Improved Smoking Target Detection Algorithm Based On YOLOv3

ZhaoWei1, YanXuZhu2, QiXuanLi3 and CaiWang4
1Department of Information Science, Beijing University of Technology, Beijing, 100020, China
2Department of Information Science, Beijing University of Technology, Beijing, 100020, China
Corresponding author’s e-mail: weizhao@email. bjut. edu. cn

Abstract. For the detection of smoking behavior, in order to make accurate state judgments, a smoking target detection algorithm based on the improvement of YOLOv3 is proposed. By introducing Mosaic data to enhance the background of the rich image, the batch_size is increased, and Mish is used as the activation function. To improve the generalization ability of the model, the DIOU_Loss function is used as the activation function to effectively use the information of the center point of the bounding box to improve the accuracy of target detection. Through self-made smoke detection data sets, experiments are carried out, and the experimental results show that the algorithm can accurately detect To smoking behavior, improve the MAP and AP of training.

1. Introduction
According to the statistics of the World Health Organization, among the 7.5 billion people in the world, the number of smokers has reached 1 billion, and the current number of smokers in China is 350 million, and it is rising at an alarming rate every year. Smoking seriously harms humanity, physical health and public property safety. In 2014, the National Health and Planning Commission drafted the "Regulations on Smoking Control in Public Places (Draft for Review)", which stated that smoking in all public places is strictly prohibited. Use manual supervision or smoke detection to supervise and detect smoking. By adopting these methods, not only is human and financial resources, but it is inevitable that human negligence or poor sensitivity of smoke detection equipment can cause fires to damage public property. Through the use of computer vision technology, combined with the most used currently Deep learning, the use of optical cameras to collect video image information, and extract the feature information of video images, is efficient and real-time. At this stage, more algorithms for detecting smoke or gestures that appear in smoking have been proposed. In the literature [2] Proposed a segmentation method based on the combination of motion information in the video sequence and two skin color detections. The real-time performance is poor, and the segmentation effect of gesture images in complex environments is not good; the document [3] proposed to combine the Gaussian mixture model with the frame difference The method is combined with the algorithm to separate the dynamic smoke area according to the characteristics. Although it has a good real-time effect, it is easily interfered by the completion factor; in the literature [4], the machine learning algorithm is used by analyzing the characteristics of the contact time of the face and hand samples. The decision tree is classified, and the applicability is better. Although the effect is better in a
complex environment, it is less effective when it is applied to a single environment.

In view of the above detection methods for smoking, whether it is for smoke or hand feature detection, there are obvious shortcomings. This paper uses a self-made smoking data set as the research object, and proposes an improved YOLOv3 detection algorithm. First, the data set is data. Enhancement, the introduction of Mosaic data to enhance the background of rich images, increases the size of the data set, avoids over-fitting and other phenomena, and reduces the requirements for the GPU

2. YOLOv3 algorithm model

2.1. Introduction to YOLOv3 structure

With the development of computer vision technology, target detection tasks have also been widely used in various industries, and many algorithm models have been derived, such as Faster RCNN[5], SSD[6], and YOLO[7]. With the development of technology With continuous improvement, YOLOv1, YOLOv2, and YOLOv3 have appeared. YOLOv3 algorithm is currently the most used algorithm model. The backbone of the algorithm model uses the Darknet53 feature extraction network. The Darknet network contains 53 convolutional layers, from layer 0 to layer 0. 74 layers, 53 layers are convolutional layers, and the rest are res layers, and integrate the ideas of Resnet. This structure can avoid problems such as gradient disappearance and gradient explosion. Darknet is a fully convolutional layer network without a pooling layer And the fully connected layer, as shown in Figure 1 below, the 1, 2, 8, 8, 4 in the figure represent several repeated residual networks, and the downsampling of the network is achieved by setting the step size of the convolution to 2. Every time after passing through this convolutional layer, the size of the image will be reduced to half. The implementation of each convolutional layer includes convolution + BN + Leaky relu, and each residual network module is added after one zero padding.

![Figure 1. Darknet-53 network structure diagram](image)

The input of YOLOv3 needs to be reshaped, and the size is unified to the same size of 416*416*3. Three different scales of prediction results can be obtained through the Darknet network. Each scale corresponds to N channels and contains the predicted information. Each The prediction corresponds to 85 dimensions, which are 4 (coordinate value), 1 (confidence score), and 80 (coco category number), as shown in Figure 2. Finally, YOLOv3 uses K-means clustering to obtain the size of the prior frame. According to the standard frame of the data set, 9 sets of a priori boxes are calculated, and they are evenly distributed to 3 different scales. The points in each scale correspond to 3 sets of a priori boxes respectively, thus, the positions of the 3 feature a priori boxes are obtained.
2.2. Target bounding box prediction
The YOLOv3 network uses \((4+1+c) \times k\) convolution kernels of size 11 to perform convolution prediction in the three feature maps, where \(k\) is the number of preset bounding box priors, and \(c\) is the prediction target. The number of categories, \(4k\) parameters are responsible for predicting the offset of the target bounding box, \(k\) parameters are responsible for predicting the probability of the target contained in the target bounding box, and \(ck\) parameters are responsible for predicting the \(k\) preset bounding boxes corresponding to \(c\) target categories. Probability. Figure 3 shows the prediction process of the target bounding box. The dotted rectangular box in the figure is the preset bounding box. The solid rectangular box is the predicted bounding box calculated by the offset predicted by the network. Among them are the center coordinates of the preset bounding box on the feature map, \(w\) and \(h\) are the width and height of the preset bounding box on the feature map. The center offset of the bounding box predicted by the network and the width-to-height scaling ratio are the target of the final prediction.

