Proceedings of CONF-MLA 2024 Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/111/2024CH0103

End-to-End Roadway Disease Recognition Based on Transformer Architecture

Zhouheng Deng

School of Automation, Central South University, Changsha, China

jthswd0923@gmail.com

Abstract. Road disease is a significant factor causing traffic accidents. Timely detection and recognition of road disease is significant for maintaining road safety and reducing traffic accidents. Therefore, it is urgent to study an automatic pavement disease recognition method with accuracy and real-time. However, existing real-time target detectors generally use a convolutional neural network-based architecture, and these detectors usually require post-processing of NMS, which makes the detectors difficult to optimize and unstable, resulting in a delay in reasoning speed. Therefore, we design a target detection model based on Transformer, which uses MobileNet as the backbone network to simplify the network structure and enables interaction and integration of features through an efficient hybrid encoder, which reduces the computational load and does not reduce the detection accuracy. Iou-aware query selection optimizes the object query vector in the Transformer structure and reduces the number of candidate boxes that the model needs to process. The experimental results show that we have achieved good results on the RDD2022 dataset.

Keywords: Transformer, Pavement damage detection, Target detection.

1. Introduction

With the construction of urban roads and the increase in the number of automobiles, road traffic safety has become a very urgent problem. According to the previous research, more than 12,000 people die in road accidents globally each year, and road conditions are an important factor in road accidents[1]. In the past, pavement distresses were usually found by manual inspection, which is both time-consuming and costly. Therefore, a low-cost and automatic road disease detection method is urgently needed.

At present, the mainstream target detection algorithms include R-CNN series, YOLO series, SSD, etc., which have achieved good results in terms of target detection accuracy based on different structures and principles. However, when performing targeted detection of roadway damage, it is necessary to detect and identify road diseases through video surveillance by vehicle or road maintenance equipment, and real-time target detection is very important. Existing real-time target detectors generally use a convolutional neural network-based architecture, and these detectors usually require post-processing of NMS (non-maximum suppression), which makes the detectors difficult to optimize and unstable, resulting in a delay in inference speed[2]. However, the target detection algorithm based on Transformer does not need to be processed after NMS[3]. As an end-to-end target detection method, Transformer not only has good real-time performance, but also has good detection accuracy. Existing Transformer-based

DOI: 10.54254/2755-2721/111/2024CH0103

target detection algorithms such as DETR have achieved good results in terms of detection accuracy, but there are still challenges in this area that need to be further addressed:

- Most of the existing Transformer-based models have a complex network structure and use an end-to-end learning approach, which is computationally very large as each prediction frame needs to be processed independently, resulting in algorithms that require more computational resources and are difficult to converge.
- Road disease detection models usually need to be deployed on mobile or edge devices, and the existing models are usually large in volume and need to handle every possible prediction frame, which makes the algorithms run slowly and are not suitable for real-time detection.

To overcome the aforementioned issues, a network structure of road disease target detectors based on RT-DETR is proposed in this research. Considering the requirements of application scenarios, a lightweight improvement scheme is designed for the RT-DETR network model, as follows:

- An efficient hybrid encoder is used in the neck of the model, and the single scale Transformer encoding and multi-scale feature fusion can further improve the computing efficiency while learning language features well. In the detection header, the IoU is used to sense the query selection and improve the accuracy of model detection and identification to the target by focusing on high-quality candidate boxes.
- MobileNet is used as a backbone network for feature extraction and the lightweight improvement is carried out through deep separable convolution and point-by-point convolution, which is possible to reduce the calculation amount without reducing the accuracy, speeds up the model reasoning speed, and makes it better deployed on mobile devices.

2. Related work

In this section, we will review some historical machine learning and deep learning-based methods for roadway disease detection, as well as some Transformer-based object detection algorithms.

2.1. Machine learning method

Machine learning methods can generally be divided into two steps, one is to design a feature extractor suitable for a specific task, and the other is to classify the extracted features. For example, SHHanzaei et al. used the local variance invariant operator to extract the crack edge in the tile crack detection statistical method to obtain the crack geometric features, and finally used the support vector machine classifier to identify different crack types[4]. These methods have achieved good results in many crack detection problems. However, in machine learning methods, it is necessary to manually design special feature extractors to extract different crack features. In general, these artificially designed feature extractors have poor generalization ability, can not have good robustness, and the detection effect is poor when the illumination changes greatly and the background is complex.

