Faster Detection Method of Driver Smoking Based on Decomposed YOLOv5

Fangfei Shi¹, Hui Zhou¹, Chunyang Ye¹ and Jianbin Mai¹

¹ School of Computer and cyberspace security, Hainan University, Haikou Hainan, 570028, China.

sssfangfei@163.com;zhouhui@hainanu.edu.cn;cyye@hainanu.edu.cn;598993434@qq.com

Abstract. Aiming at the problem of large computational resource consumption in the existing object detection network for driver smoking detection, a decomposing YOLOv5 network (Dec-YOLOv5) is proposed for optimization. This method uses singular value decomposition (SVD) to split the standard convolution in the pre-trained YOLOv5 network into two simpler convolution operations to reduce the computational cost. The optimized network generated after decomposition does not need to be retrained, which can reduce the number of parameters and calculations while maintaining the detection accuracy of the pre-trained model. The experimental results show that the detection time of Dec-YOLOv5 is only 80% of the original YOLOv5 when the overall detection accuracy reaches 93.5%. At the same time, compared with the current mainstream object detection model, the Dec-YOLOv5 network has a better ability to express the characteristics of the driver’s cigarettes, and the detection accuracy is higher.

1. Introduction

As people's requirements for travel quality increase, uncivilized driving behaviour has become an important regulatory content. Smoking is one of the most common uncivilized driving behaviours, which not only harms the health of passengers, but also easily causes traffic accidents. Therefore, detecting the smoking behaviour of the driver becomes more and more important.

At present, the mainstream approach of driver's smoking detection mainly uses computer vision technology. According to the different detection targets, it is divided into smoke detection and cigarette detection. The research of JinLan et al. [1] combines the characteristics of the slow change of cigarette smoke and uses the background difference method to extract the smoke area. Shrivastava et al. [2] extracted the colour features of smoke according to the moving target and combined them with the K-means clustering algorithm [3] to improve the detection effect. However, it takes a lot of time to extract these complex image colour and texture features, so it is inefficient in practical applications.

In recent years, convolutional neural networks have achieved good performance in object detection tasks [4-5]. The feature of this kind of method is that it can automatically extract the effective high-dimensional features of the target from the image. Therefore, some scholars have proposed to use convolutional neural networks to detect driver's smoking behaviour.

Object detection based on convolutional neural networks is mainly divided into One-stage and Two-stage. Among them, the One-stage object detection algorithm only uses one neural network for detection, mainly by predicting the bounding boxes and presence probabilities of objects in different regions of the image, and then assigning weights to each bounding box. The advantage of this method is that there is
no need to extract multiple ROI and only one full-image convolution and regression is required. For example, the method proposed by Mao et al. [6] using the YOLOv3 [7] to detect cigarettes around the face is based on the One-stage method. In the Two-stage method, a series of target regions are generated by the region generation network, and then the samples are classified and modified by a convolution neural network. Although the object detection model based on the convolutional network can significantly improve the detection accuracy, the increase of the convolutional layer will increase the number of network parameters, which consumes a lot of time, restricting the development and application of the cigarette detection model on vehicle-mounted equipment with limited computing resources.

On the basis of the above research, we propose a decomposed YOLOv5 (Dec-YOLOv5) network for driver smoking behaviour detection based on the YOLOv5 network and singular value decomposition (SVD). This method decomposes the standard convolution in the pre-trained yolov5 network into a two-step convolution combination of depth-wise convolution and point-wise convolution with less computation and parameters by SVD, so as to effectively reduce the high computational resource consumption in the original network. In addition, the weight of the new two-step convolution in the method is calculated from the weight of the original convolutional layer, so the detection accuracy of the original network can be maintained without retraining.

This paper uses the video frame images of the smoking behaviour of the driver as the sample to conduct experiments on the designed network. Firstly, the Dec-YOLOv5 network proposed in this paper will be compared with the original YOLOv5 network in terms of detection accuracy, time, and model size to verify the optimization effect of Dec-YOLOv5. Secondly, we will compare the detection accuracy of Dec-YOLOv5 with the mainstream object detection networks such as Faster R-CNN [8], YOLOv3, YOLOv4 [9], etc. to illustrate the accuracy advantage of Dec-YOLOv5 network detection.

2. The convolutional network optimization method

To optimize the parameters and calculation of the model, the common method is to adjust the design structure of the model or decompose the model parameters. MobileNet proposed by Howard et al. [10] is adopted the first idea. They proposed to build a faster network through a deep separable convolutional process. On the other hand, Zhang et al. [11] proposed a method to decompose the standard convolution into a combination of pointwise convolution and the standard convolution with a smaller output channel. They use SVD to decompose the weight of the original convolution to obtain the channels and weights of the new convolution to accelerate the convolution network.

