ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.169

Algorithm Research on Vehicle Type Classification Based on Effi-YOLOX

You Zhou [1], Leilei Ma [2]

[1,2] School of Automation, Xi'an University of Post and Telecommunications

Abstract. In order to improve the feature extraction and recognition capability of the model for vehicle images in different environments, this paper proposed a model named Effi-YOLOX to processing the vehicle type classification task. In this paper, the migration parameters based on Imagenet are used to initialize the EfficientNetv2 model, and the backbone of EfficientNetv2 is used as the backbone of Effi-YOLOX. Furthermore, MHSA(Multi-head Self-Attention) is used to replace SE-attention so that the model can extract more comprehensive features. On this basis, PANet+ is used to replace the PANet of the original YOLOX model; The Coordinate Attention(CoordAttenton) is introduced to make the model focus more on important features. The experimental results show that the Effi-YOLOX is superior to the existing classical models in the vehicle type classification task.

Keywords: Vehicle Type Classification, Convolutional Neural Network, Transformer, Attention Mechanism

1. Introduction

Since the beginning of the 21st century, the economy of China has entered a stage of rapid development and people's living standard has been greatly improved. Along with the infrastructure development, the number of vehicles on the road has increased dramatically, causing traffic jams and more traffic accidents, leading to the increasing pressure on urban traffic management. Therefore, Proposing an accurate vehicle type recognition algorithm have gradually become a research hotspot in the field of intelligent transportation. However, in the current traffic monitoring conditions, problems such as different weather conditions, similar vehicle types and complex backgrounds pose considerable difficulties for the vehicle type classification task, which therefore requires models that can extract more useful features from images.

At present, there are three main methods of vehicle type classification, the first one is classification by hardware, such as the vehicle type recognition based on information fusion in multiple sensor networks proposed in the literature[1], but the actual deployment process of this method is complicated and the maintenance steps are cumbersome; the second one is to designed image features manually, such as HOG features[2], and then use a classifier to complete the classification task, but this method can only extract shallow features of the vehicle and information, and is susceptible to the road environment and has poor robustness. Another is the classification method based on deep learning, which has been widely used in image classification tasks in recent years. It extracts the depth features of images by feeding a large number of images into a convolutional neural network for training, and the classification performance it achieves far exceeds that of traditional classification methods. For example, Jie Zhang et al[3] combined support vector machine(SVM) and deep convolutional network to design a vehicle type classifier for complex backgrounds, Yongjie Ma et al[4] combined SVM with traditional convolutional neural network AlexNet to propose a new vehicle type recognition method, which has improved speed and accuracy compared with the traditional model, and literature[5] designed improved YOLO as YOLO-vocRV for classification detection recognition of vehicle models and achieved relatively high accuracy.

2. Vehicle Type Recognition Algorithm

2.1. Overall Structure of Effi-YOLO

In order to improve the accuracy of vehicle type classification in the field of deep learning, this paper proposes a model named Effi-YOLOX based on the improved YOLOX. The model structure is shown in Figure 1, and the modules of the Effi-YOLOX are introduced as follows.

Backbone: In this paper, the backbone of lightweight network EfficientNetv2 is used to replace Darknet53 of YOLOX, which reduces the complexity of the model. The MHSA in Transformer is used to replace SE-attention, so that the network can not only focus on local information, but also global information.

Neck: Replacing the PANet[6] of original model with PANet+, which further enhances the utilization of features by the model; introducing CA(CoordAttention)[7] attention mechanism to the neck of the model, with a smaller computational overhead enabling the network to filter the features that the model should pay more attention to.

Head: YOLO series' backbones and feature pyramids continuously evolving, their detection heads remain coupled, while in YOLOX. Replacing YOLO series' head with a decoupled[8] one which improve the performance of YOLO..