2.2. Deep learning method

In recent years, many deep learning-based target detection methods at home and abroad have achieved better results compared to traditional machine learning methods. For example, Young-Jin Cha et al. used sliding windows and convolutional neural network methods to detect pavement cracks. The researchers compared the feature extraction capabilities of convolutional neural networks, Canny's and Sobel's algorithms and finally found that the crack classification method using convolutional neural networks achieved the best accuracy[5]. These methods are all two-stage detection methods, which have excellent performance in detection accuracy but can not meet the real-time requirements at the same time. In view of the fact that the two-stage detection method generally does not have a faster detection speed, some scholars have proposed a single-stage detection method. For example, Maeda et al proposed a single-stage road damage detection model based on the original SSD algorithm, which could detect multiple pavement cracks in real time[6]. Sadra Naddsf-Sh et al.proposed a real-time road crack detection model based on EfficientDet[7], and its detection speed is up to 178FPS.

DOI: 10.54254/2755-2721/111/2024CH0103

2.3. Transformer-based approach

The pavement disease detection method based on Transformer is a target detection algorithm based on an attention mechanism. The Transformer network offers powerful modeling capabilities for both text processing and natural speech processing. Compared with sequential models such as recurrent neural networks, this network can model the relationship between elements efficiently. In addition, the Transformer network also supports parallel processing, which can significantly increase reasoning speed. Xiong R et al. 's research shows that a Transformer network has a larger model capacity than a convolutional neural network[8]. That is because the self-attention mechanism in the Transformer network provides dynamic weights to the model, which is reflected in that the weight of the model will change according to different input sequences. Therefore, the network can learn the characteristics of input data and the distribution law of labels more deeply. The model's self-attention mechanism processes the input in such a way that attention is dynamically assigned to different locations of the input to better capture the correlation between elements.

3. Method

In this section we describe the various components of the model in detail, which consists of several components: backbone network, hybrid encoder, transform decoder and prediction header.

3.1. Backbone network

MobileNet is a convolutional neural network architecture with a relatively simple structure that enables efficient image processing on mobile devices with limited computational resources[9]. Its core idea is to reduce parameters and calculations by depth-separable convolution and point-by-point convolution to achieve light weight. For road disease detection tasks, it is important to have a sufficiently lightweight target detector in order for the target detector to be better applied to mobile devices liable to cell phones, camera surveillance devices, etc.

We tested the performance of the model using each of the three versions of MobileNet as the backbone network. One of them, deep convolution used in MobileNetV1, is a special kind of convolution operation which, unlike traditional convolution, is performed independently for each input channel. After deep convolution, the number of channels in the image changes, and a point-by-point convolution operation is required in order to facilitate the fusion and adjustment of features with other layers. MobileNetV2 introduces the inverted residual structure and linear bottlenecks, which provide more expressive power while still having a small model size and fast inference speed. MobileNetV3 further improves the network structure and performance. MobileNetV3 introduces several improvements, including a stronger nonlinear activation function, a lightweight attention model SE, and a NAS-based implementation of MnasNet[10].

3.2. High-efficiency hybrid encoders

An efficient hybrid encoder is used in the neck part of the model, where the features from the last three stages of the backbone network are used as inputs to this encoder after feature extraction of the image using the backbone network. This hybrid encoder contains two modules, a feature fusion module that uses the structure of a convolutional neural network and a feature interaction module that contains the attention mechanism. Where the first module employs a coding layer of an ordinary Transformer which contains standard Deformable Attention and FFN (Feedforward Neural Network). Deformable Attention in the model is realized by improving the self-attention mechanism. In Deformable Attention, for each key-value pair, the model computes an offset and then uses this offset to adjust the response's value toward its surrounding position for the attention computation. Specifically, the model interpolates for each value toward its surrounding locations to obtain a new set of response values. A bilinear interpolation approach is generally taken for the calculation. The model then performs an attention computation on this new set of response values and applies the resulting attention weights to the original values.

DOI: 10.54254/2755-2721/111/2024CH0103

3.3. IoU-aware query selection

The object query vectors in the model are a set of learnable embeddings that learn the most likely category of the object and the most likely size and location of the detection frame during the training process, and their optimization is done in the decoder stage. However, these object query vectors have no explicit physical meaning, which makes them difficult to interpret and optimize. The main idea of IoU-aware query selection is to reduce the amount of candidate frames processed by the model by pre-screening candidate frames with higher IoU values and to increase the focus on the important candidate frames in order to achieve higher detection accuracy and efficiency. Query selection initializes the object query by selecting the top ranked N features from the encoder by calculating the classification score. For the case where the classification scores are inconsistent with the confidence release, IoU-aware query selection can be performed by imposing a constraint on the model during training so that it allows consistency between IoU scores and confidence scores. Under this constraint, the model selects the top-ranked prediction frames based on the classification scores of the N encoder features that also have high IoU scores. The constraint can be expressed by equation (1):

$$\begin{aligned} L(\hat{y}, y) &= L_{box}(\hat{b}, b) + L_{cls}(\hat{c}, \hat{b}, y, b) \\ &= L_{box}(\hat{b}, b) + L_{cls}(\hat{c}, c, IoU) \end{aligned} \tag{1}$$

where $\hat{y} = \{\hat{c}, \hat{b}\}$ and $y = \{c, b\}$ denote the predicted and true labels, respectively, c denotes the category, and b denotes the detection frame.

