Journal of computer science and software applications
https://www.mfacademia.org/
ISSN:2377-0430
Vol. 3, No. 4, 2023

Enhanced YOLOX-L Algorithm with CBAM for Construction Safety Detection

Victoria Turner

Widener University

Victoria99@widener.edu

Abstract:

Many safety accidents in the construction industry are attributed to unsafe behavior, particularly the incorrect use of safety protective equipment. To address this issue, this paper proposes an enhanced YOLOX-L algorithm, which integrates the Convolutional Block Attention Module (CBAM) to improve the Neck component of the original network. This modification enhances the network's ability to extract features at various scales. While maintaining real-time detection speed, the mAP (mean Average Precision) trained with a custom dataset reaches 85.96%, which is approximately 1.1% higher than the original YOLOX-L algorithm. This effectively enables fast and accurate detection of unsafe behavior.

Keywords:

YOLOX-L; Unsafe Behavior Detection; Attention; CBAM.

1. Introduction

The construction industry is a high-risk industry with frequent safety accidents. According to statistics from the Ministry of Housing and Urban-Rural Development, although the number of safety accidents and deaths in the construction industry has decreased in recent years, it is stillrising. According to the accident causal chain theory [1], people's unsafe behavior is the main and direct cause of accidents. According to the statistics of the National Security Bureau [2], more than 70% of safety accidents are caused by improper wearing of safety protective equipment and personnel illegally invading dangerous areas. Safety protective equipment suchas helmets, seat belts, reflective clothing, etc. play an important role in protecting personal safety. However, because of their weak safety awareness, some workers do not wear them correctly every day, which leads to tragedies.

Therefore, in order to reduce the occurrence of safety accidents, it is necessary to realize real- time supervision of unsafe behaviors of construction workers. Especially in recent years, with the continuous development of deep learning algorithms, many scholars have begun to use high-performance deep learning algorithms to complete high-precision real-time detection of unsafe behavior of workers. In this paper, the YOLOX-L detector will be used as the basic network, and the attention module will be added on this basis to strengthen the feature extraction ability of small targets and improve the detection accuracy of small targets such as helmets and reflective clothing.

2. Related Works

The unsafe behaviors to be detected in this paper according to actual needs include not wearing a helmet, not wearing reflective clothing, etc. In the field of computer vision, such situations that need to detect multiple unsafe behaviors at the same time belong to the multi-target detection problem. The traditional sensor-based the automatic identification method cannot meet such needs, and the manual inspection method not only consumes a lot of manpower and material resources, but also cannot achieve the effect of real-time detection. Therefore, a target detection algorithm based on deep learning needs to be used.

At present, the target detection algorithms based on deep learning are mainly divided into two categories. One is the R-CNN series algorithm based on the candidate frame (Region Proposal), which belongs to the two-stage algorithm. This kind of algorithm first generates a series of algorithms by the algorithm. As the candidate frame of the sample, the convolutional neural network is used for sample classification and frame regression. Common algorithms include R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], R-FCN [6], etc.; another type of It is a one-stage algorithm based on YOLO [7] and SSD [8]. This kind of algorithm does not need to generate candidate frames, but only uses a convolutional neural network to directly predict the categories and positions of different targets, and convert the target frame positioning problem into regression. Problem handling, common algorithms include SSD, RetinaNet, YOLO series, etc. The two-stage algorithm improves the detection accuracy on the premise of sacrificing the detection speed, while the one-stage algorithm directly uses CNN to extract features to complete the classification and positioning, and the detection speed is significantly improved, but the accuracy is low. Liu Xinyi et al. [9] improved the Faster R-CNN algorithm in the detection of safe clothing on contaminated sites, and improved the detection speed of the model by introducing the L2 regular term into the regression loss function, but the detection time still reached 44ms, according to the construction According to the actual needs of the site, the one-stage algorithm has stronger comprehensive performance and is more widely used. At present, there are many researches on detecting single unsafe behavior based on one-stage algorithm. For example, Huang et al. [10] increased the feature map scale on the basis of the original YOLO v3 algorithm, optimized the priori dimension algorithm of a specific helmet data set, and improved the Loss function, the final mAP trained in the helmet data set reached 93.1%, an increase of 3.5% compared with the original YOLO v3 algorithm, and the detection speed was also improved. In addition, the literature [11,12,13,14] are all tests on the wearing of safety helmets, and a few literatures have begun to test the reflective clothing [15,16] However, the research on the simultaneous detection of multiple target behaviors and multi-scale targets still needs to be further improved.