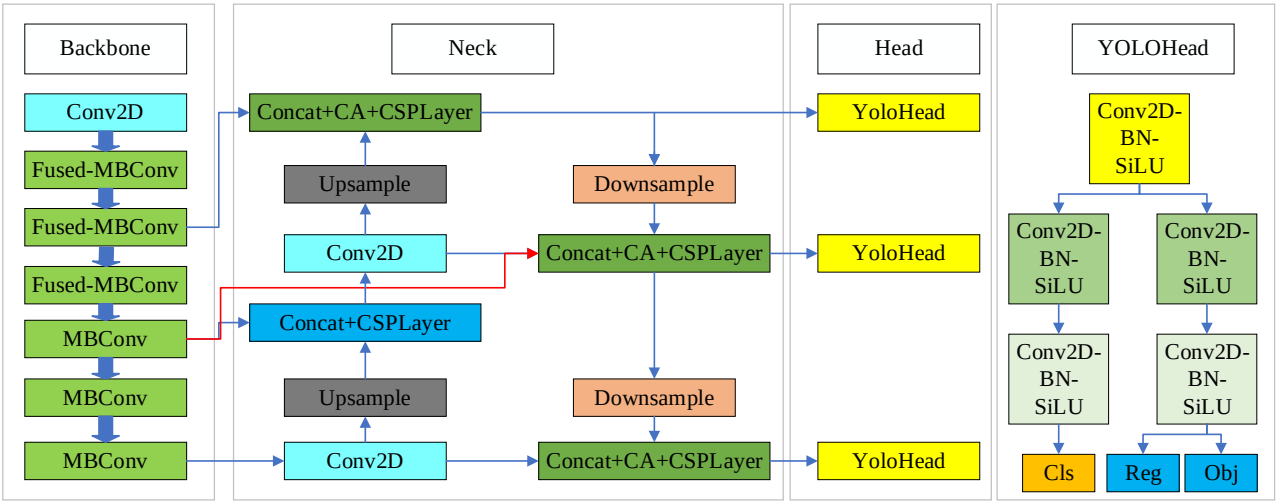

Fig. 1: Overall structure of Effi-YOLOX

2.2. Backbone

EfficientNetv2 achieves a compromise between model complexity and accuracy by reducing the width of channel and expanding the depth of network. The backbone of EfficientNetv2 is mainly consists of the Fused-MBConv and MBConv, and their structures are shown in Figure 2 and Figure 3.

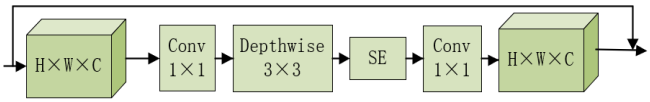

Fig. 2: Structure of MBConv

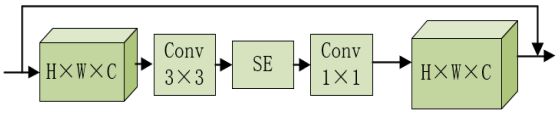

Fig. 3: Structure of Fused-MBConv

Firstly, the MBConv module up-dimensions the feature map by 1×1 conv (the convolution kernel size is 1×1)layer, and then performs deep separable convolution on the result obtained[9], with that, adds the SE(squeeze and excitation) attention mechanism, and finally reduces the dimension of the feature map by 1×1 conv layer. The depth-wise separable convolution has less number of parameters compared to traditional convolution, and the drop ratio is shown in (1).

$$r = \frac{c_{in} \times k \times k + c_{in} \times c_{out}}{c_{in} \times c_{out} \times k \times k} = \frac{1}{c_{out}} + \frac{1}{k \times k} \tag{1}$$

However, depthwise separable convolution needs to save more intermediate variables than ordinary convolution, and a lot of time is spent on reading and writing data, which leads to a slower training speed, so EfficientNetv2 uses Fused-MBConv in the shallow layers, and uses MBConv after the fifth stage, so as to achieve a balance between training speed and number of parameters.

In order to prompt the performance of the network for processing high-dimensional global features, the Multi-Head Self-Attention (MHSA) layer is used in the MBConv to replace the SE-attention, while the heads is 4. In order to make the attention operation position aware, Transformer based architectures typically make use of a position encoding On this basis, the model can be attributed to attention not only taking into account the content information but also relative distances between features at different locations[10]. The , , are obtained by linear transformation of input X. The and in the Fig.4 represent the height and width of the relative position encoding, respectively. The q, k, v, r represents query, key, value and position encodings respectively. The A and M represent element wise sum and matrix multiplication respectively. For the sake of reducing the amount of computation, MHSA is only used in stage7 which the resolution is 10×10.

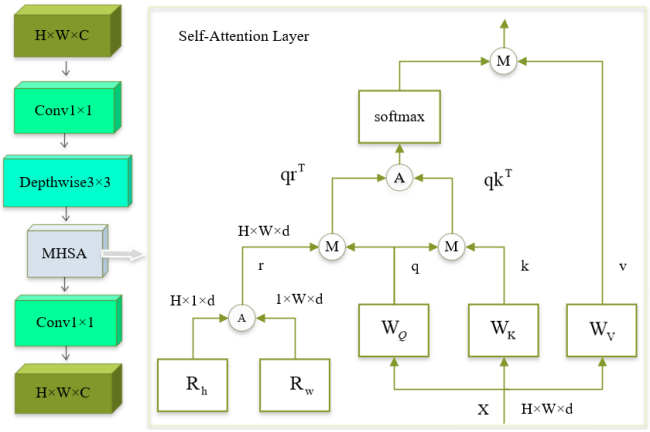

Fig.4: Structure of MBConv-T

2.3. Neck

The Neck of Effi-YOLOX is used to achieve the aggregation of semantic features at three scales, which makes it possible for the model to recognize the targets of different sizes. The three-scale feature maps(40×40×64 , 20×20×160, 10×10×256) extracted from stage4, stage6, and stage7 of the EfficientNetv2 backbone are selected, input them into the PANet+ module for feature aggregation. As shown in Figure 5, PANet adds bottom-up enhancement on the basis FPN (Feature Parymid Network)[11], which is conducive to improving the accuracy of vehicle detection. The introduction of down sampling in PANet leads to information loss of output features, PANet+ follows the idea of BiFPN[12] on the basis of PANet, and introduces mid-scale feature bridging, through the tensor stitching, it can realize feature compensation at the cost of a small amount of complexity, which is more conducive to improving the detection accuracy of vehicles in real scenes.

