Real Time Detection of Drivers' Smoking Behavior Using the Improved YOLO-V4 Model

Kaixin Zhao(✉)

College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, China
kxz9575@163.com

Abstract. Drivers' smoking behavior is one of the causes of traffic accidents. The traditional sensor-based smoking detection methods are expensive. Therefore, the method based on deep learning is used for smoking detection. However, due to the limitation of GPU computing hardware, deep learning detection algorithm is difficult to deploy. To solve this problem, this paper designs an improved YOLO-V4 lightweight target detection algorithm model, namely, SmokeNet, to detect smoking behavior, which not only has high recognition accuracy, but also can meet the conditions of real-time detection on the edge devices. First, we established a smoking data set, according to the characteristics of the dataset, we reconstructed the network structure of yolov4, reduced a large number of convolution layers, and retained only one head detection layer. Then, attention mechanism is introduced to improve the model fitting ability. Finally, we deploy the trained model in Jetson Xavier NX. Experiments show that the detection accuracy of smokenet is slightly lower than that of the original yolov4 model, but the size of the model is only 1/10 of that of yolov4, and the detection speed is increased by 57%.

Keywords: Attention · Deep Learning · Real-Time Detection · Smoking Detecion · Yolo

1 Introduction

As a part of the Driver Monitoring System (DMS), smoking detection should be judged timely and accurately. Traditional smoking detection methods include sensor-based detection as well as image-based detection. The sensor-based method mainly detects the concentration of smoke produced by cigarette burning in the air or collects and analyzes the activity or breathing signals of smokers throught wearable device sensors [4] to judge whether there is smoking behavior. As in [9], the author collects the respiratory signal and hand-to-mouth gesture patterns of volunteers through a non-invasive wearable device, and then analyzed the collected data for smoking detection, this method requires high sensitivity and high cost of the sensor. The smoking detection method based on image processing [5] is to separate cigarettes and smoke from the background, and then judge according to the specific physical features such as contour, shape, color and texture. As in [8] detected smoking by analyzing the RGB spatial color features of

K. Subramanian et al. (Eds.): CTMCD 2022, ACSR 99, pp. 126–134, 2023.
https://doi.org/10.2991/978-94-6463-046-6_16

smoke, extracting the foreground images, and analyzing the area changes and distance change of face and smoke images. These methods are easily affected by illumination and environmental conditions.

Since AlexNet was proposed in 2012 and won the champion of Imagenet Classification Competition, CNN (Convolutional Neural Network) and deep learning have attracted the attention of many researchers. In terms of target recognition, many CNN-based target detecion algorithms have been proposed. At present, object detection has made great achievements in medical treatment, agriculture and transportation, However, object detection is rarely used in smoking detection. [7] proposed an improved Faster R-CNN algorithm, which utilizes the detection of motion-specific objects to classify driver actions exhibiting great intra-class differences and inter-class similarity. [3] applied the Yolov2 algorithm to detect cigarette objects when drivers smoked. These method overcomes the problems of light interference and proves that the smoking detection method based on object detection is superior to the means based on sensor and image processing in the field of cost, accuracy and robustness. However, modern neural networks need a large number of GPUs for large-scale training and cannot work in real time. Therefore, we propose an improved yolov4 smoking detection algorithm, which can operate on the traditional GPU in in real time, and only needs a conventional GPU for training. The specific improvement methods are as follows: reduce the network depth of the backbone of yolov4 model and the size of the input image to obtain faster inferenve speed, and introduce the attention structure to improve the fitting ability of the network model. Besides, according to the characteristics of video datasets, only one head feature layer is reversed in yolov4 prediction module, which effectively avoids the detection interference caused by small target learning. The rest of this paper is designed as follows: The second part introduces the object detection and attention mechanism related work. The third part introduces some improvement methods of the model. The four part introduces the results of experiment and model comparison and Sect. 5 give conclusions and future work.

2 Related Work

2.1 Yolo Series

The object detection algorithm includes one-stage detection and two-stage detection. Two-stage object detection methods, such as Faster R-CNN [10] are mainly composed of regional recommendation network (RPN) and classification network, which divide the classification task and regression task into two parts. Although the detection accuracy is high, the model requires more time and computation to train and test. One-stage object detection methods, such as Yolo series, have faster inference speed and stronger practicability, which can meet the conditions of real-time object recognition. Yolo series adopt many excellent detection technologies to unify target classification and location into a regression problem. The state-of-art version Yolov4 [1] not only has high detection accuracy for ordinary large targets, but also can effectively detect small targets such as cigarettes, so we also built a smoking detection model based on Yolov4.

