International Journal of Cognitive Informatics and Natural Intelligence
Volume 17 • Issue 1

An Intelligent Detection Approach for Smoking Behavior

Jiang Chong, Hunan Woman's University, China*

ABSTRACT

Smoking in public places not only causes potential harm to the health of oneself and others, but also causes hidden dangers such as fires. Therefore, for health and safety considerations, a detection model is designed based on deep learning for places where smoking is prohibited, such as airports, gas stations, and chemical warehouses, that can quickly detect and warn smoking behavior. In the model, a convolutional neural network is used to process the input frames of the video stream which are captured by the camera. After image feature extraction, feature fusion, target classification and target positioning, the position of the cigarette butt is located, and smoking behavior is determined. Common target detection algorithms are not ideal for small target objects, and the detection speed needs to be improved. A series of designed convolutional neural network modules not only reduce the amount of model calculations, speed up the deduction, and meet real-time requirements, but also improve the detection accuracy of small target objects (cigarette butts).

KEYWORDS

Computer vision, Real-time, Robustness, Small object detection, Smoking detection

INTRODUCTION

With the continuous advancement of technology, smoking detection methods have also been continuously improved. Traditional smoking detection methods are usually detected by physical means such as smoke sensors and wearable devices. Mobile health technologies are being developed for personal lifestyle and medical healthcare support, of which a growing number are designed to assist smokers to quit (Ortis et al., 2020). However, these methods have many limitations: one is that the concentration of smoke in outdoor scenes is greatly diluted and cannot be sensed by the smoke sensor; the other is that wearable devices are expensive to perform detection and need to be owned by everyone. In addition, the movement trajectory and speed of multiple parts of the limbs are judged in this method, the pattern is match with the smoking behavior, and then the matching degree is judged through machine learning classification methods such as support vector machine (SVM). The detection accuracy and efficiency of this type method are relatively low (Senyurek et al.,2019).

DOI: 10.4018/IJCINI.324115 *Corresponding Author

In addition to using physical equipment to detect smoking, some scholars detect smoking by using traditional graphics object detection methods. This type of method is divided into three steps (Wu et al., 2010): First, different sizes and step length sliding windows are set, and then all the windows are slided in each position on the image. For each window, the feature of the object to be measured is extracted through the histogram of oriented gradient (HOG) or scale-invariant feature transform (SIFT) method, and finally the classification algorithm is used for each sliding window to perform classification, such as SVM, Adaboost, etc., and the sliding window with the highest score is selected as the detection result. However, this type of method has the following disadvantages: firstly, the detection effect is not ideal, it is easy to be interfered by other objects, and the positioning is not accurate, relying on the preset sliding window size and sliding step length; secondly, there is a large amount of calculation in this method, and it needs to perform feature processing and classification judgment for each sliding window; finally, the method and process of manually extracting features are more complicated and do not have generalization.

With the rapid development of computer and video processing technology, the intelligent off-site law enforcement of taxis has become possible. However, there is still a lack of intelligent analysis technology for illegal taxis. An automatic detection algorithm is proposed for smoking behavior (Huang, Jia, Liu, 2020). First, the proposed brightness screening rules are used to reduce the processing time of the image enhancement part; secondly, Haar-Adaboost and the proposed segmented histogram matching algorithm are combined to realize the recognition of the taxi window area; a set of representative features are designed to identify smoking smoke and smoke shaking actions, including the movement trajectory of the center of mass of the smoke, the area growth rate, the ratio of the smoke convex hull to the contour circumference, the area ratio of the circumscribed rectangle within the contour, and the frequency and time interval of the smoke shaking, and the support vector machine is finally used for feature classification. In order to detect smoking behavior in time and make accurate state judgments, a smoking behavior detection algorithm is proposed based on multi-task classification (Cheng et al., 202). This algorithm combines multi-task convolutional neural networks, cascaded regression and residual networks, multi-task convolutional neural network algorithm and the regression tree method based on gradient enhancement learning (RET cascade regression) are used to quickly locate the region of interest (ROI) in the mouth; on this basis, the residual network is used to detect the target and identify the state in the ROI. In view of the slow speed, false detection rate and high hardware occupancy of the classic convolutional neural network smoking detection algorithm, a fast smoking detection is proposed based on faster region with convolution neural networks (Faster R-CNN) (Han & Li, 2020). The face is detected and the detected face image is used as the cigarette detection area to reduce the target detection area and filter out targets similar to cigarettes. The image segmentation method is used to conduct a preliminary cigarette inspection on the face area and to determine whether there is a cigarette. The Faster R-CNN algorithm is used to detect the cigarette target on the image, that initially judges that there may be cigarettes and determine whether there is smoking behavior.

The AlexNet network model was born in 2012 and won the ImageNet image classification competition in that year (Krizhevsky, Sutskever, & Hinton, 2012). As a result, both academia and industry have paid extensive attention to the application of deep learning in the field of computer vision. Such as face recognition, vehicle detection, etc. In this paper, the smoking detection problem is classified as a target detection problem, that is, the position relationship between pedestrians and cigarette butts is located to determine whether there is smoking. In this paper, a lightweight smoking detection network model is designed by drawing on the high-performance detection algorithm of YOLO (you only look once) (Redmon et al., 2016). Multi-level and different feature map vectors are combined in the model, the attention mechanism module and the disabled difference module and SPP (spatial pyramid pooling) module are increased, the original network structure is improved, and the detection accuracy of small targets is improved. At the same time, it reduces the convolution kernel parameters of the model, thereby reducing the amount of model calculations and speeding up the final

model deduction speed, the requirements of real-time detection are meet. Aiming at the problem of model robustness, the robustness of the model is improved by training data enhancement, changing the loss function and activation function, adding regularization methods, and using context information.

RELATED WORK

Smoking detection belongs to target tracking detection. Traditional target detection algorithms include VJ (Viola and Jones) cascade detectors (Viola & Jones, 2001), HOG detectors (Dalal & Triggs, 2005), and DPM (deformable parts model) models (Felzenszwalb, Mcallester, & Ramanan, 2008), etc. They have large amounts of calculation, complicated manual feature extraction, the weak characterization performance and the poor generalization ability of the model, these make it difficult to solve the smoking detection problem in different scenarios. As a "natural" filter, the convolution kernel in the convolutional neural network has superior feature extraction capabilities, which is also one of the main factors for its disruptive breakthrough in the field of computer vision. In addition, the data sets of multiple scenarios are used for training, the convolutional neural network model has a strong generalization ability. Therefore, deep learning has become the preferred solution in the field of target detection.