2.1. Decomposition convolution structure optimization method

Inspired by Zhang's research, we also use SVD to decompose the convolution weights to optimize the computation and parameters of the network, the structure of our method is shown in Figure 1.

![Figure 1. Cigarette target detection network based on YOLOv5s.](image-url)
It can be seen from Figure 1, in the original deep convolutional network, the output \((X \ast Y \ast n)\) is calculated from the combination of input \((X \ast Y \ast c\) and the standard convolution \((n \ast k_w \ast k_h \ast c)\). In our method, the feature map is first input to the depth-wise layer obtained by decomposition, and then input to the pointwise layer to get the output.

The difference between our method and Howard et al. is that our method is based on the pre-trained network and does not require retraining. And we add the scaling factor \(r\) to adjust the relationship between model compression and accuracy during the decomposition, so that the decomposed structure can maintain accuracy while optimizing the number of parameters and the amount of calculation. The following Equations 1 to 2 are the ratio of the parameter quantity and the calculation quantity of the decomposed convolution and the original standard convolution respectively.

\[
\frac{C_r+k_w+k_h+1+n+1+1+C_r}{n+k_w+k_h+C} = \frac{1}{n} + \frac{1}{k_w \ast k_h} \tag{1}
\]
\[
\frac{X \ast Y \ast +1+k_w+k_h+C_r+X \ast Y \ast C_r+1+1}{X \ast Y \ast C_r+k_w+k_h+n} = \frac{r}{n} + \frac{r}{k_w \ast k_h} \tag{2}
\]

The above Equation 1 shows the comparison of the number of parameters of the decomposed convolution and original standard convolution. Assuming that the output channel of the network is 64, and the kernel is \((3 \ast 3)\), it can be seen from Equation 1 that the decomposed parameters are about \(\frac{1}{9}\) of the original convolutions. The calculation amount in Equation 2 is the same, the calculation amount of the decomposed convolution also becomes about \(\frac{1}{9}\) of the original when the \(r\) is 1. It can be seen that this method can be greatly optimized in terms of parameters and calculations compared with the original standard convolutional layer.

2.2. Decomposition of weight of convolution layer

In our method, the original convolution has been pre-trained, so we try to decomposition the standard convolution into the depth-wise and the pointwise convolution, where the weight of the new convolution is calculated from the original pre-trained convolution. The following Equations 3 to 5 show the mathematical expression of the convolution decomposition process.

\[
\text{Standard}_{(n \ast k_w \ast k_h \ast c)} \Rightarrow \text{Point}_{(n+1 \ast 1 \ast C_r)} \text{Depth}_{(C_r \ast k_w \ast k_h \ast 1)} \tag{3}
\]
\[
M_{(n+m)} = U_{(n-m)} \Sigma_{(n-m)} V^T_{(m-m)} \tag{4}
\]
\[
(A^T A) \psi = \lambda_i \psi \Rightarrow \psi = \sqrt{\lambda_i} \psi \Rightarrow \psi = \frac{1}{\sigma_i} A \psi \tag{5}
\]

The above Equation shows the weight decomposition process of convolution. Among them, Standard in Equation 3 is the weight of pre-training convolution. Point and Depth respectively represent the combination of pointwise and depth-wise convolution cascade after decomposition. In Equation 4, \(M\) is obtained by the transformation of pre-training convolution Standard, \(U\) is a square matrix with left singular value inside obtained by SVD, \(\Sigma\) is a diagonal matrix with 0 values except the singular values on the diagonal, \(V^T\) is the transpose of characteristic matrix \(V\), and its internal elements are all right singular vectors. The singular value is similar to the eigenvalue, the eigenmatrix \(\psi\) obtained from the eigenvalue method of the matrix in Equation 5 is the right singular vector in \(V^T\). And \(u, \sigma\) in Equation 7 are respectively the left singular vectors and singular values of \(U\) and \(\Sigma\) in Equation 4, so it can be known that the singular values in the matrix of \(U\) and \(\Sigma\) dot multiplication and the \(V^T\) matrix are both the eigenvalues of the matrix \(M\).