Yolo detector is usually composed of Input, Backbone, Neck and Head, as shown in Fig. 1. The Input is a color image with RGB channels. Normally, the larger the input

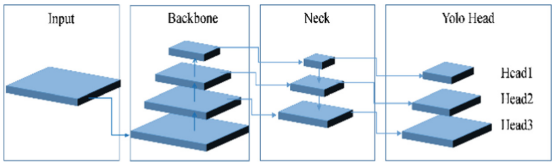


Fig. 1. The structure of Yolo.

image size, the richer the feature information and the slower the detection speed. The backbone is a network structure for feature extraction of input images. Yolov4 adopts CSPDarknet53 network structure. The neck consists of several bottom-up paths and several top-down paths, which are used to collect feature maps at different stages. The head is the output of the algorithm, which can predict the bounding box and category of the objects. In order to detect targets of different sizes and scales in the image, Yolo algorithm will adopt a method of multi-scale feature map fusion to obtain the final predicted feature map. As can be seen from Fig. 1, Yolo will detect the target on three feature scales of the picture, and finally fuse them. Take the input image size of 416 pixel $\times$ 416 pixel as an example, yolo model will generate three head layers with the size of 13 pixel $\times$ 13 pixel, 26 pixel $\times$ 26 pixel and 52 pixel $\times$ 52 pixels respectively. Head layer1, head layer2 and head layer3 are used to detect large, medium and small objects respectively. Finally, only one head layer that best matches the scale of the object will predict the result.

2.2 SE Block

Researchers have proved that introducing attention mechanism into CNN can improve the accuracy of image classification and object detection [2]. By embedding the channel attention mechanism in CNN, the receptive field of the network feature extraction layer is able to be enhanced and the network feature extraction ability can be improved.

SENet [6] introduces Squeeze and Excitation (SE) blocks, it uses global average-pooled features to compute the channel-wise attention and enhances the representation ability of feature graph in a computationally computationally efficient manner, the structure of SENet is shown in Fig. 2, its mainly composed of three steps:

Squeeze: The H $\times$ W $\times$ C size feature maps were transformed into 1 $\times$ 1 $\times$ C real columns, with the aim of giving each feature map a global receptive fields, so that the low-layer networks with lower receptive field sizes could also obtain global information.

Excitation: The real sequence was fed into the bottleneck structure composed of two full connection layers to construct the correlation between channels. The first full connection layer would reduce the dimensionality on the feature graph channel number C to obtain a 1 $\times$ 1 $\times$ (C/r) vector and, after activation of the Relu function, the second full connection layer would be promoted to the original dimension, yielding a 1 $\times$ 1 $\times$ C vector and obtaining the corresponding normalized weights using the Sigmoid activation function. These weights can be regarded as the importance of each channel after feature selection.

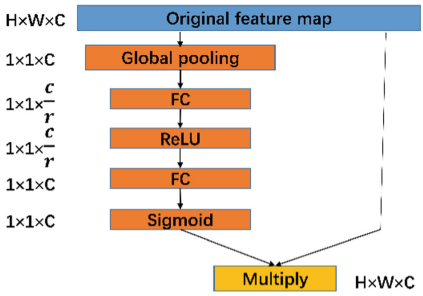


Fig. 2. Add Attention Mechanism to Feature Map.

Assign weights: The recalibration of the original features was done by multiplying by individual weights of the previous feature channels. As shown in Fig. 2, add the feature map of SE block to replace the original feature map.

3 Improvement Method and Materials

3.1 Build Datasets

At present, there is no public dataset of drivers' smoking behavior. Therefore, we take the driver smoking video in the real scene as the dataset, mainly through the camera placed on the dashboard in front of the driver. Finally, we captured 600 smoking pictures from the video. In order to increase the richness of the experimental datasets and enable the model to accurately predict under different illumination and other environmental conditions, all pictures collected were randomly preprocessed in terms of brightness, rotation, mirror and image definition, as shown in the Fig. 3. Finally, our enhanced dataset has 2,378 images, which are then numbered and annotated. Draw a bounding box and manually classify categories. We take 80% of the total number of datasets as the training set, 10% as the testing sets and 10% as the verification sets.

3.2 The Proposed Algorithm

Different head layer outputs are visualized by Grad-CAM, as shown in Fig. 4, head layer3 best matches the target scale, so it pays the strongest attention to the target and the result will be output by head layer3. In the dataset established in this paper, the size of cigarettes in the image basically does not change and all prediction results will be output by only one head layer. Therefore, we propose a detection model called smokenet. Figure 5 shows the overall architecture of the SmokeNet model. The CBL is the basic module in the network, where C represents a convolution layer, B represents a batch normalization layer and L represents a Leaky-ReLU activation function. The Res divides the low-level features into two parts, and then fuses the cross-layer features, which can effectively improve the ability of convolution feature extraction. The SE is the attention mechanism introduced. As described in Sect. 2.1, SE block can improve the fitting ability of the model.