Classic feature extraction networks are used in target detection such as VGG (visual geometry group) (Simonyan & Zisserman, 2014), GoogLeNet (Szegedy et al., 2016), ResNet (He et al., *2016)*, etc. In the field of image classification, their applications have achieved remarkable results in pre-training network structures, this is because of its powerful feature extraction capabilities, it can complete difficult multi-image classification tasks through the large number of extracted features. Target detection also requires a large number of image features, so the backbone of the detection model usually uses *GoogLeNet's Inception* structure, *ResNet's* residual structure, etc., which can not only avoid problems such as the disappearance of the gradient when the neural network backpropagates and updates the weights, it can also speed up the model convergence.

Since the development of target detection algorithms, there have been two major schools, two-stage and single-stage detection algorithms. Representative algorithms of the former mainly include Faster RCNN (region convolutional neural networks) (Ren et al., 2015), FPN (feature pyramid networks) (Lin et al., 2017), RFCN (region fully convolutional networks) (Dai et al., 2016), and Cascade RCNN (Cai & Vasconcelos, 2017). Faster RCNN is taken as an example, this type of algorithm first extracts image features through a basic convolutional neural network and outputs feature maps. Then in the RPN (region proposal network) network, softmax is used to predict 2×k scores for each position in the input feature map, where k is the number of anchors preset in this article, and 2 represents the foreground and background. At the same time, border regression is used to predict the position of each feature map with $4 \times k$ coordinate regression feature matrix, the anchor box of the foreground sample is closer to the true value through transformation, and then in the candidate network layer (proposal layer), non-maximum value suppression (non maximum suppression, NMS) (Neubeck & Gool*,* 2006*)* and score sorting are used to screen and generate region proposals. The candidate region and the previously obtained feature map information are integrated, a proposal feature map is generated through ROI (region of interest) pooling, and it is transfered to the fully connected layer, the final object classification and frame regression positioning are completed. Representative algorithms of the latter include YOLO, SSD (single shot multibox detector) (Liu et al., 2016), RetinaNet (Lin et al., 2017), and EfficientDet (Tan et al., 2020). YOLO is taken as an example, this type of algorithm converts the classification problem into a regression problem without the need to extract the candidate region step, but directly obtains the location and category of the target through the convolutional neural network. After the basic convolutional neural network extracts image features, it directly performs target classification and frame regression positioning on each feature map, and anchor frames also are used to accelerate the frame regression, and the output vector is subjected to non-maximum suppression to obtain the final prediction result. The two types of algorithms have

their own advantages and disadvantages. The single-stage algorithm has a faster model deduction speed, but it is slightly inferior in terms of prediction accuracy. In contrast, the two-stage detection algorithm has a higher target detection accuracy, its model deduction speed is slower.

Vision-based smoking detection is susceptible to interference from image noise, which leads to false detection, and the target of cigarette butts is small, which is difficult to find and identify. Therefore, there are few smoking detection methods based on target detection in the academic world, and the related work and theories are not perfect. Based on the basic idea of target detection theory, in this paper, cigarette butts are regarded as the target to be inspected, it is also suitable for electric spark detection. The structure of the convolutional neural network is designed, and it is trained in our data set. Compared with the classic deep learning detectors YOLO, SSD, Faster RCNN, it has higher detection accuracy and detection speed for the detection of smoking behavior. In addition, in some public data sets, the algorithm model of this article also has better performance.

SMOKING DETECTION DIFFICULTIES AND SOLUTIONS

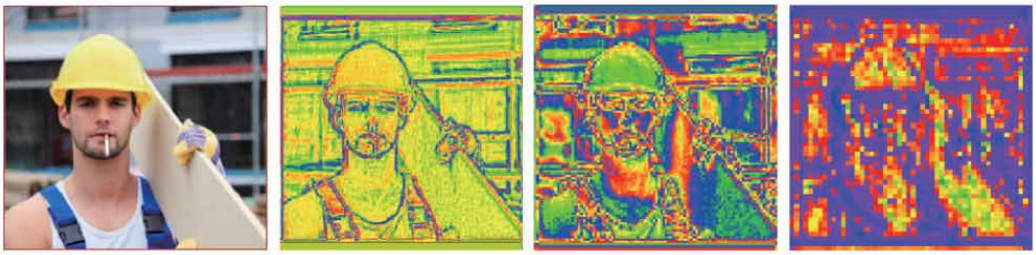
The method of completing smoking detection through deep learning target detection has the following difficulties:

Small Target Detection

First, it is necessary to give relevant definitions for small target objects. In the Microsoft COCO data set, there is a description of small objects, which refers to objects with a target area less than 32×32, and the unit is pixel. Due to the low resolution of small target objects, it is more prone to interference from blur, jitter in the picture shooting process than large objects, and its anti-noise ability is weak. Even for common graphics noise such as salt and pepper noise, Gaussian noise, it is also easy to cause a greater degree of interference to the target object, and it is difficult to completely restore the characteristics of the small object target through denoising means. Secondly, small target objects are restricted by their own size and carry less graphic information, so in the process of extracting features, very few features can be extracted.

For the problem that small target objects are difficult to find, the shallow feature map vector of the model can be used. Feature map is a basic concept related to target detection. After the input image passes through a convolution kernel or pooling kernel, the output two-dimensional matrix vector is the feature map. Another concept closely related to the feature map is the receptive field. The so-called receptive field refers to the size of the region mapped on the input image by the output vector of each position on the feature map output in each layer of the convolutional neural network. The size of the receptive field is positively correlated with the size of the core in the convolutional layer or pooling layer that the original input image has passed through. The larger the receptive field, the more "feeling" the area of the vector at each position. Larger it is, the more it can capture deep-level, high-dimensional hidden features and features of large objects; on the contrary, feature maps of small receptive fields usually capture the features of shallow details and features of small objects. For the convolutional neural network, the feature map of the shallow network has a smaller receptive field, and smaller target objects can be found. In the same way, if the model has more convolutional neural network layers, the deeper feature map receptive field is larger, and a vector in the feature map represents the feature of the larger area of the original image, then the surrounding background of small objects or the features of other objects are incorporated into the feature representation, so that the model loses the ability to find small objects. Shallow networks usually extract features that tend to be more detailed, including edges, shapes, etc., while deep networks extract more abstract features, as shown in Figure 1 (the left image is the original image, from left to right where the feature image is that the depth of the convolutional layer increases sequentially). In this regard, the idea of the FPN network can be used for reference, and the deep and shallow feature maps can be merged, which not only preserves the features of small objects as much as possible, but also can detect objects under

