Cigarette end detection based on EfficientDet

Jianhao Liao¹, Jiayu Zou², Ao Shen³, Jinfu Liu⁴, Xiaofei Du⁵

¹School of Automation, Wuhan University of Technology, Wuhan 430063, China
²School of Information Engineering, Wuhan University of Technology, Wuhan 430063, China
³School of Automation, Wuhan University of Technology, Wuhan 430063, China
⁴School of Automation, Wuhan University of Technology, Wuhan 430063, China
⁵School of Automation, Wuhan University of Technology, Wuhan 430063, China

Corresponding author’s e-mail: 289963@whut.edu.cn

Abstract - In the process of building a harmonious society in China, smoking in public places is difficult to curb. However, with the continuous development of computer vision, object detection algorithm plays a key role in the above video detection scene. However, if the detection process is not real-time, it can not be applied to smoking scenes in public places, giving people timely judgment. After analyzing a variety of target detection algorithms, this paper considers that EfficientDet can achieve high efficiency under a wide range of resource constraints, and can easily carry out multi-scale feature fusion, so that it has high speed and accuracy in the process of target detection. In this paper, Efficientnet is used as the backbone feature extraction network, and a set of fixed scaling coefficients are used to scale the depth, width and resolution of the network for preliminary feature extraction, so as to obtain three effective feature layers. Then, in order to improve the prediction level and enhance the feature extraction, the three effective feature layers are transferred to bifpn with cross-scale connection optimization. The prediction results can be used as the evaluation results of the model. The evaluation results show that EfficientDet has high mAP.

1. Introduction

In recent years, target detection algorithm has been widely used in many fields, such as traffic navigation, traffic monitoring, aerospace industry, industrial detection and so on. With the continuous improvement of the accuracy and speed of the target detection algorithm, the consumption of human resources and materials can also be reduced. It can even detect specific targets in many occasions, such as smoking in public places. Target detection of cigarette end. Because of the large scale size, the application of target detection algorithm in many emerging areas, such as autonomous driving vehicles, robots, and so on, will also be limited. The higher computation cost will limit the development of its algorithm. For example, a new NAS-FPN detector proposed by Golnaz Ghaisi[1] and so on requires more parameters and higher FLOPs to reach the highest accuracy. Although some more EfficientDetector structures have been developed before and after, such as Mask R-CNN proposed by Kaiming he[2], Joseph Redmon[3], and SSD proposed by Yolo, Wei Liu[4], etc., but Mask R-CNN ignores low-level features, which is not conducive to the improvement of accuracy, and the detectors like Yolo and SSD need to sacrifice the accuracy to improve the model efficiency. In addition, many researches have ignored the problem that different resource constraints are needed in different practical applications from mobile devices to data.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
centers. Therefore, this paper uses an Efficient net based network structure, which can change the network structure through the expansion of EfficientDet. This method can detect objects under the condition of relatively wide resource constraints, and the efficiency and accuracy of EfficientDet model are higher than other network models, which will be more important for the development of computer vision to the meaning of.

2. EfficientDet Model building design

2.1 Construction of model architecture

The model structure of the whole Efficientdet first selects Efficientnet as the backbone network, then takes the proposed bifpn as the feature network, extracts the third-level to the seventh level multiscale features \{ C_3, C_4, C_5, C_6, C_7 \} from efficientnet, and repeatedly applies the two-way feature fusion from top to bottom. Then these fused features are input into the class box network, and the object class and boundary box are generated respectively forecast. For a neural network, it is assumed that all layers must be extended uniformly by the same constant proportion. In order to extend the network length (L_i), width (W_i) and resolution (R_i; T_i) without changing the EL predefined in the baseline network during the model scaling, the model precision optimization problem of the model under given resource constraints can be expressed as

\[
\max_{d,w,r} \text{Accuracy} \ U(d, w, r)
\]

\[
s.t. \ U(d,w,r) = \bigcap_{i=1:s} E_i \left( X_{r=R_i;T_i;w=W_i} \right) \quad (1)
\]

Memory \ (U) \leq \text{target\_memory}

FLOPS \ (U) \leq \text{target\_flops}

Where w; d; r are multiples of network width, depth and resolution respectively \hat{E}_i; \hat{L}_i; \hat{W}_i; \hat{R}_i; \hat{T}_i \ are predefined parameters in baseline network. In this paper, the composite coefficient \eta \ is used to scale the width, depth and resolution of the network to achieve higher accuracy and efficiency

Depth: \quad d = \alpha \eta

Width: \quad w = \beta \eta

Resolution: \quad r = \gamma \eta

s.t. \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \quad (2)

\eta \ is a specific coefficient, which can be used to determine the total amount of resources used. The above \alpha, \beta, \gamma \ are constants obtained from a small range of web search, which determine the specific situation of resource allocation.

