ISSN: 2664-9640

Empirical evidence on efficientnet object detection

Kaige Chen

School of mechanical and Transportation Engineering, Taiyuan University of Technology, Shanxi,030002, China;

chenkaige1232021@163.com

Abstract

Target detection has always been a hot topic in the field of computer vision. Its generation, development, and continuous update iterations and optimizations have prompted technical innovations again and again. Through the development and introduction of various target detection algorithms, this paper summarizes the development and introduction of various target detection algorithms. Since 2006 In 2009, Hinton proposed to use neural networks for autonomous learning, to the later single- and double-stage algorithm of the system, and then to the new convolutional neural network EfficientNet by applying compound scaling technology, as well as the context of the optimized EfficientDet network. , and finally verified the good effect of EfficientDet by applying the framework EfficientDet-Gs under EfficientDet to the BDD100K test set.

Keywords

Target detection, EfficientNet, EfficientDet , BDD100K.

1. Introduction

The so-called target detection is to extract the positioning target from the stored images and videos by analyzing the geometric characteristics of the recognized object, and accurately determine the collective category of the selected target. Objects also have different shapes and degrees of conspicuousness, color changes, and occlusion by other objects. Therefore, more accurate and efficient target recognition is widely concerned.

In order to allow computers to efficiently obtain image information, since Hinton proposed to use neural networks to automatically learn high-level features in multimedia data in 2006, target detection based on deep learning has become an important research hotspot in the field of computer vision. Large-scale image databases such as COCO, Pascal VOC, Labelme, etc., on the other hand, build LeNet, AlexNet, VggNet and other networks to obtain a richer image database, improve the performance of convolutional neural networks, and promote the accuracy of multimedia target recognition. effectiveness.

In 2014, Girshick et al. used the region-based convolutional neural network for the first time to use deep learning in target detection, which improved the accuracy and achieved the overall optimization of performance and efficiency. Pyramid pooling network and Fast R-CNN algorithm, and later Faster R-CNN and R-FCN algorithms are also two-stage target detection algorithms.

Algorithms based on two-stage need to perform replacement training to obtain the shared convolution parameters of the region proposal network and detection network so far, so they all need to consume time. Redmon et al. proposed a framework called YOLOV1, which utilizes the entire Feature maps to predict the confidence of multiple categories and computational frameworks, a one-step framework based on global regression, which directly maps pixels to bounding box coordinates and classification frequencies, reducing time overhead. Later,

ISSN: 2664-9640

because YOLO was difficult to deal with combined objects, Liu et al. proposed the SSD algorithm, which achieved three times better accuracy than Faster R-CNN on PASCAL VOC and COCO. However, there is still a problem that the size and shape of the preselected box cannot be directly obtained through learning, and need to be set manually. There are few low-level feature convolution layers, and there is a problem of insufficient feature extraction.

Efficientnet proposes a composite scaling jointly determined by the composite coefficients in the three dimensions of width, depth and resolution, and by designing a standardized convolutional network expansion method, it can not only improve the recognition accuracy, but also fully save resources. In the same search method of MnasNet, the most original EfficientNet-B0 was searched, and through continuous optimization algorithms, EfficientNetB0-B7 was developed to the recent EfficientNetV2, and its processing speed and operating efficiency have been greatly improved. Now based on the direction of trick fusion and improvement, the traditional efficientnet neural network has produced a new star - EfficientDet. The improved model achieves more advanced accuracy with fewer parameters. Not only can it be easily deployed on the vehicle platform and realize fast reasoning, but also further improve the detection accuracy of small targets. At the same time, as a target detection algorithm with high model efficiency and support for compound scaling, it always achieves better efficiency than existing technologies under a wide range of resource constraints, and has great development potential in the field of autonomous driving detection.