Figure 1. Feature maps of different scales

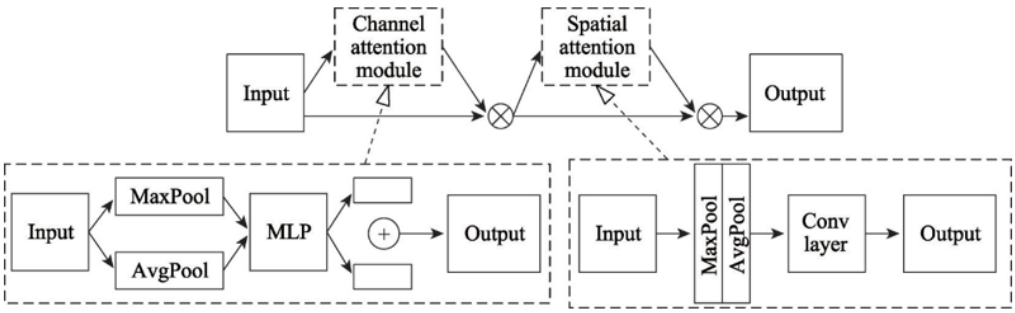


different scales well. The fusion here refers not to the addition of the vector values corresponding to the two-dimensional space of the feature map, but the expansion at the dimensional (channel) level, which can be understood as the "stacking" of multiple feature maps.

The attention mechanism can also enhance the model's ability to detect small objects. The attention model was initially applied to machine translation tasks. In 2017, SENet (squeeze-and-excitation networks) won the last ImageNet image classification competition through the designed attention module (Hu et al., 2020), marking the attention mechanism successful luck in the field of computer vision. In 2018, on the basis of SENet, the attention model CBAM (convolutional block attention module) designed an attention module that combines the two dimensions of spatial position and feature channel (Woo et al., 2018), and achieved better results, as shown in Figure 2.

The feature channel represents the number of feature maps in a certain layer, which is equal to the number of convolution kernels in this layer. Assuming that the input of the feature map is c (the number of feature channels) × h (the height of the feature map) × w (the width of the feature map), average pooling and maximum pooling of h and w of the feature map are performed into one dimension, and then these two c × 1 × 1 output vectors are added to obtain a c × 1 × 1 output vector, which is applied to the original input feature map, a convolution multiplication operation is performed to enhance the attention of the feature channel dimension. The feature attention mechanism of spatial location applies average pooling and maximum pooling to the input feature map respectively, and then stitches the two in the feature channel dimension. At this time, the feature map becomes 2c × h × w, and then it is input to a convolution layer with the number of convolution kernels c, and finally the original feature map is convolved and multiplied to enhance the attention of the spatial position. Experiments have proved that the attention module can help the convolutional neural network to extract more robust features (Hu et al., 2020; Woo et al.,2018). In this paper, the attention module is designed to solve the problem that small target objects are difficult to capture. Its design draws on the design ideas of CBAM. The upper

Figure 2. Convolutional attention module



part of the module is enhanced by fusing the feature map vectors through the maximum pooling layer and the average pooling layer. In addition, the attention of the spatial location is enhanced in the lower part by adding the feature map vectors that have passed through the maximum pooling layer and the average pooling layer, the attention of the feature channel dimension is enhanced.

Lightweight Feature Extraction Network

Smoking detection needs to pay attention to real-time problems, and it is necessary to detect and warn smoking behavior in time. Because the smoking action and process are relatively short, if the detection is not performed in real time and the response is made in time, it is easy to miss the detection. Object detection is a combination of two subtasks, namely image classification and frame positioning, both of which require a large number of features of the object to be inspected. Therefore, the detection model usually consists of two parts, the backbone network and the heads that perform the detection. The backbone network usually uses a large number of convolutional layers to extract features, and the head uses the extracted features to complete the target location and classification process. The time loss of model deduction mainly lies in the backbone network. The reason why the scale of the backbone network is getting larger and larger is to learn more complex nonlinear mapping relationships and extract more potential features. However, as the number of layers and the amount of parameters increase, the amount of calculations inevitably increases, so it is necessary to design a lightweight feature extraction network on the basis of ensuring that sufficient target features can be extracted.

In fact, to deploy detection models in some scenarios, the computing power of hardware resources needs to be considered, which is another important factor in designing lightweight feature extraction networks. For example, when the model is deployed on a development board such as Nvidia TX2, since the computing power of the development board is far less than that of a dedicated graphics card for deep learning, the deduction speed of the model will be further slowed down, so the design of a lightweight feature extraction network is more effective significance. In this paper, the residual module and the SPP module are designed and combined with the common convolutional layer, a backbone feature extraction network is formed. Compared with the backbone feature extraction network composed of conventional convolutional neural network layers, the backbone network in this paper improves the feature extraction ability without increasing the amount of parameters.

The residual module ResNet has achieved remarkable results in image classification tasks. Its specific structure is shown in Figure 3. ResNeXt borrows and inherits the residual idea of ResNet (Xie et al., 2017), and it proves its proposed grouping multi-cardinality path through experiments, the structure can learn more features without increasing the complexity of the parameters, the classification accuracy is improved and the number of model hyperparameters is reduced. Therefore, as shown in Figure 4(a), in this paper, the ResNeXt module is used for reference, and the residual module is designed in the backbone network of the feature extraction of the detection model, instead of the traditional simple convolutional network layer. The module first completes the down-sampling process of the feature map through a convolution operation with a convolution step length of 2, and then it refers to the ResNeXt structure, 32 sets of multi-path structures are established to complete grouped convolutions. At the same time, the stacking of the convolution of two residual groups is used, and the feature learning ability of the network is further enhanced through the double-layer residuals, the gradient explosion and gradient disappearance are avoided.