2.1.1 EfficientNet

The backbone network of Efficientnet is composed of MBConv. The specific steps of compound extension method are as follows:

Step 1: fix \eta \ to 1, search \alpha, \beta, \gamma \ grid through formula (1) and formula (2), and find the proportion with the highest efficiency.
Step 2: Keep α, β, γ unchanged, and the baseline network is expanded with different η by formula (2). And the specific effects of different expansion methods are shown in figure (1).

![Figure 1](image)

Figure 1. (a) baseline network; (b) network expansion by increasing width; (c) network expansion by increasing depth; (d) network expansion by increasing resolution; (e) network expansion by setting hybrid parameters.

### 2.1.2 BiFPN network

![Figure 2](image)

Figure 2. (a) FPN: using simple top-down fusion; (b) ASFF: adopting complex bidirectional fusion; (c) bifpn: searching for effective blocks in FPN and then repeatedly stacking, so as to flexibly control the size of FPN.

2.1.2.1 Comparison of three FPN models

The fundamental purpose of multi-scale feature fusion is to fuse features at different resolution scales. The three FPNs shown in Fig.2 adopt different methods for feature fusion. As shown in Fig. 2 (a), the FPN network mainly solves the multi-scale problem in object detection. The problem is that the traditional FPN network has always been limited to one-way information flow. Therefore, as shown in Fig.2 (b), ASFF adds an additional bottom-up aggregation network to the FPN network, and adopts a more complex two-way fusion technology. In order to further improve the efficiency of the model, this paper adopts a variety of cross-scale connection optimization schemes: since a node has only one edge but does not have feature fusion, it can hardly contribute to feature fusion. Therefore, we first delete all nodes of single input edge, and then add additional edges from the original input when the input to output nodes are at the same level all the time. Edge to the output node to obtain more feature fusion.
technology without additional burden. Each path in the bidirectional path is taken as a layer of feature network layer, and the same layer is repeated many times to achieve higher-level multi-scale feature fusion. Finally, the bifpn structure is obtained, as shown in Fig.2 (c).

2.1.2.2. The principle and method of weighted feature fusion

Nowadays, many models assume that the input features have the same weight, but in fact, when different input features have different resolutions, their contribution to the output is not the same. Therefore, before fusing the features with different resolutions, we often need to change them to the same resolution before superimposing them. In order to solve this problem, this paper provides an additional weight for each input during feature fusion to let the network learn the importance of each input feature, so we decide to adopt fast normalized fusion:

\[ O = \frac{\sum_i W_i}{\sum_i \sigma^{\zeta_i} + \zeta} \cdot I_i \]  

(3)

The Wi in the above formula will use relu to modify the linear element function so as to stabilize the output of Wi greater than 0. There is no problem of gradient disappearance caused by the application of sigmoid. And \( \zeta = 0.0001 \) is a small value set to avoid numerical instability. Like softmax fusion, all weights will be normalized to a probability in the [0,1] interval, but the corresponding fast normalization processing will show high efficiency.