Just like eigenvalues, singular values are arranged from large to small in the eigenmatrix and decrease very fast, so that the sum of the singular values in the previous part can represent most of the singular values. Therefore, adding the scaling factor \(r\) in the process of weight decomposition can achieve the purpose of using a smaller convolution combination to represent the original convolution.
3. Driver smoking detection method based on decomposing YOLOv5 network

3.1. Cigarette target detection model based on YOLOv5

Compared with the previous detection networks such as YOLOv3 and YOLOv4, the advantage of the YOLOv5 is that it can control the depth and width of the structure through code, so it can generate a series of versions of different sizes such as YOLOv5x, YOLOv5l, YOLOv5m, and YOLOv5s. This also makes the YOLOv5 model have stronger robustness in practical application. Considering the computational consumption and the portability of the model, we choose YOLOv5s with the minimum depth and width to build a cigarette target detection model. Figure 2 shows the structure of the detection network based on YOLOv5s.

![Feature extract network](image)

The core of this model is the convolutional network of the Backbone and Neck modules. The convolutional layer and sub-sampling layer in the network will extract the high-dimensional positive features of the target in the image from different scales after learning. Among them, the convolutional layer can enhance the input data and reduce noise, and the sub-sampling layer can help reduce the amount of data processing. And we will also optimize this part of the model.

3.2. Design of driver smoking fast detection model based on decomposed YOLOv5 network

Combining the YOLOv5 network and the proposed decomposition convolutional structure method, we constructed the Dec-YOLOv5 network as shown in Figure 3. The image sample of the driver smoking is first input into the YOLOv5 as shown in Figure 3a for pre-training. After the pre-training, the SVD method is adopted to decompose the standard convolution of the Backbone in the pre-training network to get the new depth-wise layer and the pointwise layer, and generate the network as shown in Figure 3b.
Figure 3. The method we proposed in this paper
After the optimization network is generated, the test samples will be input into the network in Figure 3b. The optimized convolution layer can extract the high-dimensional positive features of the samples. Then the Prediction network will generate feature graphs of different levels of the targets to be detected according to the extracted features. Finally, the coordinate information and the predicted value of the category of the targets in the image can be obtained by combining the NMS method.

Since the function of the first convolution layer in the Backbone network and the convolution layer in the Prediction network is to sample the input samples and output the multi-level and high-dimensional positive characteristics respectively, they are not considered in the convolution decomposition. In order to prevent the sharp decline of detection accuracy, the decomposition optimization is only carried out in the middle of the Backbone network.

4. Experiment and analysis

4.1. Experimental data and experimental settings

Since there is no public driver's smoking behaviour data set, the data in this paper mainly come from videos of various smoking postures taken from different angles, weather and lighting conditions in real scenes. A total of 120 minutes of smoking behaviour data were included in the videos, among which video was obtained by using (1280 * 720) pixels to shoot 1.2 min of smoking driving of volunteers. A sample of the self-photographed data in this paper is shown in Figure 4.

![Figure 4. A sample of driver video data taken by this paper](image)

In the experiment, the scaling factor $r$ of the Dec-YOLOv5 model is set to 5, the initial learning rate is set to 0.0001, and the maximum iteration times Epochs is set to 1000. The training strategy of degenerated learning rate is adopted to reduce the Epoch by 10 times every 20 iterations. The experiment will divide the data set into three sets: training, calibration, and testing. After the training, the model will first perform calibration optimization on the calibration dataset before testing the performance on the test dataset.

To prove the effectiveness of the method proposed in this paper, different experiments are designed to verify it in this section. The details are as follows: (1) When the network super parameters are set the same, the driver smoking detection methods based on Dec-YOLOv5 and YOLOv5 are compared in terms of detection accuracy, average precision mean, detection time, and model size. (2) The average detection results of different smoking driving videos were compared between Dec-YOLOv5 and other common object detection methods.

4.2. Performance comparison between Dec-YOLOv5 and YOLOv5

In this experiment, Dec-YOLOv5 is compared with the YOLOv5 before optimization in the case of the same network super parameters and experimental data. The Precision and mAP are used to measure the performance of driver smoking detection. In addition, it also uses the memory occupied by the model and the time it takes to detect 100 frames of driver smoking images to compare the model size and the computing resource consumption. Table 1 shows the experimental results of their comparison.
Table 1. The performance comparison between Dec-YOLOv5 and YOLOv5.

| Method      | Detection Time(s) | Model Storage size(MB) | Precision (%) | mAP (%) |
|-------------|-------------------|------------------------|---------------|---------|
| Dec-YOLOv5  | 4.0               | 25.4                   | 93.5          | 87.6    |
| YOLOv5      | 4.9               | 27.0                   | 93.8          | 88.4    |

It can be seen from Table 1, when the YOLOv5 network before optimization was deployed and tested on our devices and datasets, the required memory size was 27.0MB and the detection time was 4.9s. In contrast, the Dec-YOLOv5 network proposed in this paper can optimize the deployment memory and detection time of the model to 90% and 80% of the former when the detection accuracy is only 0.3% less than the former. It can be proved that the Dec-YOLOv5 network can effectively optimize the problem of high computational loss while maintaining the detection accuracy of the pre-trained YOLOv5 network to the maximum extent. At the same time, since the Dec-YOLOv5 method does not need to be retrained, it is more robust for the research on the deployment of high-precision detection models in devices with limited computing resources.