IoU-aware query selection can reduce the number of candidate frames that the model needs to process, which reduces the computational cost and time consumption. At the same time, by focusing on high-quality candidate frames, the accuracy of the model's prediction of the target can also be improved

4. Experiment

For the training of the model, it was chosen to use the road damage dataset *RDD2022*[11]. For the RDD2022 dataset, the unlabeled images in it were excluded, and the training and validation sets were divided in the ratio of 7:3 for training. After configuring the environment load the pre-training weights on a piece of *NVIDIA V100 GPU* for training. The training results using different versions of MobileNet as the backbone network and their comparison of each parameter with the RT-DETR benchmark model are shown in Table 1:

Table 1. Model performance and parameters using different backbone networks.

Backbone network	Parameters(M)	Training time	Inference speed (FPS)	mAP(%)
ResNet50	42.8	72h	24.7	62.2
MobileNetV1	12.1	42h	37.8	45.3
MobileNetV2	9.3	48h	34.3	49.4
MobileNetV3	27.5	52h	33.8	54.4

The results of the experiment showed that MobileNetV3 achieves the best results in lightweight improvements. Compared with the relatively simple structure of V1 and V2, the introduction of SE module and other improvements have complicated the backbone network structure to a certain extent, but from the results of these improvements, the benefits outweigh the drawbacks, while reducing a certain number of parameters and training time to keep the model accuracy does not drop too much, and the inference speed of the model reaches a very good level.

In Fig. 1 and Fig. 2 we can see the PR curves of the model with MobileNetV3 as the backbone network and the various loss curves during training, respectively. The PR curves in Fig. 1 show that the model has the worst ability to recognize category 3 and the best ability to recognize category 2. In Fig 2, the curves of the model's performance on the training and validation sets are relatively similar, indicating that the model has good generalization ability. The giou-loss loss curve decreases rapidly

DOI: 10.54254/2755-2721/111/2024CH0103

with the rise of the number of training generations and then tends to flatten out, indicating that the predicted box gradually overlaps with the actual box, and the predictor can gradually find the approximate location of the roadway disease. The cls-loss curve rises rapidly at the beginning of the training period and then decreases slowly, which indicates that the gradient disappearance or gradient explosion problem may

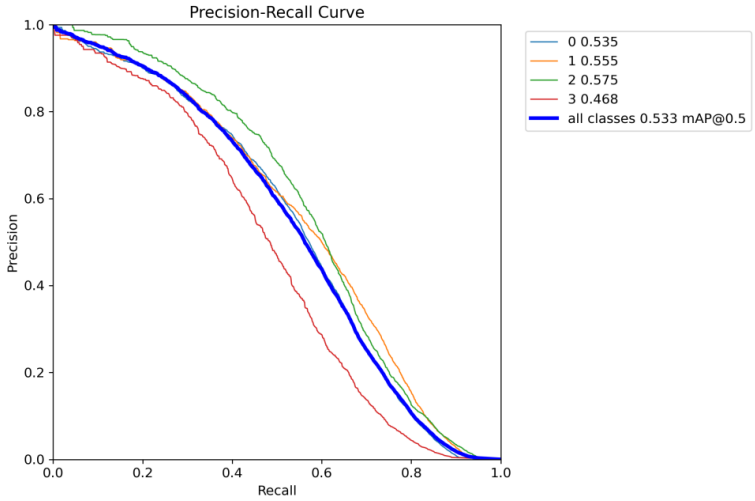

Figure 1. PR curves trained with MobileNetV3 backbone network.

occur at the beginning of the training period, or it may also have something to do with the learning rate setting is related to the dataset, but in the end the model accurate classification probability can still be gradually improved. l1-loss curve changes with the rise of the number of training generations rapidly decreases and then tends to flatten out, indicating that the model is well fitted and robust to outliers. The model's accuracy, recall and mAP50 indicators also show a normal trend of increase, and the mAP50 indicator finally stabilizes at about 54%.