SPPNet (Spatial Pyramid Pooling) was originally designed to solve the problem that when the convolutional layer and the fully connected layer are connected, the input feature maps of different sizes cannot produce fixed-length feature representations, resulting in the failure of the convolutional layer output and the fully connected layer input connection (He et al., 2014). In this paper, the potential feature fusion feature of SPPNet is used, and the maximum pooling kernel of non-synchronization length is used to act on the input feature map, the output feature maps of different sizes of receptive fields are obtained, and then through the vector splicing of feature channel dimensions, multi-level features are combined to strengthen the network's ability to learn abstract features and deep semantic

Figure 3. ResNet module (left) and ResNeXt module (right)

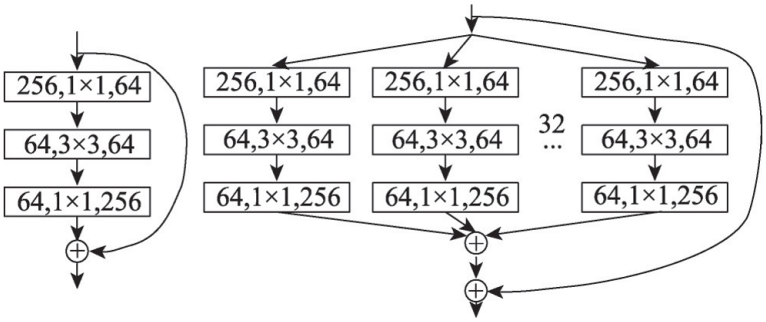
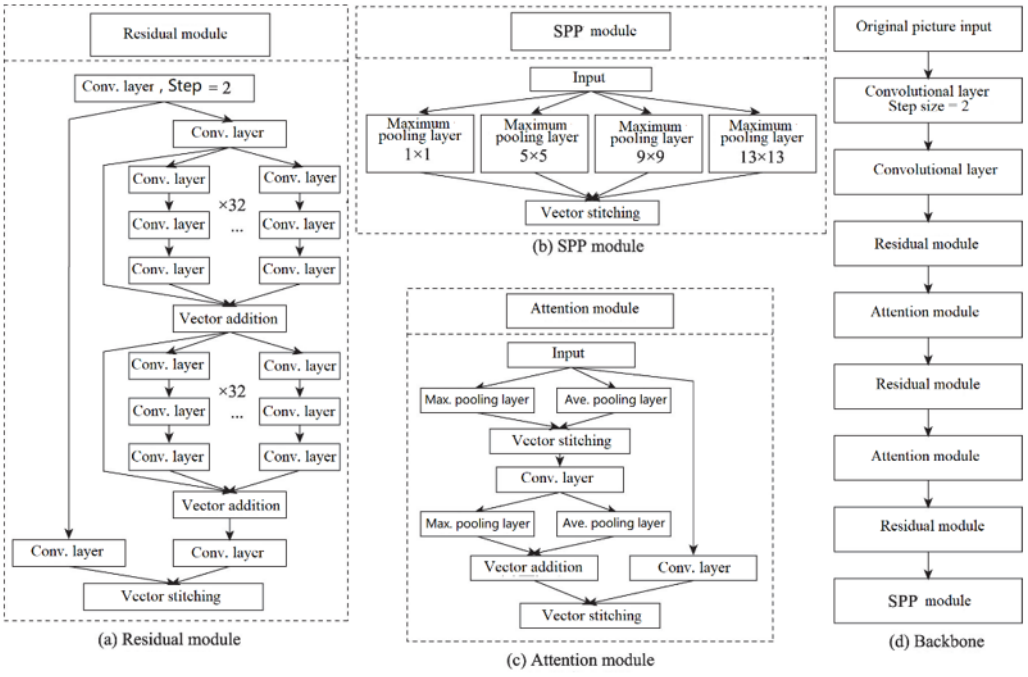


Figure 4. Basic network module of detection model



features. Compared with performing vector stitching after performing multiple convolution operations, the same effect is achieved while significantly reducing the complexity of model operation and the amount of parameter calculation. Therefore, the SPPNet module is used for reference, at the end of the backbone network, the SPP module is designed in this paper. As shown in Figure 4(b), it contains four different-sized pooling core modules. After the input feature maps pass through the four pool layers, four types of feature maps with different areas of receptive fields will be generated. Among them, the feature map with 1×1 pooling core has the smallest receptive field, and the feature map with 13×13 pooling cores has the largest receptive field range, it is proportional to the size of the pool core. The SPP module makes up for the shortcomings of insufficient sampling times of the overall network by fusing the feature map vectors of four different receptive fields, and it is conducive to discovering the overall characteristics of the target object and deep semantic features.

Model Robustness

In addition to the main body of the deep learning target detection algorithm model consisting of a convolutional neural network, it also includes conventional components, such as activation functions, loss functions, and regularization methods. The activation function provides the model with nonlinear modeling capabilities. The convolution and pooling operations are just matrix operations, only linear modeling and calculation of high-dimensional spaces, and the activation function applies the nonlinear mapping of input and output to make the nerve network model learns and performs nonlinear function fitting. The sub-problems of target detection can be attributed to target classification and coordinate regression positioning. Neither the classification nor the regression function can be just a linear function. Therefore, it is necessary to choose a stronger activation function Mish to replace the original activation function *(Misra, 2019)*. The Mish activation function is equation (1):

$$f\left(x\right) = x \times \tanh\left(\ln\left(1+e^{x}\right)\right) \tag{1}$$

The loss function measures the quality of the model's prediction, and measures the difference between the predicted value and the real value, which is the target of model training. The loss function describes the problem to be solved in this article through formulas, and the loss function of target detection can be divided into two parts, classification loss and border regression loss. Classification loss function selects the binary cross entropy (BCE) loss function in the YOLO algorithm. The loss function of this part includes two parts, confidence loss and category loss, as shown in equation (2):

$$\begin{aligned} L_{class} &= -\lambda_{obj}\sum_{i=0}^{S\times S\,anchor-1}\sum_{j=0} I_{ij}^{obj}\left[t_{conf}\ln\left(pred_{conf}\right)+\left(1-t_{conf}\right)\ln\left(1-pred_{conf}\right)\right] \\ &-\lambda_{noobj}\sum_{i=0}^{S\times S\,anchor-1}\sum_{j=0} I_{ij}^{noobj}\left[t_{conf}\ln\left(pred_{conf}\right)+\left(1-t_{conf}\right)\ln\left(1-pred_{conf}\right)\right] \\ &-\lambda_{obj}\sum_{i=0}^{S\times S\,anchor-1}\sum_{j=0} I_{ij}^{obj}\left[t_{clsf}\ln\left(pred_{cls}\right)+\left(1-t_{cls}\right)\ln\left(1-pred_{cls}\right)\right] \end{aligned} \tag{2}$$

Wherein, S represents the size of the output feature map, *anchor* represents the number of anchor frames that each feature map vector is responsible for predicting, λ_{obj} and λ_{noobj} represent the penalty factor, t_{conf} and $pred_{conf}$ represent the true value and predicted value of the object confidence, respectively, t_{cls} and $pred_{cls}$ represents the true value and predicted value of the object's category, respectively. I_{ij}^{obj} represents whether there is an object to be tested in the j-th anchor box at the i-th position. If there is an object, its value is 1, and if it does not exist, it is 0.