2. Neural network-EfficientNet

The efficientnet launched by Google in 2019 can be said to be the fastest convolutional neural network in the field of computer vision today. From the original VGG16 to today's Xception, in addition to improving the existing neural network in the number of stacked layers, it also There are three points that are quite important for the improvement of an algorithm. First, the proposed network must be trainable and ensure sufficient convergence. Second, the scale of the input parameters should be appropriately small, which is convenient for later training and speed improvement. The third point is to innovate the structure of the neural network and learn more important things. Efficientnet is based on these four aspects, using less parameter training to get better recognition.

2.1.1. Features of the model

Referring to other networks, the existing Efficientnet has outstanding features, which are reflected in the following three points:

Use residual neural network to increase the depth of neural network, through deeper neural network

Change the number of feature layers extracted by each layer, realize feature extraction of more layers, get more features, and increase the width

By increasing the resolution of the input image, the network can learn and express things more abundantly, which is conducive to improving the accuracy.

2.1.2. The development history of EfficientNet

Efficientnet uses MBCConv in MoblieNet V2 as the backbone network of the model, and also applies the squeeze and excitation method in SENet to optimize the network structure.

Table 1: Three Scheme comparing

model	input size	Width factor	depth factor	convergence speed
EfficientNetB0	224x224	1.0	1.0	0.2
EfficientNetB1	240x240	1.0	1.1	0.2
EfficientNetB2	260x260	1.1	1.2	0.3

ISSN: 2664-9640

EfficientNetB3	300x300	1.2	1.4	0.3
EfficientNetB4	380x380	1.4	1.8	0.4
EfficientNetB5	456x456	1.6	2.2	0.4
EfficientNetB6	528x528	1.8	2.6	0.5
EfficientNetB7	600x600	2.0	3.1	0.5

Among the various networks in the history of ImageNet, Efficientnet can be said to have achieved crushing in effect:

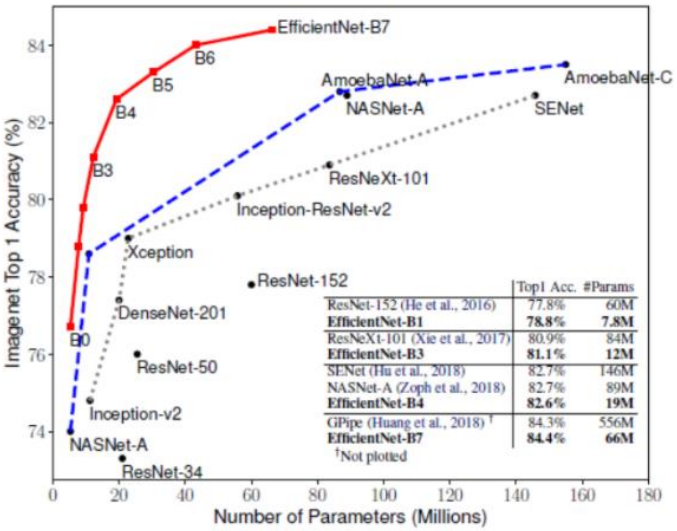


Figure 1: Model Size vs. ImageNet Accuracy

However, larger networks with larger widths, depths, or resolutions tend to achieve higher accuracy, but the accuracy gains quickly saturate after reaching 80%, demonstrating the limitations of only dilating in a single dimension. The dimensions of model dilation are not completely independent. For example, for larger resolution images, deeper and wider networks should be used, which means that each dilation dimension needs to be balanced instead of dilation in a single dimension.

2.2. Application areas and improvements of Efficientnet

At present, efficientnet has been involved in many fields that require accurate image recognition, supervised learning, and deep learning, covering many fields such as unmanned driving, biomedical science, climate recognition, material flaw detection, image classification, etc.

In driverless road condition recognition [5] , the BDD100K dataset, currently the largest and most diverse autonomous driving dataset, is used, covering about 100,000 pictures, various scene types and three different types of scenes in one day. Time-of-day image annotations, which are also applied in different scenarios in extreme weather. The application set covers 10 categories of items such as cars, buses, pedestrians, bicycles, trucks, motorcycles, cyclists, and traffic lights. After removing the corresponding categories, training, verification, and testing are performed at a set ratio Nice effect.

