Traffic Sign Detection Based on the Fusion of YOLOR and CBAM

Qiang Luo [1], Wenbin Zheng[1,2,*]

[1] *School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, Sichuan, China*
[2] *V.C. & V.R. Key Lab of Sichuan Province, Sichuan Normal University, Chengdu, China, 610068*

Abstract

In the field of traffic sign recognition, traffic signs usually occupy very small areas in the input image. Generally, the Convolutional Neural Networks (CNN) based multi-layer residual networks are used to extract the feature information from these small objects, which often leads to the feature misalignment in the process of feature aggregation. Moreover, most CNN-based algorithms made use of only explicit knowledge, not implicit knowledge. In this paper, a novel method (named YOLOR-A) that combines YOLOR with CBAM is proposed. The CBAM attention mechanism module is integrated to focus the important object. This method can add implicit knowledge into model, which realizes the translation mapping of the feature kernel space and solve the problem of feature misalignment in traffic sign detection. The experimental results show that the proposed method achieves 94.7 mAP, 57 FPS on TT100k dataset, satisfying the real-time detection and outperforming the state-of-the-art methods.

Keywords

Traffic sign detection, Implicit knowledge, Attention mechanism, Feature alignment

1. Introduction

Driver assistance systems and autonomous vehicles have been widely used[1]. As a sub-module, the traffic sign detection system plays an important role in improving driving safety. For the task of traffic sign detection, traffic signs usually only occupy a small proportion of the input image, while extracting high-dimensional features requires multi-level down-sampling, which leads to the loss of characteristic information of small traffic signs[2]. Although the residual structure can alleviate the information loss in the down-sampling process, the residual information fusion process[3] is an indiscriminate combination of context information, which often leads to misalignment in the feature aggregation process[4]. However, the use of implicit knowledge is a good solution to this problem. In deep learning, implicit knowledge refers to the observation-independent knowledge implicit in the model, which can help the model to utilize feature information more effectively. Wang et al.[5] integrated implicit and explicit knowledge into a unified matrix factorization framework for customer volume prediction. Belzen et al.[6] used the implicit knowledge in the neural network to assist in the analysis of protein sensitivity features to achieve protein functional anatomy.

This paper proposes a novel method (named YOLOR-A) that combines YOLOR[7] (You Only Learn One Representation) and CBAM[8] (Convective Block Attention Module).The CBAM attention mechanism is used to focus on the important traffic sign region, and the implicit knowledge is integrated to solve the misalignment problem.

2. YOLOR-A for Traffic Sign Detection

The YOLOR-A model is composed of a backbone feature extraction network, neck network, and recognition head. Backbone uses the network architecture based on CSPDarknet53[9], the core of Neck

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

EMAIL: 1642652860@qq.com (Qiang Luo); zhengwb@cuit.edu.cn (Wenbin Zheng);
ORCID: 0000-0002-7619-648X (Qiang Luo); 0000-0002-9183-796X (Wenbin Zheng);

is the structure of Feature Pyramid Networks and Path Aggregation Networks (PAN[10]), the head uses the structure of YOLO[11] detector, the Align feature alignment module is added to Neck, and the Pre-prediction refinement module is added to head. The YOLOR-A model framework is shown in ***Figure 1***. Then, the CBAM attention module is added after the Neck network to refine the small object features and improve the recognition accuracy.