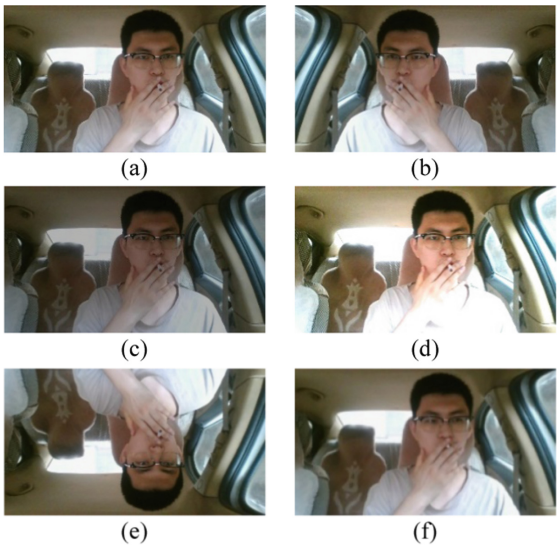


Fig. 3. Image augmentation methods: (a) original image, (b) horizontal mirror, (c) Darken, (d) Brighten, (e) 180° clockwise rotation, and (f) blur processing.

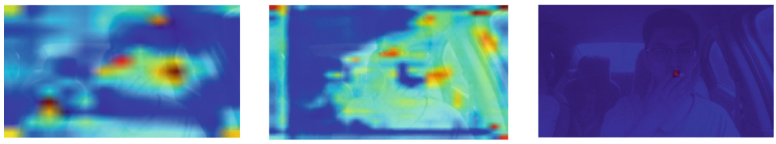


Fig. 4. The visualization results of Grad-CAM: (a) Head layer 1, (b) Head layer 2, (c) Head layer 3.

4 Experimental and Discussion

In order to verify the effectiveness of our improved algorithm, we compare SmokeNet with yolov4. We have adopted three indicators, such as model size, fps and mAP to assessment the detection performance of the model. The model size reflects the demand for computing resources and the large models need more GPU memory and GPU processors. The fps indicates how many pictures the model can process per second, which directly reflects inference speed of the model. mAP is proposed in Pascal VOC dataset, which represents the detection accuracy of the algorithm, and its mathematical definition is shown in the following Eq. (1–3):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

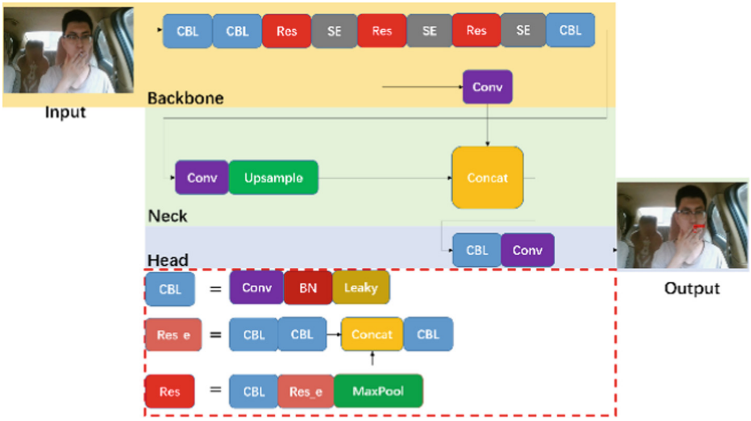
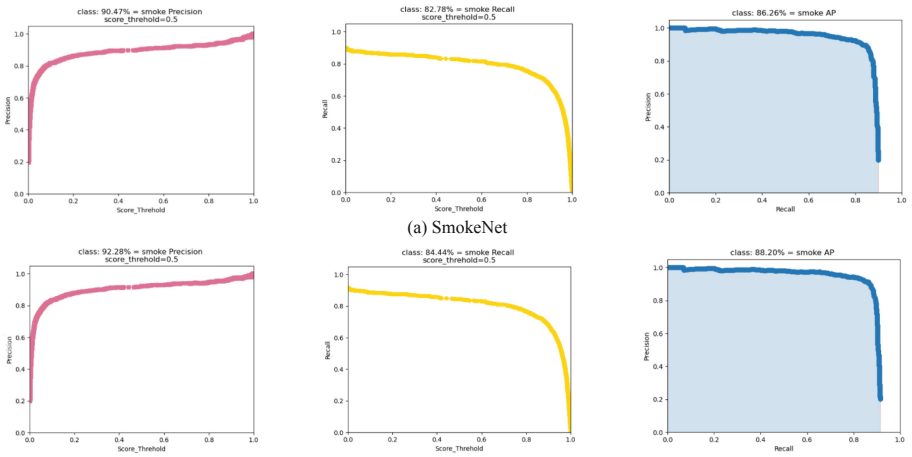