![Figure 3. Bounding box prediction diagram of YOLOV3](image)

2.3. Improved YOLOv3 algorithm
Mosaic data enhancement simply means to sort four pictures through random scaling, random attenuation, and random arrangement. The Mosaic data enhancement method greatly enriches the detected data set, and at the same time reduces the use of GPU, directly calculates the data of 4 pictures, so that the Mini-batch size does not need to be large, and a GPU can achieve better results.

Compared with the LeakyRelu function, the Mish function has no upper bound and lower bound: no upper bound is a feature required by any activation function, because it avoids gradient saturation that leads to a sharp drop in training speed; no lower bound Attributes help to achieve a strong regularization effect. Non-monotonic function: This property helps to maintain a small negative value, thereby stabilizing the network gradient flow. Leaky ReLU, cannot remain negative due to its difference of 0 Value, so most neurons are not updated. Infinite-order continuity and smoothness: Mish is a smooth function with good generalization ability and effective optimization ability of results. As show in Figure 4:
Figure 4. Diagram of Leaky ReLU function and Mish function

The DIOU_Loss function considers the information of the center point distance of the bounding box based on IOU and GIOU. Its loss formula is:

$$L_{DIOU} = 1 - IOU + \frac{s^2(b, b')}{c^2}$$

DIOU loss retains some of the advantages of IoU loss and GIOU loss. DIOU loss can directly minimize the distance between the two target frames. So it converges much faster than GIOU loss; for the situation that contains two boxes in the horizontal and vertical directions, DIOU loss can make the regression very fast; DIOU can also replace ordinary IoU. The evaluation strategy applied to the NMS, makes the results obtained by the NMS more reasonable and effective.

3. experiment

3.1. Data sets
This article uses a self-built data set method to construct the data set required for model training. The total number of data sets used is 9424 smoking images, and each image contains the action of smoking. As shown in Table 1:

Table 1. Experimental data

| Data sets    | Number of pictures | Number of smokers |
|--------------|--------------------|-------------------|
| Training sets| 7583               | 10721             |
| Testing sets | 1841               | 2574              |

Before dividing the training data set, we will enhance the data. The Mosaic data enhancement is introduced. Simply put, the four pictures are sorted by random scaling, random attenuation, and random arrangement.

3.2. Network training
During the experiment, the input size of the image is uniformly resized to 416*416, DIOU_Loss is used as the loss function, the Batch is set to 16, the test is set to 1, the learning rate is set to 0.001, and the maximum number of iterations is set to 28000, and The number of iterations is reduced to one-tenth of the original when the number of iterations is 25000 and 28000, so as to complete the network training of the improved YOLOv3 model, its LOSS change process, and the change curve of IOU are shown in Figure 5.
3.3. Quantitative evaluation

Commonly used target detection evaluation indicators include AP, Precision (accuracy), Recall (recall rate), IOU (intersection ratio) and other indicators to determine the detection performance of the two evaluation algorithms.

TP represents the number of matches between the predicted box and the true label box; FP means that the background is predicted as the number of targets; FN represents the number of targets to be detected but not detected by the model; TN means that it is a background, and the model has not detected the number of targets.

Precision precision is also called accuracy, and the calculation formula is:

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

Recall recall rate is also called recall rate, the calculation formula is:

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

Precision changes with the change of Recall, so AP is often used as the performance indicator of the comprehensive evaluation algorithm. IOU is called the intersection and union ratio, which refers to the ratio of the intersection and union between the box predicted by the model and the real box. In our target detection task, if the predicted frame of our algorithm and the real frame are cross-matched, then our algorithm is OK, indicating that our predicted frame is accurate, this threshold is 0.5, you can set the higher the setting, the more accurate the bounding box. In YOLOv3, this IOU is used to filter our a priori box, and the IOU is also used for threshold judgment when calculating the model effect mAP of the test set.

This article is the training process carried out on GeForceRTX2080x4GPU. The model will be evaluated on average IOU, AP and other indicators, as shown in the following table:

| Model               | Input Size | AP   | Average IOU |
|---------------------|------------|------|-------------|
| Improved YOLOv3 model | 416*416    | 88.79% | 79.07%      |
| YOLOv3 model        | 416*416    | 68.73% | 61.61%      |

The quantitative evaluation results show that the model used in this article can improve the behavior of smoking detection. In order to further visually display and evaluate the effect of the test, the test results performed are shown in the following Figure6. Observing the test results, we can see that the improved YOLOv3 model can well detect the behavior of smoking, and all indicators have been improved.
4. Conclusion
This article analyzes the model structure of YOLOv3, uses the modified model for the smoking detection task, and conducts network training. During the experiment, through quantitative and more intuitive observation of the test image, the experimental results are obtained. The improved YOLOv3 in this article can greatly improve the MAP, AP and other indicators of the detection task.

References
[1] Chao, W.(2013) Research on gesture recognition algorithm for smoking behavior. Yanshan University, Qinhuangdao.
[2] Pan G Z, Yuan Q, Fan C X, et al.(2015)Cigarette-smoke detection based on Gaussian mixture model and frame difference. Computer Engineering &Design, (5): 1290—1294.
[3] Tsai H C, Chuang C H, Tseng S Pet al. (2013)The optical flow-based analysis of human behavior-specific system. In: International Conference on Orange Technologies. New York. pp. 214-218.
[4] Ren S, He K, Girshick R, et al.(2015)Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. Vancouver. pp.91-99.
[5] Liu W, Anguelov D, Erhan D, et al.(2016)Ssd: Single shot multibox detector.in: European conference on computer vision.Germany.pp.21-37.
[6] RedmonJ, Divvala S, Girshick R, et al.(2016)You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. New York. pp. 779-788.