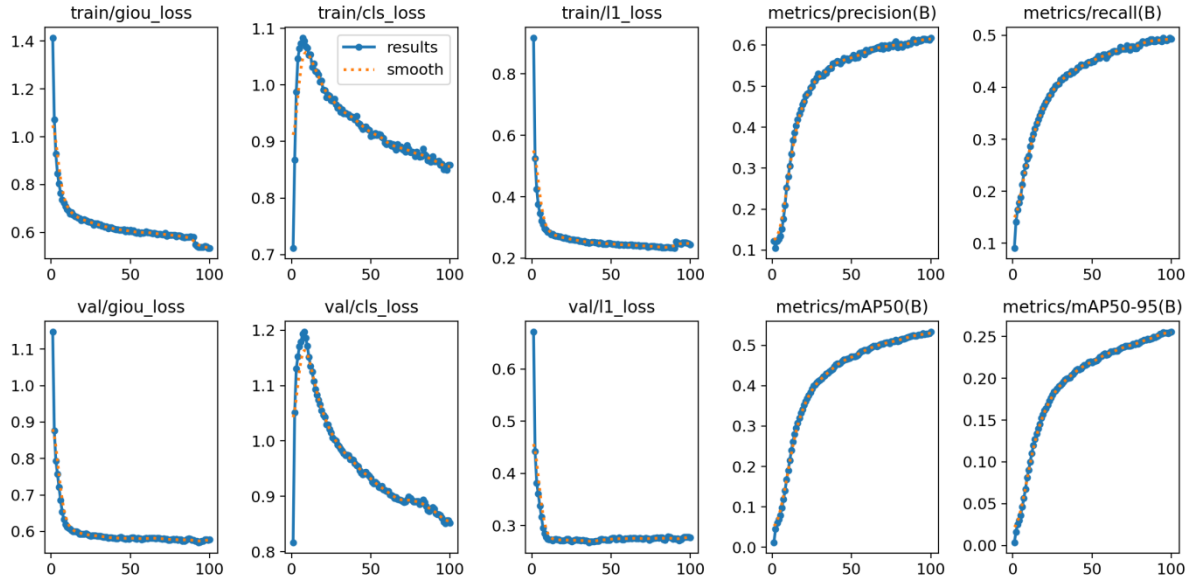

Figure 2. Plot of training results using MobileNetV3 backbone network.

DOI: 10.54254/2755-2721/111/2024CH0103

5. Conclusions

In this paper, a pavement disease recognition model based on RT-DETR is designed. The model realizes feature interaction and fusion at different scales in the encoder stage, which cuts down the computation amount without degrading the detection accuracy. The object query vectors in the Transformer structure are optimized by IoU-aware query selection, which reduces the amount of candidate frames that the model needs to process. For the purpose of making the road disease detector more lightweight for mobile and embedded devices, this paper improves the RT-DETR model by replacing the original backbone network with MobileNet, which performs the computation through two special convolutions that reduce the amount of computation without degrading the accuracy and speed up the model inference, enabling better deployment on mobile or embedded devices.

References

[1] Ding W, Zhao X, Zhu B, et al. An ensemble of one-stage and two-stage detectors approach for road damage detection[C]//2022 IEEE International Conference on Big Data (Big Data). IEEE, 2022: 6395-6400.

[2] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[4] Hanzaei S H, Afshar A, Barazandeh F. Automatic detection and classification of the ceramic tiles' surface defects[J]. Pattern recognition, 2017, 66: 174-189.

[5] Cha Y J, Choi W, Büyüköztürk O. Deep learning-based crack damage detection using convolutional neural networks[J]. Computer-Aided Civil and Infrastructure Engineering, 2017, 32(5): 361-378.

[6] Maeda H, Kashiyama T, Sekimoto Y, et al. Generative adversarial network for road damage detection [J]. Computer-Aided Civil and Infrastructure Engineering, 2021, 36(1):47-60.

[7] Naddaf-Sh S, Naddaf-Sh M M, Kashani A R, et al. An efficient and scalable deep learning approach for road damage detection[C]//2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020: 5602-5608.

[8] Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture[C]. International Conference on Machine Learning.PMLR,2020:10524-10533.

[9] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[10] Tan M, Chen B, Pang R, et al. Mnasnet: Platform-aware neural architecture search for mobile[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2820-2828.

[11] Arya D, Maeda H, Ghosh S K, et al. Rdd2022: A multi-national image dataset for automatic road damage detection[J]. arXiv preprint arXiv:2209.08538, 2022.