The early target detection algorithm YOLO uses the MSE (mean square error) loss function, and Faster RCNN uses the L_1-smooth loss function as the border regression loss function. However, the L_n norm is not accurate to measure the regression loss. In 2019, the DIoU (distance intersection over union) loss function was proposed (Zheng et al., 2020), which makes the frame regression process of target detection faster and more accurate than the previous loss function. The loss function is as equation (3):

$$L_{regression} = 1 - \frac{\left|B\cap B^{gt}\right|}{\left|B\cup B^{gt}\right|} + \frac{\rho^{2}\left(b,b^{gt}\right)}{c^{2}} \tag{3}$$

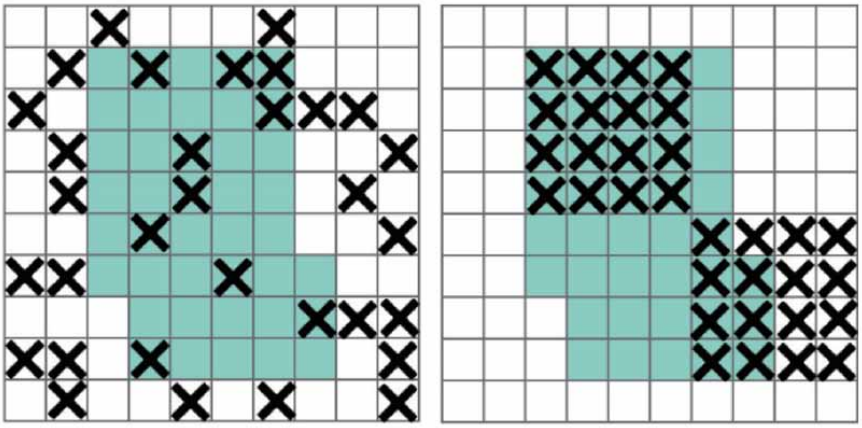
Wherein, B and B^{gt} represent the predicted value and true value of the frame respectively, $B \cap B^{gt}$ and $B \cup B^{gt}$ represent the intersection area and union area of the two, b and b^{gt} represent the center points of B and B^{gt}, and $\rho^2(\cdot)$ represents the Euclidean distance, and c is the diagonal length of the smallest rectangular box covering B and B^{gt}. From this, the overall loss function of the detection model is obtained as equation (4):

$$L_{detection} = L_{class} + L_{regression} \tag{4}$$

The regularization method commonly used in machine learning is Dropout [4]. The input neurons are ignored with a certain probability during the forward propagation of the neural network. By simulating the phenomenon of human forgetting, the model is prevented from overfitting and the model is more robust. Dropout is usually widely used for regularization of fully connected layers, but its effect on the convolutional layer is not obvious. The activation units in the convolutional layer are spatially related. Even if Dropout randomly discards the vectors in some positions of the feature map, the object information can still be transmitted to the next layer through the convolutional network. Therefore, in this paper, the DropBlock method is used to complete the regularization constraint on the feature map, and with a certain probability, the neighboring block in the map is ignored instead of a certain point, as shown in Figure 5.

For deep learning algorithms, the importance of the data set is self-evident, and the quality of the data set determines the quality of the detection model. Generally speaking, the larger the amount of data, the more scenes to be tested are included, the stronger the generalization performance of the detection model, and the detection accuracy will increase accordingly. Through training data enhancement techniques such as Mixup, CutMix *(*Yun et al., 2019*)*, multi-scale scaling, translation, rotation and symmetry, etc., the diversity of training data is further improved and the model training data is prevented from being too single. Finally, the context information is used to set two categories to be detected, pedestrian and smoke. After the output of the final prediction result vector is suppressed by non-maximum value, by first selecting the pedestrian detection frame with greater confidence, the prediction vector with the smoking detection frame and the pedestrian detection frame IoU (intersection over union) less than 0 is eliminated in reverse, the false detection rate of cigarette butts is redued, the detection accuracy is improved.

Figure 5. Dropout regularization (left) and DropBlock regularization (right)



MODEL STRUCTURE DESIGN

This chapter will introduce the overall smoking detection model structure and detection process which are proposed in this paper. The model draws on the single-stage detection idea of the YOLO algorithm, and the deep learning model is directly used to perform target classification and border regression positioning of the object to be measured, the position and category of the object are obtained .

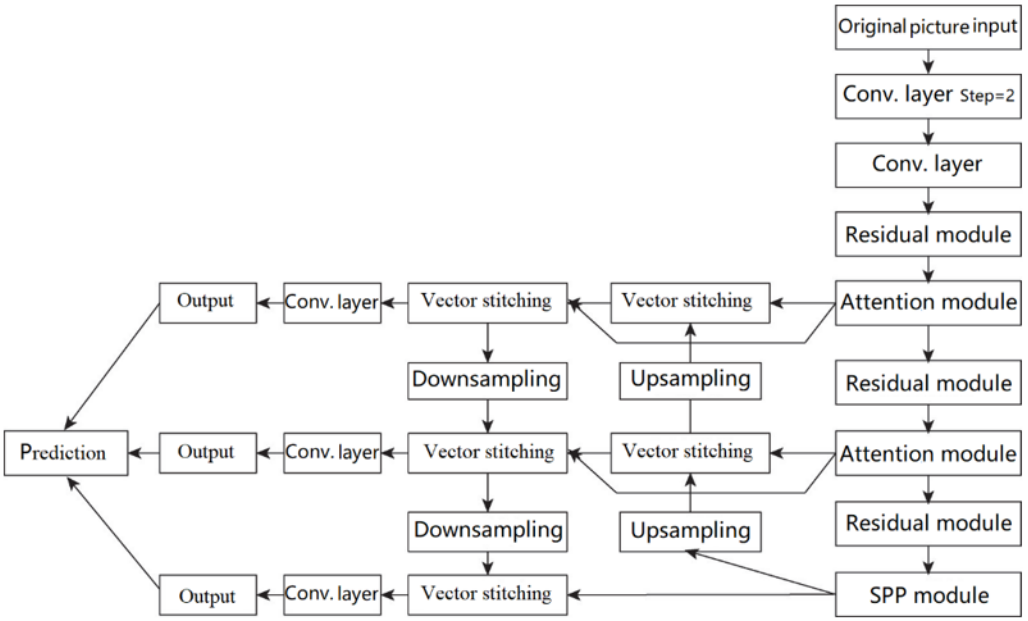
Overall Network Structure

The backbone network part of feature extraction is shown in Figure 4(d), which is composed of convolutional layer, residual module, SPP module and attention module, it learns and extracts features of the object to be measured. Figure 6 is the overall network architecture of the smoking detection model in this paper. This paper draws on the BiFPN (bidirectional feature pyramid network) structure in Google's EfficientDet model, and it combines the output vectors of feature maps of three different receptive field scales. The size of the down-sampled feature map is correspondingly reduced to half of the original feature, and the size of the feature map is enlarged through the up-sampling method of bilinear interpolation, in order to complete the vector mosaic of the feature channel dimension with the original feature map. Finally, three-scale output results are obtained to cover cigarette targets of different sizes as much as possible. Non-maximum suppression algorithms, context information calculation processing, and prediction confidence ranking are used to process and summarize the three-layer output vector, interference items are eliminated, the final target category and position are obtained in the image.

