features are beneficial for subsequent classification and detection At the same time, in order to reduce the influence of wrong labels, outliers, and noise on the classification boundary that often occur in the training data set, a bidirectional harmonic log loss function is introduced. The robustness of the model is improved, and the accuracy of detection is also improved. The next step will be to improve the problem of missed detection of small and distant objects and misdetection of objects covered by obstacles during the detection process.

References

[1] Feng Xiaoxia. Research on image recognition algorithm based on deep learning [D]. Taiyuan University of Technology, 2015.

[2] Liu Bin. Research on face detection based on adversarial deep learning [D]. Qingdao University of Science and Technology, 2018.

[3] Lu Zhen. Research on video-based obstacle detection methods in vehicle-assisted safe driving [D]. University of Electronic Science and Technology of China, 2016.

[4] Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, et al. (2014) OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks [C] 2014.

[5] Girshick, Ross, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[6] Girshick R. Fast r-cnn [C] // Proceedings of the IEEE International Conference on Computer Vision. 2015

[7] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [C] // Advances in neural information processing systems. 2015.

[8] 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.

[9] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C] // European Conference on Computer Vision. Springer International Publishing, 2016.

[10] Shen Z, Liu J, et al. DSOD: learning deeply supervised object detectors from scratch [C] // IEEE International Conference on Computer Vision, 2017.

[11] Huang, G., Liu, Z., Maaten, L.V.D., et al. (2017) Densely Connected Convolutional Networks [C] // IEEE Conference on Computer Vision and Pattern Recognition, 2017.