2.1.2.3. Parameter configuration of compound scaling

\( \eta \) is the composite coefficient that controls all other scaling dimensions. Bifpn, box / class prediction network and input image resolution are scaled up according to the following equations (4) (5) (6), respectively, to obtain Efficientdet-D0 (\( \eta = 0 \)) to D3 (\( \eta = 3 \)), as shown in table 1.

| Input Size | Backbone Network | BiFPN #channels | #layers | Box/class #layers |
|------------|------------------|-----------------|---------|------------------|
| D0( \( \eta = 0 \) ) | 512 B0 | 64 | 3 | 3 |
| D1( \( \eta = 1 \) ) | 640 B1 | 88 | 4 | 3 |
| D2( \( \eta = 2 \) ) | 768 B2 | 112 | 5 | 3 |
| D3( \( \eta = 3 \) ) | 896 B3 | 160 | 6 | 4 |

The width and depth of BiFPN are scaled by this formula:

\[ W_{\text{BiFPN}} = (1.35^\eta) \cdot 64 \]  

(4)

2.1.3 Box/class prediction network

The width of the box / class prediction network is equal to that of bifpn, and the depth (number of layers) is increased linearly by the following equation:

\[ D_{\text{box}} = D_{\text{class}} = 3 + \lfloor \eta / 3 \rfloor \]  

(5)

2.1.4 Input image resolution

The resolution of the input image is increased linearly by the following equation:

\[ R_{\text{input}} = 512 + 128 \cdot \eta \]  

(6)

Finally, the overall network architecture of EfficientDet is obtained as shown in Figure 3.
3. The result and discussion of EfficientDet training

In order to verify the model, the data set of smoking is made in advance and divided into training set and test set according to the appropriate proportion of 4 / 5. The training set is used to train the model on GPU, and the test set is used to evaluate the effect of the model. In this paper, the average accuracy is used to evaluate the model, and mAP is selected as the measurement standard. The implementation platform and environment are shown in Table 1, and the important parameters are shown in Table 2.

Table 2. implementation platform and environment

| Processor | Intel (R) Core (TM) i5-8300H CPU @2.30GHz |
|-----------|---------------------------------------------|
| System    | Windows 10 (64 bit)                          |
| Programming Language | Python3.7.4                                  |
| Environment | Anaconda3 (64 bit)                           |
| Framework Construction | Tensorflow2                                  |

Table 3. part parameters

| BATCH_SIZE | 4 |
|------------|---|
| Lr         | 1e-3 |
| Init_Epoch | 0  |
| Freeze_Epoch | 50 |
| Epoch      | 60  |

3.1 The mAP results of EfficientDet

We tested the Efficientdet D0 to D3 after training on the test set, and the results show that the maximum mAP is 44.32% and the running speed is 30% faster on GPU. Figure 4 shows the mAP results of...
EfficientDet-d3 ($\eta = 3$) tested on the test set after training.

| Model          | mAP  |
|----------------|------|
| EfficientDet-D0 | 32.4 |
| EfficientDet-D1 | 38.3 |
| EfficientDet-D2 | 41.1 |
| EfficientDet-D3 | 44.3 |

Table 4. MAP results of EfficientDet (D0-D3)

![Figure 4. the mAP results of Efficientdet on smoking detection.](image)

3.2 smoking test results of Efficientdet applied to smoking

Finally, we apply the model to the cigarette data set we prepared, and decode the prediction picture. Although the prior frame can represent the location information and the size information of the frame, it is limited and cannot represent any situation. Therefore, we need to adjust the EfficientDet, that is, to adjust the position of the upper left corner and the lower right corner of the corresponding prior frame. When the overall prediction framework is adjusted, the final framework can be obtained by ranking scores and non maximum inhibition screening. The predictive effect of EfficientDet in smoking detection is shown in Figure 5.
4. Conclusion
The EfficientDet target detection method used in this paper can detect the target under a wide range of resource constraints, and under this wide resource constraints, compared with some of the current technologies, the efficiency and accuracy of the model are higher. In general, Efficient net's compound scaling needs to increase depth, width and resolution simultaneously when designing larger networks. Compared with previous object detection and semantic segmentation models, it has less parameters and higher success rate. Effective detector can effectively detect cigarette butts of people who have smoking behavior in public places, so as to curb the phenomenon of smoking in public places. The development of EfficientDet is very beneficial to the research of object detection and the development of diversity of applications. It is believed that it will be of great significance to the development of computer vision.

Acknowledgments
The research presented here was supported by "National innovation and entrepreneurship program for college students" (Item Number No. S202010497230).

References
[1] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[2] Ghiasi G, Lin T Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 7036-7045.
[3] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. European conference on computer vision. Springer, Cham, 2016: 21-37.
[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[5] Tan M, Le Q V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks [J]. 2019.