Smoking Detection Process

In this paper, the real-time streaming protocol (RTSP) is used to read video frames at a certain frame interval, and the frames are simply pre-processed and input into the smoking detection model designed in this paper. The model is used to calculate the cigarette butt and the location of the

Figure 6. Overall network structure of detection model



pedestrian with high confidence. First, NMS is used to eliminate redundant detection frames, and then context information correlation algorithm is used to calculate the IoU of each cigarette butt prediction coordinate and the pedestrian prediction coordinate, and exclude the prediction box whose calculated value is less than a certain threshold, which means that the cigarette butt is not related to the pedestrian. There is no smoking. Finally, the captured video frames of smoking phenomenon are marked and saved locally, and warning information is issued to relevant personnel. The specific detection process is shown in Figure 7.

EXPERIMENT AND PERFORMANCE ANALYSIS

Due to the lack of a corresponding smoking detection data set, the production of the relevant data set is first completed, the data set is used to train the detection model, the deep learning training framework is Darknet, the graphics card model for training is NVIDIA GeForce RTX 2080Ti, and the operating system is 64-bit Ubuntu Kylin 16, configure CUDA10.0, CUDNN7.5.

Data Set Production

In the data set produced in this article, the data comes from crawled Google and Baidu images, smoking clips in the HMDB human behavior database, and smoking videos recorded by myself. Data annotation uses Github's open source annotation tool YOLO Mark. The content of the target label is a five-tuple (*class, x, y, w, h*), where *class* represents the object category, which is an integer of type Int; $0 \leq x, y, w, h \leq 1$ represents the center point coordinates of the target object, as well as the ratio of the height and width of the labeled truth box in the original input image. The labeled values are processed through the normalization method to facilitate subsequent model deduction and calculation. The visualization of data labeling is shown in Figure 8, and the data set is divided as shown in Table 1. Disturbance items refer to samples without any smoking targets.

Experimental Results and Analysis

After the Darknet training framework is used to build the detection model of this article, the experiment is first started on the self-made data set of this article.

First, the data preprocessing process is carried out. The data enhancement methods include multi-scale scaling, translation and rotation, symmetry, random erasure, and CutMix. The number of iterations in the whole training process is 30,000, the batch size is 64, the Adam gradient optimizer is used, the initial learning rate is 0.001, the weight attenuation coefficient is 0.000 5, the regularization method uses Dropblock, and the loss function is the formula (4) in this article. Figure 9 shows the change of the loss value of this model when it is trained on the self-made data set and the change of the mAP of the training set and the validation set. It is found that the detection effect of the model on the validation set is very good. Figure 10 shows the detection effect of the trained model in the real scene.

On the data set made in this article, experiments were carried out by using YOLOv3, SSD, RetinaNet and the target detection algorithm proposed in this article. The detection results are shown in Table 2. Through the test experiment on a host equipped with Nvidia GeForce RTX 2080Ti graphics card, it can be seen that the detection algorithm proposed in this article is higher in detection accuracy (mAP) and model deduction speed (FPS) than the YOLOv3 algorithm and other representative single-

Figure 7. Flow chart of smoking detection

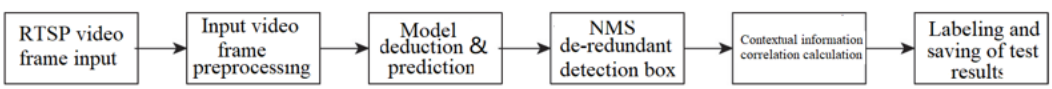
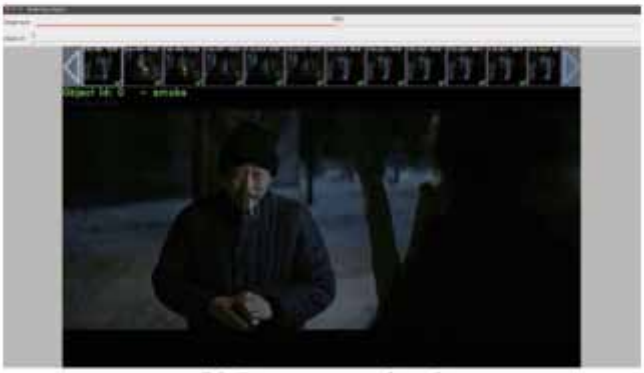


Figure 8. Data annotation

(a) Data annotation 1



(b) Data annotation 2

Table 1. Settings of dataset

Data content	Training set	Validation set	Test set
Number of smoking samples	3100	200	300
Number of interference items	500	20	50

stage detection algorithms. It shows that the algorithm in this paper has a corresponding improvement in the overall smoking target detection accuracy and detection speed.

After that, the proposed model in this article was used for training and testing on the public data set PASCAL VOC. A TITAN graphics card was added when training the model, and dual graphics cards were used for training. Table 3 describes the training and detection results of this algorithm on the PASCAL VOC dataset. The deduction process uses GeForce RTX 2080Ti graphics card. Compared with classic detection algorithms such as YOLOv3, SSD, and Faster RCNN, the algorithm in this paper still has certain advantages.

In order to verify the improvement of the algorithm in this paper in small target detection, model training and testing were carried out on the public data set Tsinghua-Tencent 100K (TT100K) traffic sign detection data set. This data set includes 3 types of traffic signs, namely prohibition signs, warning signs and indication signs, so the label category of the training data is set to 3. The training set of the data set has 6 107 pictures, and the test set has 3 073 pictures. After that, the YOLOv3