In the field of biological sciences [6] , the efficientnet network has a good application effect for lesion classification and endoscopic observation. For the classification and identification of pathological lesions, the selection of lesion images from photos taken by wireless capsule endoscopy has an effective diagnostic rate. In multi-lesion classification, after selecting a suitable dataset, the input convolution layer is roughly extracted and the batch normalization layer is introduced, and the network is quickly converged by rescaling. At the same time, the residual structure is introduced for reuse to improve the stability of the model.

ISSN: 2664-9640

In the field of medical science [7] , the attention mechanism SE module is improved into an ECA module, which is used as a convolution in the excitation operation to avoid the side effects of dimensionality reduction. The phenomenon of overfitting is eliminated, and the accuracy has reached 98.93%.

other applications such as weather recognition [8] , the average accuracy of efficientnetB7 is also 5% higher than ResNet, 39% higher than Darknet, and 11% higher than VGG16 network on average

It has a very broad application prospect.

3. Improvement of EfficientNet network

The overall framework of Efficientdet is shown in Figure 1. Efficientnet, as its backbone network, is a series of excellent frameworks for fast and high accuracy (Effi-cientnet b0-b7). Compounding Scaling is the core of the entire Efficientnet series. By defining the scaling parameter ϕ , Efficientnet unifies the depth d, width w, and resolution r under the parameter ϕ , and is controlled by ϕ to achieve the three purpose of dynamic adjustment. Due to this standardized convolutional neural network expansion method, the Efficientnet series can not only achieve higher accuracy through model expansion, but also save certain computing resources through model compression. However, considering the actual conditions of the current autonomous driving scene, there is a problem of software and hardware cost performance only using compound scaling. In order to complete most of the feature extraction tasks in complex traffic environments, the optimization of Efficientnet's own network structure is still particularly critical. The network architecture of Efficientnet-b0 is shown in Table 1. Stage1 is an ordinary convolutional layer with a convolution kernel size of 3×3, which contains Batch Normalization (BN) and Swish activation function. Stage2-Stage8 are repeating the stacking of MBConv structures. Stage9 consists of a normal 1×1 convolutional layer, an average pooling layer and a fully connected layer. As shown in Figure 3(a), MBConv [15] is the core part of Efficientnet. The overall design idea is to invert the residual structure (Inverted Residuals), which is used before the 3×3 or 5×5 depth separable convolution structure. The 1×1 convolution increases the dimension, and SE net (Squeeze-and-Excitation Networks) is added after the depthwise separable convolution, and finally a residual edge is added after the 1×1 convolution is used to reduce the dimension.