4.3. Performance comparison between Dec-YOLOv5 method and other object detection methods

In order to verify the effectiveness of the Dec-YOLOv5 method, this experiment compares the accuracy of the Dec-YOLOv5 method with other three common object detection models in the detection of different smoking driving videos in the dataset. Each video is manually counted with and without cigarettes for each frame of the video.

Table 2. Comparison of detection performance of different object detection methods.

| Video    | Faster R_CNN | YOLOv3 | YOLOv4 | Dec-YOLOv5 |
|----------|--------------|--------|--------|------------|
| Video 1  | 84.7         | 82.4   | 89.4   | 89.3       |
| Video 2  | 81.4         | 83.6   | 86.7   | 88.1       |
| Video 3  | 85.4         | 87.8   | 90.3   | 92.4       |
| Video 4  | 85.9         | 86.8   | 90.1   | 90.5       |
| Video 5  | 86.6         | 87.3   | 91.0   | 90.4       |
| Video 6  | 85.3         | 87.1   | 93.3   | 92.6       |
| Video 7  | 82.1         | 79.5   | 88.4   | 90.0       |

Table 2 shows the average detection results of Dec-YOLOv5 and some other mainstream object detection methods on 7 different smoking and driving videos in the dataset. As can be seen from the results in the table, compared with other object detection methods, the Dec-YOLOv5 method proposed in this paper has a stronger feature expression ability for driver smoking images, so the detection accuracy of cigarette targets in the images is also higher.

5. Conclusion

Aiming at the problem that the current object detection method based on deep convolutional network has high computational cost, we proposed a Dec-YOLOv5 method to optimize driver smoking detection. The experimental results show that, Dec-YOLOv5 can not only maintain the detection accuracy of the YOLOv5 effectively, but also significantly reduce the calculation cost during detection. And compared with the mainstream object detection methods, our method has stronger ability to express the texture features of the detected image, and the detection accuracy is higher.

6. Reference

[1] L. Jinlan, W. Lin, Z. Ruliang, H. Chengquan and R. Yan, “A Method of Fire and Smoke Detection Based on Surendra Background and Gray Bitmap Plane Algorithm,” 2016 8th International Conference on Information Technology in Medicine and Education (ITME), IEEE, pp. 370-374, 2016.
[2] M. Shrivastava, P. Matlani, “A smoke detection algorithm based on K-means segmentation,” 2016 International Conference on Audio, Language and Image Processing (ICALIP), IEEE, pp. 301-305, 2016.

[3] M. A. Wong, J. A. Hartigan, “Algorithm AS 136: A K-Means Clustering Algorithm,” Journal of the Royal Statistical Society, vol. 28, no. 1, pp. 100-108, 1979.

[4] T. He, Z. Zhang, H. Zhang, “Bag of tricks for image classification with convolutional neural networks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, pp. 558-567, 2019.

[5] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3431-3440, 2015.

[6] P. Mao, K. Zhang, D. Liang, “Driver Distraction Behavior Detection Method Based on Deep Learning,” IOP Conference Series: Materials Science and Engineering, IOP, vol. 782, no. 2, pp. 022012, 2020.

[7] J. Redmon, A. Farhadi, “YOLOv3: An Incremental Improvement,” arXiv preprint, arXiv: 1804.02767, 2018.

[8] S. Ren, K. He, R. Girshick, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, vol. 39, no. 6, pp. 1137-1149, 2016.

[9] A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv preprint, arXiv: 2004.10934, 2020.

[10] A. G. Howard, M. Zhu, B. Chen, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv preprint, arXiv: 1704.04861, 2017.

[11] X. Zhang, J. Zou, K. He, “Accelerating Very Deep Convolutional Networks for Classification and Detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, vol. 38, no. 10, pp. 1943-1955, 2015.

**Acknowledgments**

This work is supported by National Natural Science Foundation of China (No. 61962017, 61662019), Hainan Provincial Key Research and Development Program (No. ZDYF2020018), National Key Research and Development Program (No. 2018YFB2000805).