To sum up, in order to solve the problems of complex objective environment and numerous small targets on the construction site, and to further improve the detection ability of the model for multi-scale and multi-targets, this paper intends to select the YOLOX algorithm with the best performance as the basic network for research, and strengthen its characteristics. The extraction network part is improved to further improve the detection accuracy of small targets, making it more suitable for target detection tasks in construction site scenarios. Finally, train other models with the same dataset to verify its effect.

3. Related Algorithms and Improvements

3.1. YOLOX-L Algorithms

According to the survey, the latest research result of the first-stage algorithm is the YOLOX [17] algorithm proposed by the Megvii Science and Technology Research Institute in 2021. This algorithm adds some refreshing improvements on the basis of the previous version of YOLO, such as Decoupled Head, Mosaic combined with Mixup for data enhancement, Anchor-free mechanism, SimOTA label assignment strategy, etc. These optimizations make YOLOX a new high-performance anchor-free detector, which surpasses other detectors in speed and detection accuracy.

According to the experimental results of the original literature, when the parameters are almost the same, the AP value of YOLOX-L trained on the COCO dataset is 50%, which is 1.8% higher than that of YOLOv5-L at almost the same detection speed. The AP of YOLOX-Darknet53 is 3% higher than that of YOLOv5-Darknet53 at the same speed. In addition, YOLOX-L achieves 50% AP detection speed of 68.9FPS on a single Tesla V100, which is a very competitive value.

3.2. Convolutional Block Attention Module

Attention [17] mechanism is to apply the human perception method and the attention behavior of the human brain to the machine, so that the machine can automatically obtain important information through learning, which effectively improves the problem of model performance degradation when the image size is too large. applied to computer vision. In 2017, Hu et al. proposed the SENet (Squeeze-and-Excitation Networks) [18] model for the first time by applying the Attention idea. By explicitly modeling the relationship between feature channels, the expressive ability of the model was improved. Although the compression and excitation modules in SENet can filter out the attention for the channel, it increases a certain number of parameters and calculation, and cannot obtain the positional relationship information well. In order to improve this problem, Woo et al. In addition, spatial dimension modeling is added, and the CBAM [20] module is proposed, which enables the feature map to infer the attention map in turn along two independent dimensions, and then multiplies the attention map with the input feature map for adaptive feature optimization, saving parameters. and computing power, effectively improving the classification and detection performance of the model. Moreover, the CBAM module is a lightweight attention module that can ignore the overhead of this module, and can be seamlessly integrated into the existing CNN architecture as a plug-and-play part forend-to-end training.

The main target category detected in this study is in the upper body area of the person. For image data with a long target distance, after repeated up-sampling operations, the resolution of the feature map will gradually decrease, resulting in a decrease in the target features of the high-level feature map. However, the low-level high-resolution feature map has only weak semantic features and location details after multiple downsampling operations, and the detection accuracy is greatly reduced. Therefore, in order to strengthen the local information of the feature map, so that the model can accurately locate the key feature areas, and enhance the representation ability of the feature map, this study combines the CBAM module into the YOLOX-L model to effectively deal with the construction scene. The detection of local information of personnel such as reflective clothing further improves the detection accuracy.

3.3. Improved Algorithm

The model structure of YOLOX is mainly divided into three parts, namely the backbone feature extraction network CSP Darknet, the enhanced feature extraction part FPN (Feature Pyramid Networks, feature pyramid network) and the feature prediction part YOLO Head. In order to make it more suitable for the small target detection task in the construction site scene, an attention mechanism is added to the enhanced feature extraction network part. In theory, the attention module can be added after any feature map, but the features generated by the backbone part are the basis for feature fusion and target prediction, and the weights of the attention module need to be randomly initialized, so in order not to destroy the backbone part Weights affect the effect of network prediction. In this study, after adding the CBAM module to the main part, before the effective feature layer is input into the feature pyramid network, the neck part of the original structure is improved, and the Attention-FPN structure is proposed. After training, it is suitable for the structure of the improved algorithm model is shown in Figure 1 below.