Figure 9. Training log

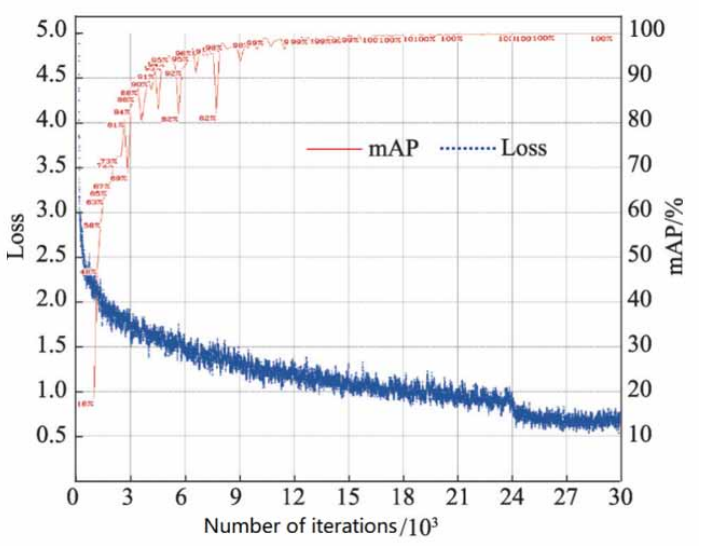


Figure 10. Detection results

(a) Detection result 1

(b) Detection result 2

Table 2. Detection results of different algorithms on proposed dataset

Algorithm	Backbone	mAP/%	FPS
YOLOv3	Darknet-53	78.7	94
SSD	VGG-16	75.4	62
RetinaNet	Resnet-50	84.6	47
Proposed	Proposed backbone	86.3	103

Table 3. Detection results on PASCAL VOC

Algorithm	Backbone	mAP/%	FPS
YOLOv3	Darknet-53	76.8	91
SSD	VGG-16	74.3	67
Faster RCNN	VGG-16	73.2	17
Proposed	Proposed backbone	79.3	98

model (the main reference and comparison model of the model in this article) and the model without AttentionBlock (this part is replaced by the ordinary convolutional layer) in this paper, the model in this paper that does not use the bidirectional fusion of multi-layer feature maps (this part is replaced with the ordinary one-way FPN network structure) and the complete model in this paper are compared experimentally. The final test results are shown in Table 4.

The detection and deduction process uses the GeForce RTX 2080Ti graphics card. After experimental comparison, the algorithm in this paper uses the Attention Block module and the structure of multi-scale bidirectional feature map fusion prediction, the recognition and detection capabilities of small targets are strengthen. Figure 11 is a comparison example diagram of the algorithm in this paper and the YOLOv3 algorithm for detecting traffic signs.

CONCLUSION AND OUTLOOK

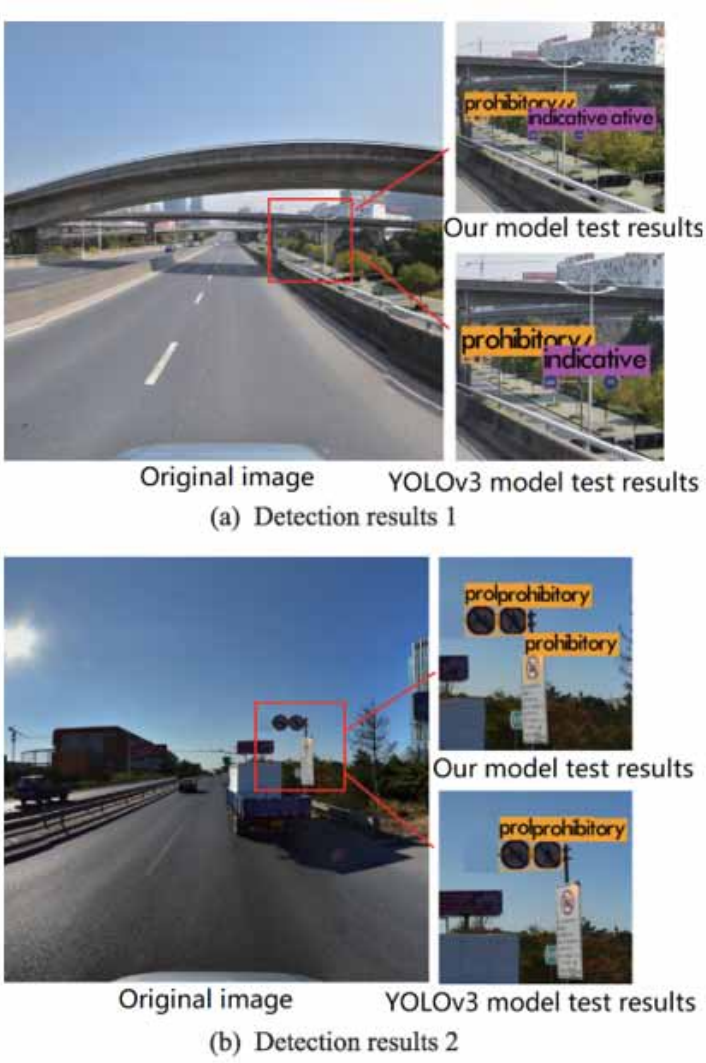
What are the hazards of smoking in public places?

1. It is easy to cause fire. A cigarette usually lasts for about 10 minutes. If the cigarette butts are randomly discarded, they can easily cause a fire if they are close to flammable materials.

Table 4. Detection results on TT100K

Algorithm	Attention module	Multi-scale bidirectional feature map fusion	mAP/%	FPS
YOLOv3	—	—	87.2	91
Proposed	√	×	90.5	104
	×	√	89.7	98
	√	√	94.3	103

Figure 11. Comparison of detection results on traffic sign



2. Pollution of the air in public places. The smoke released by tobacco burning contains more than 3,800 known chemical substances, most of which are harmful to the human body, including carbon monoxide and nicotine, which can cause various harms to the human body.
3. The harm of passive smoking. In an environment with extremely poor ventilation and exposure to a room full of tobacco smoke for only one hour, the carboxyhemoglobin in the blood of passive smokers rose from an average of 1.6% to 2.6%, which is roughly equivalent to smoking a cigarette with a medium tar content. The smoke inhaled by passive smoking contains a variety of toxic substances and carcinogens.

Based on slice computer recognition technology, the deep learning model is fully compatible with high-pixel cameras, with fine recognition details and long recognition distance. A variety of smoking behaviors' recognition are supported such as holding cigarettes and holding cigarettes in the mouth. In this paper, a detection model of smoking behavior is proposed based on deep learning, it is used in

the actual application scenarios and can quickly detect and warn smoking behavior. This model has a good detection effect on fine-grained small targets. On the one hand, in order to solve the problem of real-time model detection, the backbone network structure of the convolutional neural network is optimized for extracting image features, which not only reduces the amount of model parameters and calculations, but also speeds up the model deduction, thereby the detection speed is increased. The structure can also be used in scenarios where computing resources are limited. On the other hand, in order to improve the robustness of the model, the new activation function *Mish* (A Self Regularized Non-Monotonic Activation Function) is introduced into the convolutional layer of the model in this paper, and the regularized *DropBlock* module (A regularization method for convolutional networks) is added to the detection model to prevent the model from overfitting. Second, *DIoU* bounding box regression loss function is selected to replace the conventional root mean square error loss, the accuracy of target object positioning is improved. Finally, contextual information is used to reduce the false detection rate of target objects. The self-made data set is used to train the proposed model in this paper, the model has a good detection effect.