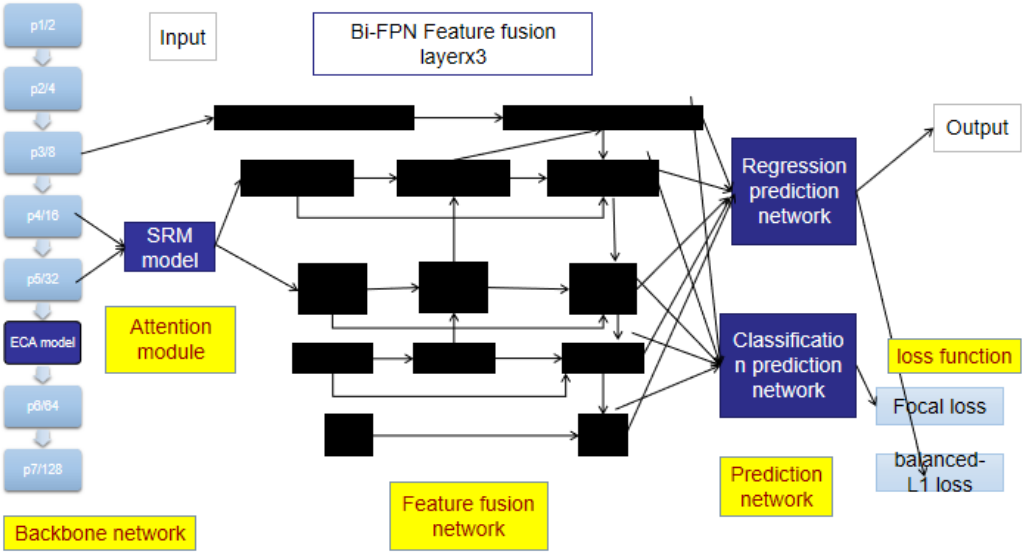


Figure 2: Network structure of EfficientDet

ISSN: 2664-9640

4. Experiments

In the autonomous driving scenario, the research of neural network tends to be applied on mobile devices. The large number of 1×1 convolutions in MBConv still have a high amount of computation, and SE net has proved to be poorly supported on mobile devices.

4.1. Efficientdet improves Efficientnet

The above are all obstacles to the application of Efficientnet to the vehicle platform. In order to overcome the resource limitation in the application, so that the deep neural network can be better deployed on the in-vehicle computing platform, and at the same time provide practical feasibility for the model expansion, this paper uses the Ghost module, ECA net and Channel Shuffle to implement MBConv Refactor .

replaces the 1×1 convolution that takes up a lot of memory and FLOPs in the original MBConv to meet the limited memory and computing resources of the on-board computing platform. Subsequently, ECA net was used instead of SE net, which not only improved the accuracy, but also further reduced the complexity of the model, so as to facilitate the deployment on the mobile terminal. However, the Ghost module may cause the information in the ordinary convolution and the Ghost map to not be exchanged, thus weakening the performance of the model to a certain extent. Therefore, in order to overcome the small amount of side effects, this paper adds channel shuffling (Channel Shuffle) after the up-dimensional convolution to help the information flow between the feature channels. Finally, this improved backbone network is called the Efficientnet-Gs series , which aims to reduce computing and storage costs while maintaining efficient feature extraction capabilities and compound scaling features to achieve the goal of applications in autonomous driving scenarios.

4.2. Effect comparison

The experiments are based on the BDD100K dataset. BDD100K (A Large-scale Diverse Driving Video Database) is the largest and most diverse autonomous driving dataset released by Berkeley AI Lab (BAIR). The BDD100K dataset contains 100,000 high-definition videos. Key frames are sampled at the 10th second of each video to obtain 100,000 images and annotate them.

The batch size of the training set is set to 8, and the iteration number is set to 1,600,000.

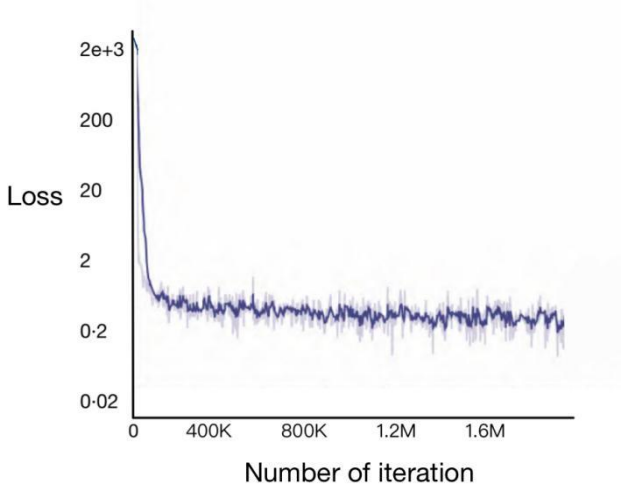
As can be seen from the figure, in the first 5000 iterations of network training, the value of the loss function drops rapidly, and then enters a stable convergence stage. In the training process of the BDD100K dataset, the method of dynamically decreasing the learning rate based on some measured values during the training process is adopted, which enables the network to learn more information from the dataset and use the best obtained during the training process. The optimal weights are used as the final weights file . After combining the reconstructed MBConv , the model size of Efficientdet-Gs-d0 is only 10.1M, which is 36% lower than the original network, which greatly reduces the resource consumption of the device and improves the cost performance of the model. Its FLOPs is 1.9B, which is 25% lower than the original network computation.