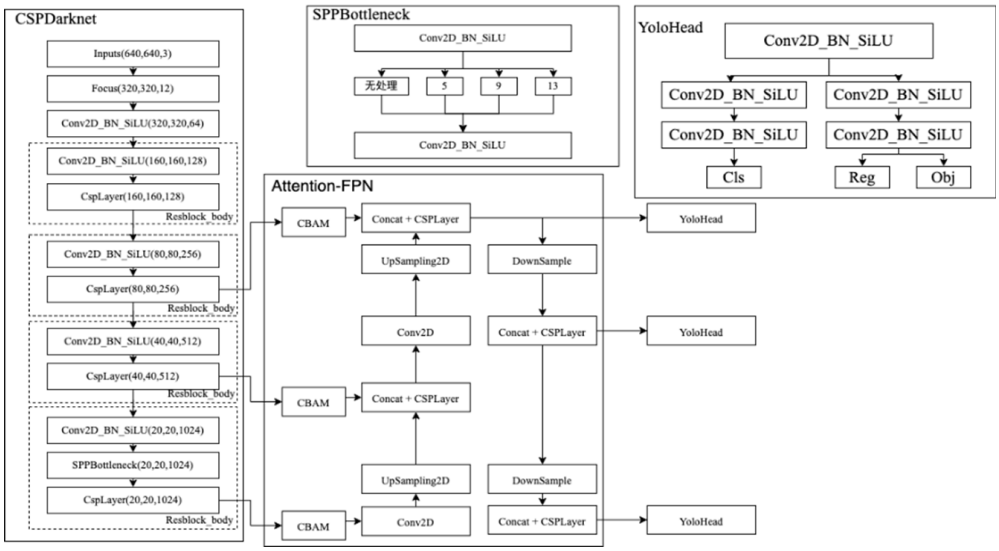

Fig 1. Improved YOLOX Model Structure

The overall structure of the improved YOLOX is shown in Figure 3. After the CBAM module outputs three feature layers after enhanced feature extraction, the shape sizes are (80, 80, 256),(40, 40, 512) and (20, 20, 1024) respectively. After the Decoupled Head prediction, each feature layer obtains three prediction results, which are the coordinates of the target frame (Reg), the target frame foreground and background judgment (Obj), and the target frame category (Cls). In Figure 4, the number of convolution channels of the Reg prediction result is 4, which represents the offset of the center point of the prediction frame compared to the feature point and the width and height of the prediction frame compared to the parameters of the logarithmic index; The number of convolution channels is 1, which represents the probability that each feature point prediction box contains an object. The number of convolution channels of the Cls prediction result is num_classes, which represents the probability that each feature point corresponds to a certain type of object, and the predicted value in the last dimension num_classes represent the probability of belonging to each class.

4. Experiment Procedure

4.1. Dataset Construction

The OpenCV method is used to achieve video frame extraction, and the frequency is to extract one frame every 2 seconds, and then the construction site data in the real scene is obtained through manual screening. In addition, in order to enrich the data set, python crawler code was also used to crawl the image data on the website according to keywords such as safety helmet", "reflective clothing", "construction", "construction site protective equipment". In the end, a total of 18,251 images were obtained. Images that meet the training requirements. Use the LabelIng image annotation tool to label the data according to five categories of person, head, head_helmet, fgy, and wcfgy, representing people, not wearing helmets, wearing helmets, reflective clothing, and not wearing reflective clothes, and then divide the training set and test set according to the ratio of 9:1.

4.2. Experiment Environment

Table 1. Experimental Setting

Experimental Environments	Version
Operation System	Ubuntu 16.04
GPU	NVIDIA GeForce GTX 1080Ti (64G)
CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
Framework	PyTorch 1.1.0
Compilation environment	Python3.7

The hardware and software environment of this experiment is shown in Table 1.