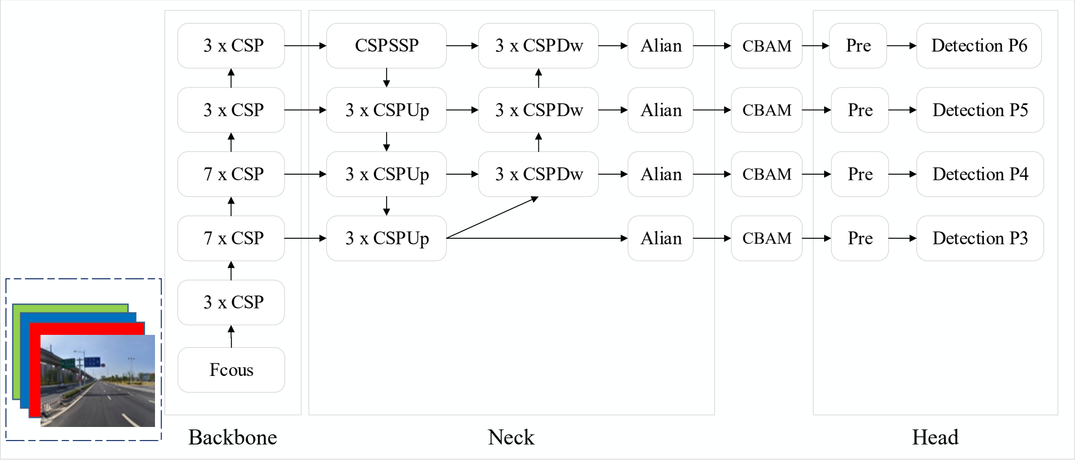


Figure 1: YOLOR-A model framework

2.1. Implicit knowledge learning module

The implicit knowledge in a neural network generally comes from the deep layer of the network, which is the knowledge implicit in the model and not affected by the input value. Therefore, the implicit knowledge representation is independent of concrete input values, which can be regarded as a set of constant tensors Z = (z_1, z_2, …, z_k). Before the introduction of implicit knowledge, the mapping relationship between objects and features can be abstracted as a point-to-point mapping relationship, as shown in **Figure *2***. The CNN-based residual network extracts feature information. In the feature aggregation stage, this simple correspondence is prone to misalignment.

As shown in **Figure *3***, after the introduction of implicit knowledge, the implicit knowledge added to the output features of the neck network structure of the model, and the features can be aligned to the network output through translation transformation, which solves the problem of misalignment in the feature aggregation process. By adding implicit knowledge to the prediction head module and multiplying it with the input features, the point-to-point mapping relationship in the original network can be transformed into a mapping of feature points to range intervals, so that different categories can achieve finer feature mapping, which facilitates the model to distinguish different categories and thus improve the classification accuracy.

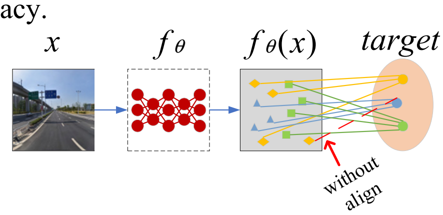


Figure 2: Network with misalignment features

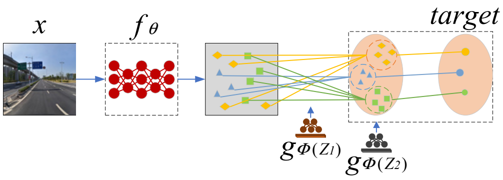


Figure 3: Network with implicit knowledge.

2.2. CBAM attention module

CBAM[8] is a simple but effective attention module. Most of the images are irrelevant foreground information in the traffic sign dataset. Using CBAM can help the model extract effective feature information and focus on the important area for traffic sign.

3. Experiment

3.1. Datasets and Evaluation metrics

TT-100k[12]: TT-100k dataset contains 16,811 images of 2048-2048, which were collected from Chinese street scenes, with a total of 234 types of traffic signs. However, the number of categories varies greatly, so this paper selects 45 categories with the highest frequency for research.

The model detection accuracy evaluation metric uses the Mean Average Precision (mAP[13]). The model detection speed evaluation metric uses Frames Per Second (FPS).

3.2. Results and Analysis

The experimental platform is Ubuntu 20.4.1 operating system, Pytorch-1.7.1 deep learning framework, and the hardware configuration is: graphics GPU NVIDIA GeForce GTX3090, 24GB video memory. The code is written in Python3.7, run on PyCharm platform.

This paper selects the classic two-stage object detection algorithm Faster RCNN, Cascade RCNN[14] algorithm; the single-stage algorithms SSD512[15], yolov5s, and the recently advanced algorithms tph-yolov5[16] and Scaled-YOLOv4[17] in the field of object detection have been compared. The results on test dataset are shown in **Table 1**.