Table 2: Performance evaluation on the BDD100K dataset

BDD100K	Car	Bus	Truck	Person	Traffic light	Bicycle	mAP	Model size	FLOPs
Efficientdet-d0	0.7270	0.6003	0.6190	0.5876	0.5364	0.4712	0.5942	15.7m	2.54B

ISSN: 2664-9640

Efficientdetd0 +SRM/ECA	0.7520	0.6015	0.6230	0.6326	0.5883	0.4706	0.6235	15.8	2.56
Efficientdet-Gs-d0	0.7536	0.6007	0.6153	0.6314	0.5792	0.4813	0.6227	10.1	1.90



References

[1] Xu Degang, Wang Lu, Li Fan. A review of typical target detection algorithms for deep learning [J]. Computer Engineering and Applications, 2021, 57(08): 10-25.
[2] Liu Zhipeng. Research on target detection algorithm based on deep learning [D]. Jiangnan University, 2021. DOI: 10.27169/d.cnki.gwqgu.2021.001981.
[3] Li Yinan. A review of deep learning target detection methods [J]. China New Communications, 2021, 23(09): 159-160.
[4] Yang Wei, Du Xuefeng, Zhang Yong, Gao Yue. A review of vehicle target detection algorithms based on deep learning [J]. Automotive Practical Technology, 2022, 47(02): 24-26. DOI: 10.16638/j.cnki .1671-7988.2022.002.006.
[5] Qiu Tianhao, Chen Shurong. Pedestrian re-identification based on EfficientNet-based bi-channel multi-scale joint learning [J/OL]. Computer Applications: 1-8 [2022-02-21].
[6] Wang Zhenya, Zhao Jihong, Wang Yanpeng, Ge Guangying, Sun Qun. Research on detection and classification of pills based on improved EfficientNet network [J]. Modern Computer, 2021, 27(28): 27-32.
[7] Zhang Jiaying. Identification and classification of skin cancer based on EfficientNet [J]. Modern Information Technology, 2021, 5(09): 13-15. DOI: 10.19850/j.cnki.2096-4706.2021.09.004.
[8] HE K, ZHANG X, REN S, at el. Deep residual learning for image recognition [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778
[9] Li Yanchen, Zhang Xiaojun, Zhang Minglu, Shen Liangyi. Object detection in autonomous driving scenes based on improved Efficientdet [J/OL]. Computer Engineering and Applications: 1-11 [2022-02-21].
[10] Chen Xijiang, Anqing, Banya. Optimizing Vehicle Detection for EfficientDet Deep Learning [J]. Journal of Nanjing University of Information Technology (Natural Science Edition), 2021, 13(06): 653-660. DOI: 10.13878/j. cnki.jnuist.2021.06.003.
[11] Wang Huiyong, Yin Ming. Research on efficient target detection algorithm based on EfficientDet [J]. Journal of Hefei University of Technology (Natural Science Edition), 2021, 44(07): 900-908.
[12] Ouyang Pengxiang. Research on efficient target detection algorithm based on improved SSD [D]. Hefei University of Technology, 2021. DOI: 10.27101/d.cnki.ghfgu.2021.001100.

ISSN: 2664-9640

[13] Bao Zhuangzhuang, Zhao Xuejun. Ship detector based on EfficientDet without pre-training SAR images [J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(08): 1664-1672. DOI: 10.13700/j.bh. 1001-5965.2020.0255.

[14] Zhang Yang, Yao Dengfeng, Jiang Minghu, Li Fanshu. Recognition of fine-grained smoking behavior based on EfficientDet network [J/OL]. Computer Engineering: 1-11 [2022-02-21]. DOI: 10.19678/j.issn. 1000-3428.0060760.

[15] WANGH, YU Y, CAI Y, et al. A comparative study of-state-of-the-art deep-learning algorithms for vehicle detection [J] IEEE Intelligent Transportion Systems Magazine, 2019, 11