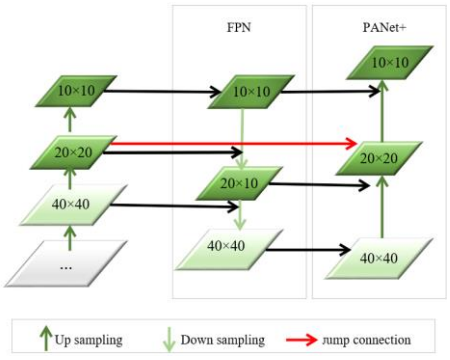

Fig.5: Structure of PANet+

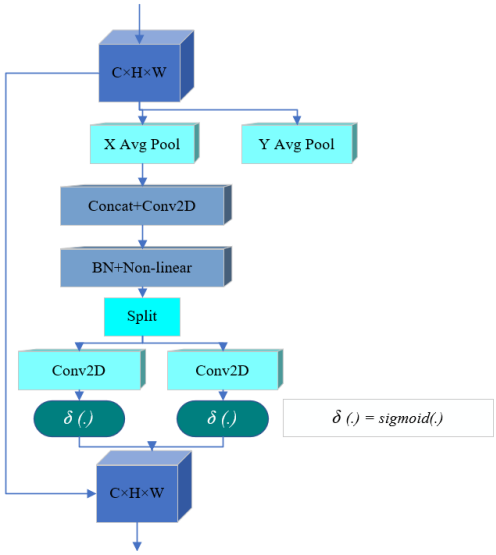

Fig.6: Structure of CA

The CA attention mechanism is introduced to the three feature maps (40×40×192, 20×20×416, 10×10×512) after the feature is superimposed, and the structure of CA attention is shown in Figure 6. By embedding the location information into the channel attention, it not only obtains cross-channel information, but also direction perception and location perception information, which can help the model to locate and identify the target of interest more accurately while avoiding incurring a large amount of computational overhead. To alleviate the loss of location information caused by globalaveragepooling2D, the channel attention is decomposed into two parallel 1D feature coding processes; the obtained feature maps are superimposed, and then 1×1 conv layer is used to generate intermediate feature maps with spatial information in both vertical and horizontal directions; then, the intermediate feature maps are divided into two feature maps along the spatial direction, on this basis, using 1×1 conv layer to convert the number of channels. Finally, the weights of the two directions are multiplied with the input feature map, so that the location information is stored in the feature map.

2.4. Head

YOLO series' backbones and feature pyramids continuously evolving, their detection heads remain coupled, while in YOLOX, replacing YOLO's head with a decoupled one as shown in Figure 7, it contains a 1 × 1 conv layer to reduce the channel dimension, followed by two parallel branches with two 3 × 3 conv layers respectively.

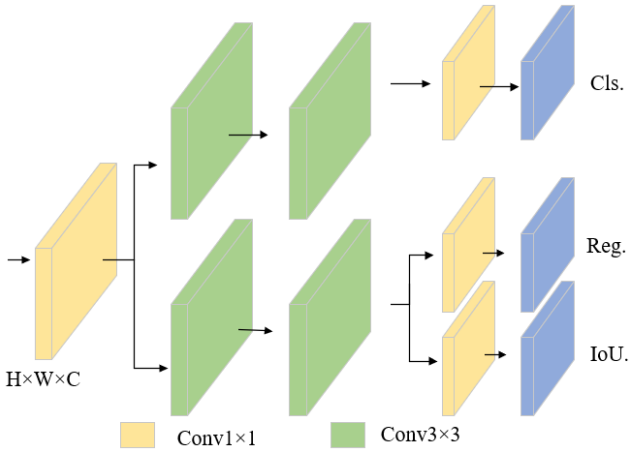

Fig.7. Structure of Head

The Reg.(H×W×4) is used to process the regression parameters of each feature point; Obj.(H×W×1) is used to determine whether each feature point contains an object; and the Cls.(H×W×6) is used to determine the kind of objects contained in each feature point. The final dimension of each feature layer is H×W×11.