5. Experimental Results

In order to verify whether the improved model proposed in this paper improves the detection performance of helmets, seat belts and reflective clothing, YOLO v4 and YOLO X-L are selected in this experiment to compare with this method, and compare the performance of different models under the same self-made data set. AP, mAP and detection time (t) per frame, the resultsare shown in Table 2 below.

Table 2. Experimental Results

	YOLO V4	YOLOX-L	Improved Model
AP(head)/%	82.98	82.4	89.96
AP(head_helmet)/%	88.35	88.99	94.44
AP(person)/%	93.67	95.41	97.8
AP(aqd)/%	78.95	82.32	72.66
AP(wdaqd)/%	85.09	87.51	89.89
AP(aqd_unknown)/%	68.31	68.48	73.62
AP(fgy)/%	86.64	83.13	85.69
AP(wcfgy)/%	82.02	82.83	83.66
mAP	83.25	84.88	85.96
t/ms	23.32	23.51	23.7

It can be seen from Table 2 that the AP value obtained from training of each category using the improved model in this study has increased to varying degrees. Therefore, the detection accuracy of reflective clothing is still relatively low. Although the detection speed of each image of the improved model is also 0.38ms slower than that of YOLO v4, the mAP trained by the model is about 1.1% higher than that of the original YOLOX-L model and about 3.7% higher than that of the YOLO v4 model. The proposed algorithm can still meet the demand for high- precision real-time detection on the construction site. The figure 2 below shows some of the prediction results.

a) b)

Fig 2. Prediction Results

6. Summary

This paper studies the unsafe behaviors such as failure to properly wear safety helmets, safetybelts, and reflective clothing at the construction site. The improved YOLOX-L algorithm based

on the attention mechanism is used to improve the detection accuracy of unsafe behaviors, while satisfying the demand for real-time detection can improve the unsafe behavior of the construction site to a certain extent. However, this paper is limited by the lack of data sets, and has not yet completed the detection of other unsafe behavior types, such as human-machine collision, etc., and further research is needed in the future.

References

[1] Heinrich H W.Industrial accident prevention: a scientific approach [M]. 2nd ed.London: McGraw- Hill Book Co, Inc, 1941.

[2] Chang X, Liu X M. Fault tree analysis of unreasonably wearing helmets for builders. Journal of Jilin Jianzhu University. Vol.35(2018) No.6, p. 67-71.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. IEEE Computer Society. IEEE Computer. Society, 2013.

[4] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[6] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.

[7] 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.

[8] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[9] Liu X.Y., Zhang B.F. Fu Y., et al. Detection on normalization of operating personnel dressing at contaminated sites based on deep learning. Journal of Safety Science and Technology.Vol.16(2020) No. 07, p. 169-175.

[10] Huang L, Fu Q, He M, et al. Detection algorithm of safety helmet wearing based on deep learning[J]. Concurrency and Computation: Practice and Experience, 2021, 33(13): e6234.

[11] FU J, CHEN Y, CHEN S. Design and Implementation of Vision Based Safety Detection Algorithm for Personnel in Construction Site[J]. DEStech Transactions on Engineering and Technology Research, 2018 (ecar).

[12] Li Y, Wei H, Han Z, et al. Deep learning-based safety helmet detection in engineering management based on convolutional neural networks[J]. Advances in Civil Engineering, 2020, 2020.

[13] Mneymneh B E, Abbas M, Khoury H. Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications[J]. Frontiers of Engineering Management, 2018, 5(2): 227-239.

[14] Wu F, Jin G, Gao M, et al. Helmet detection based on improved YOLOv3 deep model[C]. 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2019.

[15] Sun S, Zhao S, Zheng J. Intelligent Site Detection Based on Improved YOLO Algorithm[C]//2021 International Conference on Big Data Engineering and Education (BDEE). IEEE, 2021: 165-169.

[16] Linder T, Griesser D, Vaskevicius N, et al. Towards accurate 3D person detection and localization from RGB-D in cluttered environments[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)–Workshop on Robotics for Logistics in Warehouses and Environments Shared with Humans. 2018.

[17] Ge Z, Liu S, Wang F, et al. YOLOX: exceeding YOLO series in 2021[J]. ArXiv preprint arXiv: 2021, 2107. 08430.

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

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[20] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.