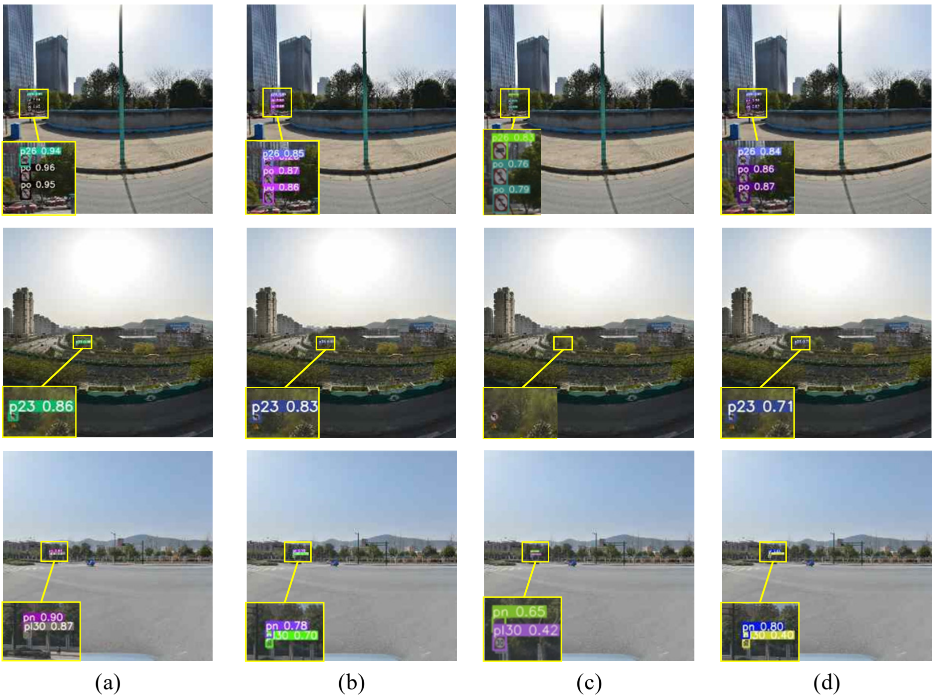
Table 1
The comparison results of different object detectors on TT100k dataset. S, M, L means small size(s<32x32), medium size(32x32<s<96x96), large size(s>96x96).

method	input size	mAP				FPS
		S	M	L	ALL	
SSD512[15]	512x512	28.6	66.6	83.8	68.3	45
Faster RCNN	800x800	13.4	63.7	83.6	59.5	28
Cascade RCNN[14]	800x800	26.5	80.6	91.4	76.1	8
yolov5s	640x640	77.5	80.6	81.4	79.2	333
ScaledYOLOv4[17]	640x640	66.4	79.3	87.7	80.5	166
tph-yolov5s[16]	1280x1280	84.6	92.6	91.5	90.5	45
YOLOR-A(proposed)	1280x1280	**91.8**	**95.5**	**97.2**	**94.7**	57

The ablation experiments show that our proposed algorithm is effective, as shown in **Table 2**.In summary, the proposed algorithm YOLOR-A combined with the CBAM attention mechanism and using the implicit knowledge for traffic sign detection, has the best detection accuracy and the competitive speed.

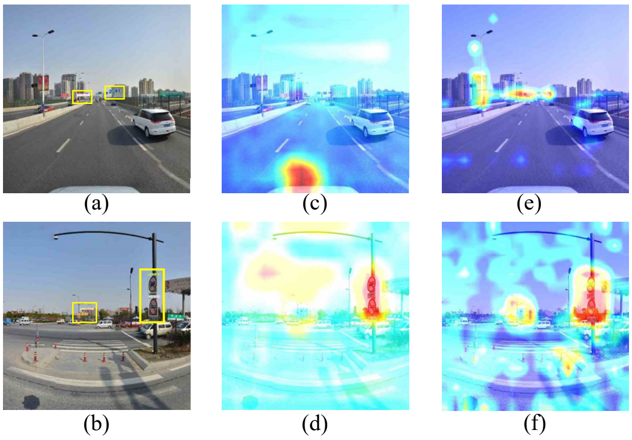
Table 2
The comparison results of different object.

methods		TT100k			
		S	M	L	ALL
YOLOR	baseline	87.6	91.7	88.3	88.1
	+Align	89.7	92.0	89.5	91.2
	+Pre	90.2	92.2	88.1	91.8
	+ (Align & Pre)	91.6	94.3	96.7	93.8
	+ (Align & Pre & CBAM)	**91.8**	**95.5**	**97.2**	**94.7**



(a) (b) (c) (d)

Figure 4: Visual detection performance of TT100k dataset. (a): The detection effect of YOLOR-A. (b): The detection effect of tph-yolov5. (c): The detection effect of ScaledYOLOv4. (d): The detection effect of yolov5.