3. Experimental Results and Analysis

In order to verify the effectiveness of this model for vehicle type recognition in actual scenarios, this paper compares this model with YOLOX and YOLOv4 using relevant evaluation metrics.

3.1. Experimental Environment and Parameter Setting

The CPU of this experimental platform is Intel Xeon Gold 5218, the memory is 32GB, the graphics card is GeForce RTX 2080Ti with 11G video memory, in addition, there is a Raspberry Pi 4B with 8GB RAM for further test the performance of models. The software environment is ubuntu18.04, cuda10.1, Tensorflow2.2.0 deep learning framework.

The dataset is the BIT-Vehicle dataset released by Beijing Polytechnic University, containing 9850 vehicle images, captured by two cameras at different times and places, with image sizes of 1600*1200 and 1920*1080, respectively. These images contain changes in lighting conditions, proportions, vehicle surface color and viewpoint. Due to shooting delays and vehicle size, the top or bottom of some vehicles are not included in the images. All vehicles in the dataset are divided into six categories: buses, microbuses, minivans, sedans, SUVs, and trucks. The number of vehicles in each category is 508, 883, 476, 5922, 1392 and 822.

3.2. Comparison of Experimental Results

This study compares the vehicle type classification results of Effi-YOLOX, YOLOX, and YOLOv4 from the three aspects of Recall, mAP, and FPS. The results are shown in Table 1.

Table 1: Vehicle Type Classification Results

Model	Recall(%)	mAP(%)	FPS/FPS(Raspberry Pi)
YOLOv4	79.44	93.87	54/1.1
YOLOX	91.26	96.14	73/2.3
Effi-YOLOX	96.59	98.32	69/4.8

The Table I show that the Effi-YOLOX can reach 96.59% Recall in vehicle type recognition tasks, which is 17.15% higher than YOLOv4 and 5.33% higher than the original YOLOX. Meanwhile, the mAP reached 98.32%, which is 4.45% and 2.18% higher than YOLOv4 and YOLOX, respectively. In addition, the FPS of Effi-YOLOX is slightly lower than YOLOX, but when they are deployed on a mobile platform that lacks computing power such as Raspberry Pi 4B, Effi-YOLOX is significantly better than YOLOX due to the smaller model width.

4. Conclusion

In order to improve the performance of convolutional neural network in the field of vehicle type recognition, this paper proposes an vehicle type classification model named Effi-YOLOX based on YOLOX. The experimental results on BIT-Vehicle dataset show that the classification performance of the Effi-YOLOX is better than the classical model.

5. References

[1] Fuliang Li,Zhihan Lv. Reliable vehicle type recognition based on information fusion in multiple sensor networks[J]. Computer Networks,2017,117.

[2] DalalN, Triggs B. Histograms of oriented gradients for human detection. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.San Diego,CA,USA.2005.886–893.

[3] Zhang Jie ,Zhao Hongdong, Li Yuhai,Yan Miao, Zhao Zetong. Classifier for Recognition of Fine-Grained Vehicle Models under Complex Background[J]. Laser&Optoelectronics Progress,2019,56(04):166-173.

[4] Ma Yongjie, Ma Yunting, Chen Jiahui. Vehicle Recognition Baesd on Multi-Layer Features of Convolutional Neural Network and Support Vector Machine [J]. Laser&Optoelectronics Progress,2019,56(14):55-61.

[5] Liu Yao.The Research on Multi-target Detection and Recognition Method Based on Impproved Darknet Framework[D]. Xi'an Polytechnic University,2019.DOI:10.27390/d.cnki.gxbfc.2019.000446.

[6] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In

CVPR, 2018. 2,5.

[7] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [EB/OL]. (2018-3-4) [2021-07-29]. https://arxiv.org/abs/2103.02907.

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In ICCV, 2019.

[9] Tianwen Zhang et al. Depthwise Separable Convolution Neural Network for High-Speed SAR Ship Detection[J]. Remote Sensing, 2019, 11(21) : 2483-2483.

[10] Srinivas A , Lin T Y , Parmar N , et al. Bottleneck Transformers for Visual Recognition[J]. 2021.

[11] Sun Haiyan. An Urban Road Recognition Method that Combines Low-level Features and High-level Semantic Knowledge[D]. North China University of Technology,2018.

[12] Xuqiang Yin et al. Using an EfficientNet-LSTM for the recognition of single Cow's motion behaviours in a complicated environment[J]. Computers and Electronics in Agriculture, 2020, 177.

[13] Li Jiaping et al. A Transfer Learning Method for Meteorological Visibility Estimation Based on Feature Fusion Method[J]. Applied Sciences, 2021, 11(3) : 997-997.

[14] Zhi Tian,Chunhua Shen,Hao Chen,Tong He. FCOS: Fully Convolutional One-Stage Object Detection.[J]. CoRR,2019,abs/1904.01355:

[15] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In CVPR, 2021.