Leakage of electrical equipment will affect the operation of the equipment, and even cause fires to cause economic losses and threaten personal safety. The electrical sparks generated by electrical leakage are different from ordinary flames, and their flashing time is shorter and the target is small, which makes it difficult for ordinary sensors to identify. The electric spark can be detected based on image background modeling, the background interference is removed during electric spark detection, the foreground is extracted, and the foreground area is segmented by size and color characteristics, so as to identify the electric spark. Electric spark is similar to a smoking scene, and the proposed method in this article can also be applied to electric spark recognition.

Because deep learning is limited by the diversity of the data set, the effect of this model in actual production applications is not perfect. In the future, the data set will be further expanded and the model will be incrementally trained, the rate of false detections and missed detections are further reduced, and detection effect of the model's practical industrial application is improved.

ACKNOWLEDGEMENTS

The project is supported by the Scientific Research Fund of Hunan Provincial Education Department (21A0603, Target Detection and Intelligent Recognition of Metal Minerals on Microscopic Image using Deep Learning Method), China.

REFERENCES

CaiZ.VasconcelosN. (2017). Cascade R-CNN: delving into high quality object detection. *arXiv:1712.00726*.

Cheng, S. H., & Ma, X. F.(2020). Smoking Detection Algorithm Based on Multitask Classification. *Acta Meteorologica Sinica*, *41*(5), 538–543. doi:10.3969/j.issn.1000-1158.2020.05.05

Dai, J. F., Li, Y., & He, K. M. (2016). R-FCN: object detection via region- based fully convolutional networks. *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, (pp.886-893). Washington: IEEE Computer Society. doi:10.1109/CVPR.2005.177

Felzenszwalb, P., Mcallester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society. doi:10.1109/CVPR.2008.4587597

Han, G. J., & Li, Q. (2020). A rapid detection algorithm for smoking based on Faster R-CNN. *Journal of Xi'an University of Posts and Telecommunications*, *25*(2), 85–91. doi:10.13682/j.issn.2095-6533.2020.02.016

He, K., Zhang, X., & Ren, S. (2016). Deep residual learning for image recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916. doi:10.1109/TPAMI.2015.2389824 PMID:26353135

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2020). Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 2011–2023. doi:10.1109/TPAMI.2019.2913372 PMID:31034408

Huang, X. P., Jia, K. B., & Liu, P. Y. (2020). Automatic Detection of Taxi Driver Smoking Behavior Based on Traffic Monitoring. *Jisuanji Fangzhen*, *37*(12), 337–344. doi:10.3969/j.issn.1006 -9348.2020.12.070

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates.

Lin, T. Y., Dollár, P., & Girshick, R. B. (2017). Feature pyramid networks for object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society.

Lin, T. Y., Goyal, P., & Girshick, R. (2017). Focal loss for dense object detection. *Proceedings of the 2017 IEEE International Conference on Computer Vision*. IEEE Computer Society.

Liu, W., Anguelov, D., & Erhan, D. et al.. (2016). SSD: single shot multibox detector. *Proceedings of the 14th European Conference on Computer Vision*. Springer.

MisraD. (2019). Mish: a self regularized non- monotonic neural activation function. *arXiv:1908.08681*.

Neubeck, A., & Gool, L. J. V. (2006). Efficient non-maximum suppression. *Proceedings of the 18th International Conference on Pattern Recognition*. IEEE Computer Society. doi:10.1109/ICPR.2006.479

Ortis, A., Caponnetto, P., Polosa, R., Urso, S., & Battiato, S. (2020). A Report on Smoking Detection and Quitting Technologies. *International Journal of Environmental Research and Public Health*, *17*(7), 2614. doi:10.3390/ijerph17072614 PMID:32290288

Redmon, J., Divvala, S., & Girshick, R. (2016). You only look once: unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society. doi:10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031 PMID:27295650

Senyurek, V. Y., Imtiaz, M. H., Belsare, P., Tiffany, S., & Sazonov, E. (2019). Cigarette smoking detection with an inertial sensor and a smart lighter. *Sensors (Basel)*, *19*(3), 570–588. doi:10.3390/s19030570 PMID:30700056

Senyurek, V. Y., Imtiaz, M. H., Belsare, P., Tiffany, S., & Sazonov, E. (2019). Smoking detection based on regularity analysis of hand to mouth gestures. *Biomedical Signal Processing and Control*, *51*, 106–112. doi:10.1016/j.bspc.2019.01.026 PMID:30854022

SimonyanK.ZissermanA. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.*

Szegedy, C., Vanhoucke, V., & Ioffe, S.. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society. doi:10.1109/CVPR.2016.308

Tan, M. X., Pang, R. M., & Le, Q. V. (2020). EfficientDet: scalable and efficient object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Viola, P. A., & Jones, M. J. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society. doi:10.1109/CVPR.2001.990517

Woo, S., Park, J. C., & Lee, J. Y. (2018). CBAM: convolutional block attention module. *Proceedings of the 15th European Conference on Computer Vision*. Springer.

Wu, P., Heieh, J. W., & Cheng, J. C. (2010). Human smoking event detection using visual interaction clues. *Proceedings of the 20th International Conference on Pattern Recognition*. IEEE Computer Society. doi:10.1109/ICPR.2010.1056

Xie, S. N., Girshick, R., & Dollar, P. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society. doi:10.1109/CVPR.2017.634

Yun, S., Han, D., & Chun, S. (2019). CutMix: regularization strategy to train strong classifiers with localizable features. *Proceedings of the 2019 IEEE International Conference on Computer Vision*. IEEE Computer Society. doi:10.1109/ICCV.2019.00612

Zheng, Z. H., Wang, P., & Liu, W. (2020). Distance- IoU loss: faster and better learning for bounding box regression. *Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference*. AAAI. doi:10.1609/aaai.v34i07.6999

Jiang Chong received her Bachelor's degree in computer science and technology from Hunan University in 2002. Then she obtained her Master's degree in computer application technology from Central South University in Changsha, China. Now she is an researcher at the School of Computer Science and Engineering, Hunan Women's University, China, and an Ph.D. candidate at Faculty of Computer Science and Information Technology, Universiti Putra Malaysian(UPM), Malaysia. Her research interests include learning, deep learning,computer vision. E-mail: jessiejch@qq.com