(a) (c) (e)

(b) (d) (f)

Figure 5: Feature visualization. (a)(b): TT100k dataset picture. (c)(d): Neck network feature visualization output without implicit knowledge and CBAM. (e)(f): Neck network feature visualization output with implicit knowledge and CBAM.

Some detection examples are shown in **Figure *4***. The algorithm YOLOR-A has the best detection effectiveness compared with tph-yolov5, ScaledYOLOv4, and yolov5, and its corresponding detected traffic signs have the highest confidence, especially for small objects.

Based on the heat map visualization experiments, are shown in **Figure *5***. we can conclude that the problem of algorithmic feature misalignment is solved with the inclusion of implicit knowledge.

4. Conclusion

In this paper, a traffic sign object detection algorithm based on the fusion of YOLOR and CBAM is proposed. This method can make use of the implicit knowledge in a neural network to overcome the feature misalignment problem, and incorporates the CBAM attention mechanism so that the object detector can focus on the important feature area for the traffic sign. The experimental results show that the proposed algorithm obtain better performance compared with other competitive algorithms.

5. Acknowledgements

This work is supported by the Natural Science Foundation of Sichuan, China (No. 2022NSFSC0571) and the Sichuan Science and Technology Program (No. 2018JY0273, No. 2019YJ0532). This work is supported by funding of V.C. & V.R. Key Lab of Sichuan Province (No. SCVCVR2020.05VS). This work is also supported by the China Scholarship Council (No. 201908510026).

6. References

[1] C. Han, G. Gao, Y. Zhang, Real-time small traffic sign detection with revised faster-RCNN, Multimedia Tools and Applications, 78 (2019) 13263-13278.

[2] L.L. Shen, L. You, B. Peng, C.H. Zhang, Group multi-scale attention pyramid network for traffic sign detection, Neurocomputing, 452 (2021) 1-14.

[3] X. Liu, Pedestrian Reidentification Algorithm Based on Local Feature Fusion Mechanism, Journal of Electrical Computer Engineering, 2022 (2022).

[4] Z.L. Huang, Y.C. Wei, X.G. Wang, W.Y. Liu, T.S. Huang, H. Shi, AlignSeg: Feature-Aligned Segmentation Networks, Ieee Transactions on Pattern Analysis and Machine Intelligence, 44 (2022) 550-557.

[5] J. Wang, Y. Lin, J. Wu, Z. Wang, Z. Xiong, Coupling Implicit and Explicit Knowledge for Customer Volume Prediction, The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)2017).

[6] J.U. zu Belzen, T. Burgel, S. Holderbach, F. Bubeck, L. Adam, C. Gandor, M. Klein, J. Mathony, P. Pfuderer, L. Platz, M. Przybilla, M. Schwendemann, D. Heid, M.D. Hoffmann, M. Jendrusch, C. Schmelas, M. Waldhauer, I. Lehmann, D. Niopek, R. Eils, Leveraging implicit knowledge in neural networks for functional dissection and engineering of proteins, Nature Machine Intelligence, 1 (2019) 225-235.

[7] C.-Y. Wang, I.-H. Yeh, H.-Y.M. Liao, You only learn one representation: Unified network for multiple tasks, arXiv preprint arXiv:2105.04206, (2021).

[8] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV)2018), pp. 3-19.

[9] A. Bochkovskiy, C.Y. Wang, H. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv:2004.10934, (2020).

[10] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition2018), pp. 8759-8768.

[11] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv:1804.02767, (2018).

[12] Z. Zhe, D. Liang, S. Zhang, X. Huang, S. Hu, Traffic-Sign Detection and Classification in the Wild, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)2016), pp. 2110-2118.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, European conference on computer vision, (Springer2014), pp. 740-755.

[14] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, Proceedings of the IEEE conference on computer vision and pattern recognition2018), pp. 6154-6162.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, European conference on computer vision, (Springer2016), pp. 21-37.

[16] X. Zhu, S. Lyu, X. Wang, Q. Zhao, TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios, Proceedings of the IEEE/CVF International Conference on Computer Vision2021), pp. 2778-2788.

[17] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, Scaled-yolov4: Scaling cross stage partial network, Proceedings of the IEEE/cvf conference on computer vision and pattern recognition2021), pp. 13029-13038.