Fig. 5. The structure of improved model.

$$\text{mAP} = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \{AP = \frac{\sum Accuracy}{N} \tag{3}$$

TP represents the numbers of true positive image. FP represents the numbers of false positive image and FN represents the numbers of false negative image. Q is the number of detection categories, Q = 1 was used in this study. The precision, recall and mAP curve of the SmokeNet and yolov4 are shown in Fig. 6. All experiments will be carried out in the embedded platform Jetson Xavier NX. The detailed results are shown in Table 1. As shown in Table 1, the model size of Yolov4 is 244 MB, and the fps is only 14. In contrast, the model size of SmokeNet is only 23 MB, which means less hardware resources are consumed. Although the detection accuracy of SmokeNet mode is lower than yolov4, the detection speed is improved by 57% and the map can reach 86%. In my opinion, we only detect one object, and the target object will not have much scale change. Therefore, the detection model does not need too many layers. For driver smoking detection, it is more important to have a faster detection speed, so the lightweight object detection algorithm is more suitable for practical application.

At the same time, we verify the influence of attention mechanism on accuracy, we added different attention mechanisms to the model and compared it with the model without attention mechanisms. It can be seen from the results in the Table 2, the attention mechanism has little effect on the accuracy of yolov4 model, because yolov4 already has a very deep convolution layer for feature extraction, and our datasets are small and the detection target is single. On the contrary, adding attention mechanism can significantly improve the accuracy of our lightweight model, and the inference speed is almost unchanged.



(a) SmokeNet

(b) Yolov4

Fig. 6. From left to right, the images representing the precision, recall and mAP curve image of the model.

Table 1. Comparison results of smokenet and yolov4.

	Model size	fps	Precision (%)	Recall (%)	mAP
YOLOV4	244 MB	14	92.28	84.44	88.20
SmokeNet	23 MB	22	90.47	82.78	86.26

Table 2. Influence of attention mechanism on model accuracy.

	mAP
Yolov4 without-attention mechanism	87.69
Yolov4-SE	88.20
Yolov4-CBAM	88.12
SmokeNet without-attention mechanism	85.98
SmokeNet-SE	86.26
SmokeNet-CBAM	86.20

5 Conclusions

In this paper, a lightweight high-precision smoking detection framework is proposed and applied to edge embedded devices. In view of the relatively fixed size of cigarette targets in vehicle video, the last multi-scale feature layer in Yolo algorithm is modified to output only one feature layer, which effectively avoids the detection interference caused by small target learning. At the same time, the attention mechanism is added to

the extracted feature layer to increase the richness of network feature information. The experimental results show that the performance of this proposed method is better than the traditional driver smoking detection method.

Although the driver smoking detection method proposed in this paper has been basically completed, the rationality and effectiveness of the model and method have been verified by experimental tests. However, some work and research contents are still worthy of further research and discussion.

Limited by the video dataset taken by ourselves, the dataset used in this experiment has several problems such as insufficient sample coverage. Therefore, in the follow-up work, it is necessary to further increase the types and quantity of samples. At the same time, we will continue to compress model, we are considering model pruning and other methods to further improve the detection speed.

References

1. Bochkovskiy A., Wang C Y., Liao H., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.
2. Cao C., Liu X., Yi Y., et al., 2015. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. IEEE International Conference on Computer Vision.
3. Chien T C., Lin C C., Fan C P., 2020. Deep Learning Based Driver Smoking Behavior Detection for Driving Safety. Journal of Image and Graphics.
4. Echebarria I., Imtiaz S A., Peng M., et al., 2017. Monitoring smoking behaviour using a wearable acoustic sensor. In 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
5. Iwamoto K., Inoue H., Matsubara T., et al., 2010. Cigarette smoke detection from captured image sequences. International Society for Optics and Photonics.
6. Jie H., Li S., Gang S., et al., 2017. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence.
7. Lu M., Hu Y., Lu X., 2020. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Applied Intelligence.
8. Odetallah A D., Agaian S S., 2012. Human visual system-based smoking event detection. Mobile Multimedia/Image Processing, Security.
9. Patil Y., Lopez-Meyer P., Tiffany S., et al., 2013. Detection of cigarette smoke inhalations from respiratory signals using reduced feature set. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.
10. Ren S., He K., Girshick R., et al., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence.