YOLOv5-AC

Subjects: Computer Science, Artificial Intelligence
Contributor: HaoHui Lv

Attention Mechanism-Based Lightweight YOLOv5 for Track Pedestrian Detection. In response to the dangerous behavior of pedestrians roaming freely on unsupervised train tracks, the real-time detection of pedestrians is urgently required to ensure the safety of trains and people. Aiming to improve the low accuracy of railway pedestrian detection, the high missed-detection rate of target pedestrians, and the poor retention of non-redundant boxes, YOLOv5 is adopted as the baseline to improve the effectiveness of pedestrian detection by model pruning, improving attention mechanism, etc.

Keywords: pedestrian detection ; deep learning ; model pruning ; context extraction module ; attention module ; DIoU_NMS

1. Introduction

As rail transportation plays an increasingly important role in China, the safety of rail transit operations has also attracted more and more attention. As a consequence, It is of great significance to carry out research on pedestrian detection and abnormal state monitoring at railway stations to ensure the safety of pedestrians. pedestrians usually move fast and irregularly on the railway track, while the target is very small and has a high degree of coincidence of body positions within the visual range of the machine's vision. In addition, complex and uncertain environmental factors such as trees, weeds, and telephone poles around the railway track have caused huge obstacles to pedestrian detection. In order to ensure pedestrian safety, we will experimentally improve our YOLOv5 through the following five points to achieve better pedestrian detection results.

(1) L1 [1] regularization is added to constrain the scaling factor of the BN [2] layer to make the activation coefficients sparse. Next, the modified model is sparsely trained to cut out the sparse layers. We end up with a very compact model with repeated cutting.

(2) In Backbone, the CEM module is introduced to fully extract the features of different scales. The CxAM module is introduced to extract context semantic information to improve recognition accuracy. The CnAM module is introduced to correct the position of F5 layer features and improve the accuracy of target box regression.

(3) DIoU_NMS is used instead of NMS to filter prediction boxes to avoid eliminating different target prediction boxes with high consistency.

(4) The researcher collected a certain number of datasets along with a certain number of relevant public datasets to provide data support for the verification of the actual effect of the improved model.

(5) According to the direction of improvement, a number of related ablation experiments were designed to verify the validity of each contribution.

2. Improved YOLOv5-AC

2.1. YOLOv5 Network Structure

The YOLO series of algorithms, from YOLOv1 to YOLOv5 [3], has been the hottest algorithm in the field of target detection due to its fast and efficient performance. The latest generation of YOLOv5's weight files is only 28 MB, which is ideal as an initial model. Therefore, YOLOv5 is selected as the experimental object for algorithm improvement in this experiment. The network structure of YOLOv5 generally follows the previous series. The feature extraction network of the backbone adopts CSPDarknet [4]. The input newly adds the focus structure, slices the input image, reduces the size, and increases the depth, which can improve the speed of feature extraction. At the same time, the CSP2 structure is deployed to the neck part to enhance the ability of network feature fusion. The optimization function adopts Adam [5] and SGD [6]. Focus and conv are the structures that mainly contain the convolution kernel and residual components. As a consequence, the

network depth can be changed by controlling the number of residual components in Conv, while the network width can be adjusted by gaining command of the number of convolution kernels in Focus and Conv. Therefore, YOLOv5 has launched four models ranging from small to large by regulating the parameters: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The YOLOv5 network structure is shown in **Figure 1**.

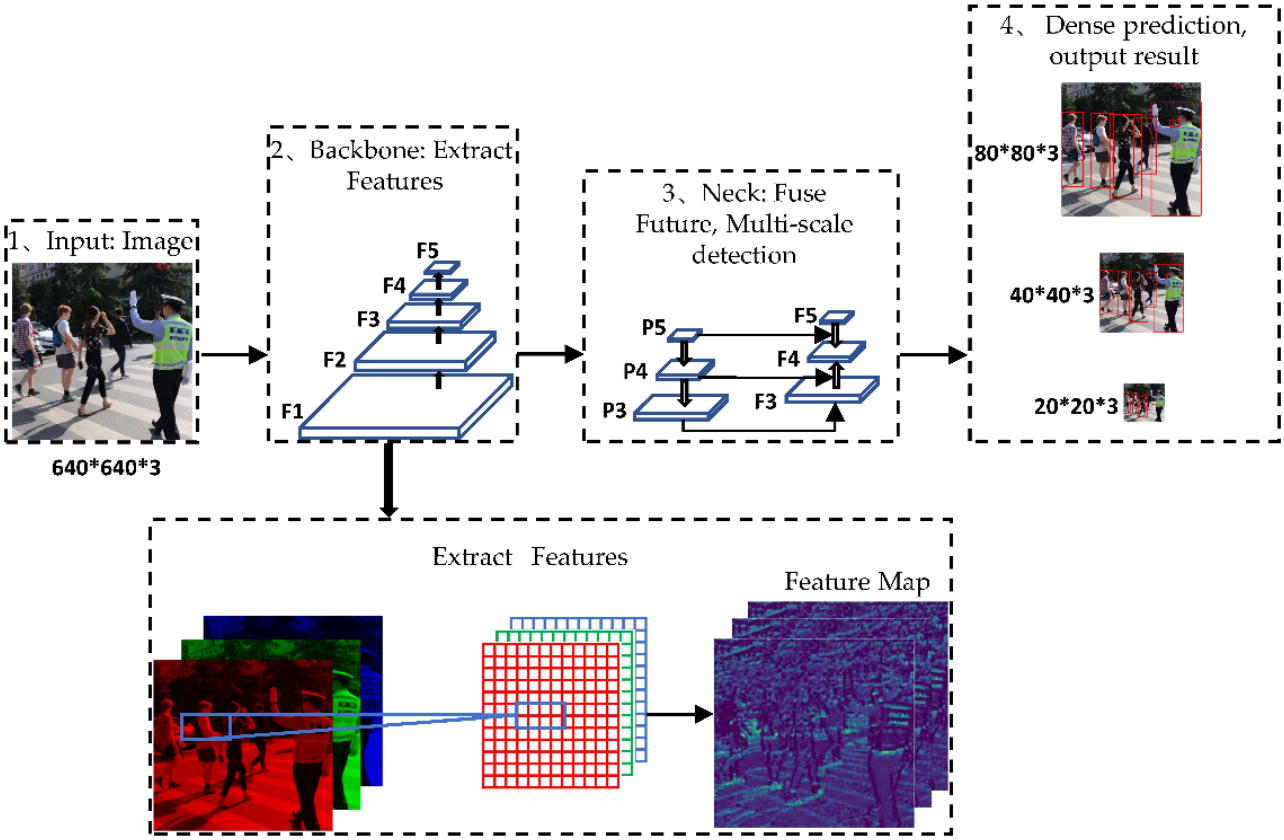

Figure 1. YOLOv5 algorithm structure diagram.

2.2. Sparse Training and Model Pruning

YOLOv5 is already a cracking lightweight detection network where the trained weight model generally does not exceed 30 MB, which is still too large for some embedded devices. If we simply choose to reduce the size of the network input, such as 640 to 320, as the size of the model is reduced accordingly the detection effect will also have a greater loss at the same time. Therefore, according to a method of network slimming proposed by Zhuang Liu et al. [7], we add L1 regularization parameters to the model to constrain the scaling factor of the BN layer, which can cause the coefficients close to 0 to become smaller. These pairs of parameter layers with little influence on forward propagation are eliminated through sparse training. We can obtain a very compact and efficient network model by repeating the above operations.

The principle of YOLOv5 network channel clipping is shown in the **Figure 2**.

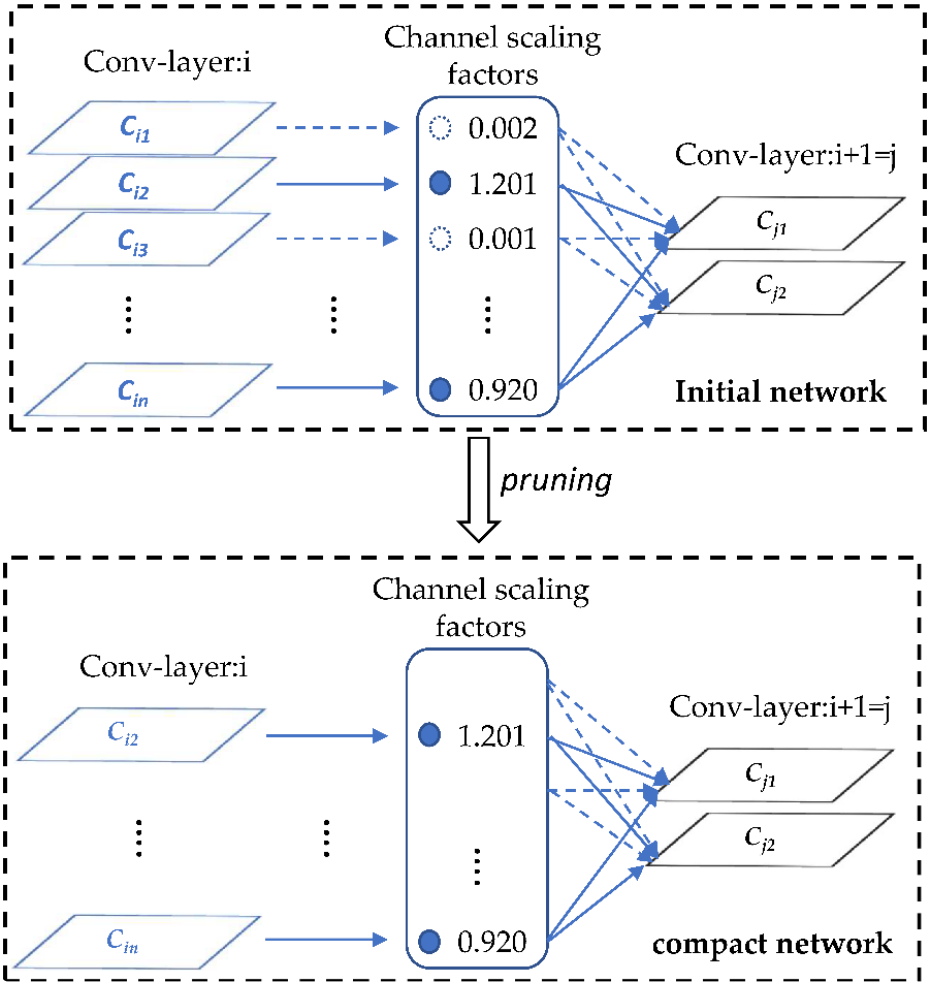

Figure 2. Principle of pruning.

Above all, the researcher append L1 regularization to the model to perform corresponding sparse training. Then, channel pruning is performed on the trained model. Ultimately, the training hyperparameters are fine-tuned to ensure the model inference results are optimal. The algorithm implementation process is shown in **Figure 3**.

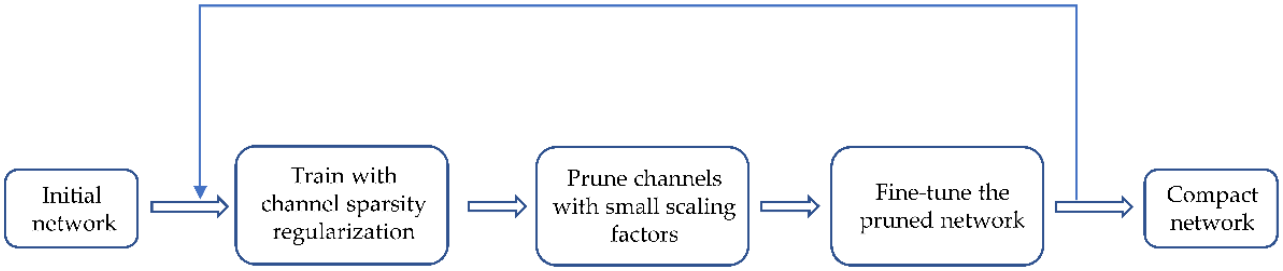

Figure 3. Process of pruning.

2.3. AC_FPN structure

The output is sent to the deconvolution layer of the FPN network for feature fusion after the feature maps are processed by the above three modules. The improved AC_FPN [8] structure is shown in **Figure 4**.

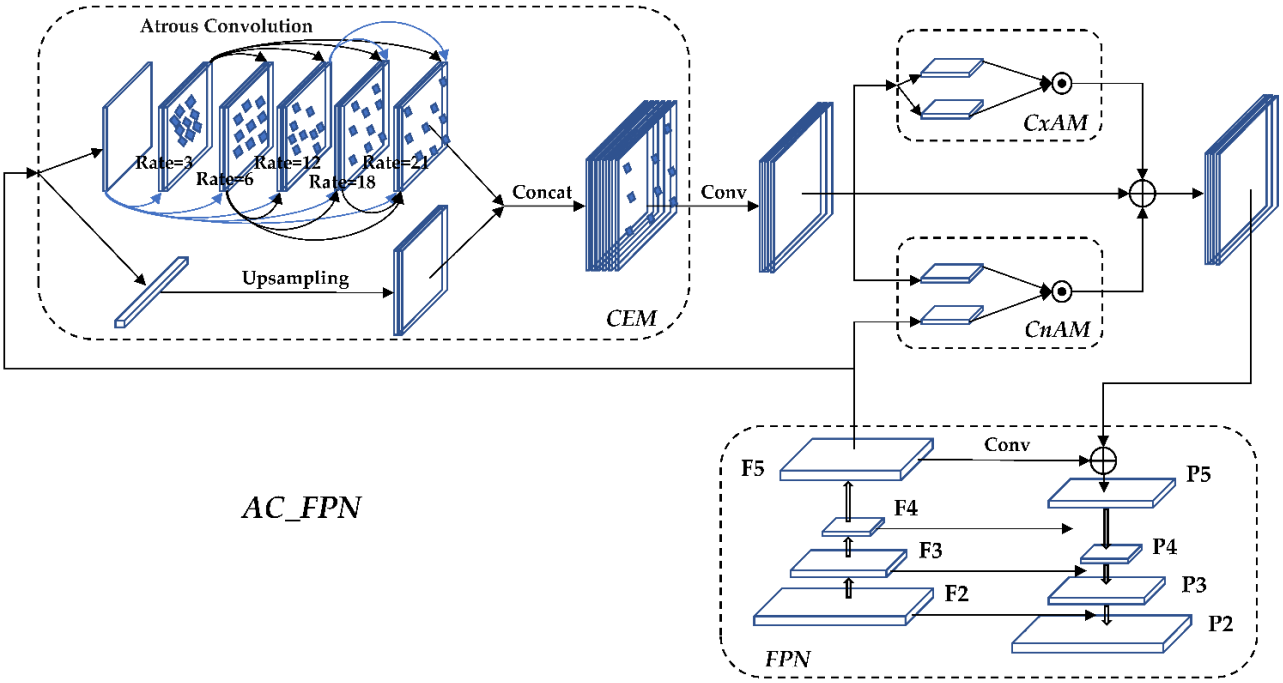

Figure 4. AC_FPN structure diagram.

2.4. Improved NMS

Non-Maximum Suppression (NMS) needs to be performed for the screening of many target boxes in the post-processing process of target detection. YOLOv5 adopts the traditional NMS [9] method. The occluded target selection box is usually removed when facing two different targets with a high degree of coincidence by using this method. For an environment with a large number of targets, where there will be many targets with a high degree of coincidence, the occlusion target candidate boxes that are obscured will be removed as redundant information by NMS, which is not suitable for models that want to detect accurately. DIoU_NMS is used to replace the NMS. DIoU_NMS introduces the parameter β of the center point distance between the two boxes. When β →∞, DIoU_NMS degenerates into traditional NMS. Otherwise, as long as the center points of the two frames do not coincide perfectly when β → 0, they will be retained by DIoU_NMS. As a consequence, the value of β can be adjusted to 0 → ∞ according to the actual situation to achieve the best effect to restrain redundant boxes. Its classification score update formula is defined as Formula (1):

$$s_i = \begin{cases} s_i, IoU - R_{DIoU}(M, B_i) < \epsilon \\ 0, IoU - R_{DIoU}(M, B_i) \geq \epsilon \end{cases}, \quad (1)$$

where si is the classification score and e is the NMS threshold, RDIoU(M, Bi) is the penalty item, M is the predicted box with the highest score, and Bi is the other box.

2.5. Improved YOLOv5-AC Network Structure

The features are further extracted by adding a context extraction model (CEM) in Backbone. CxAM is added to extract the context semantics. CnAM is applied to correct the feature positions of the F and F5 layers. Post-processing uses DIoU_NMS to replace NMS. The improved YOLOv5-AC structure is shown in **Figure 5**.

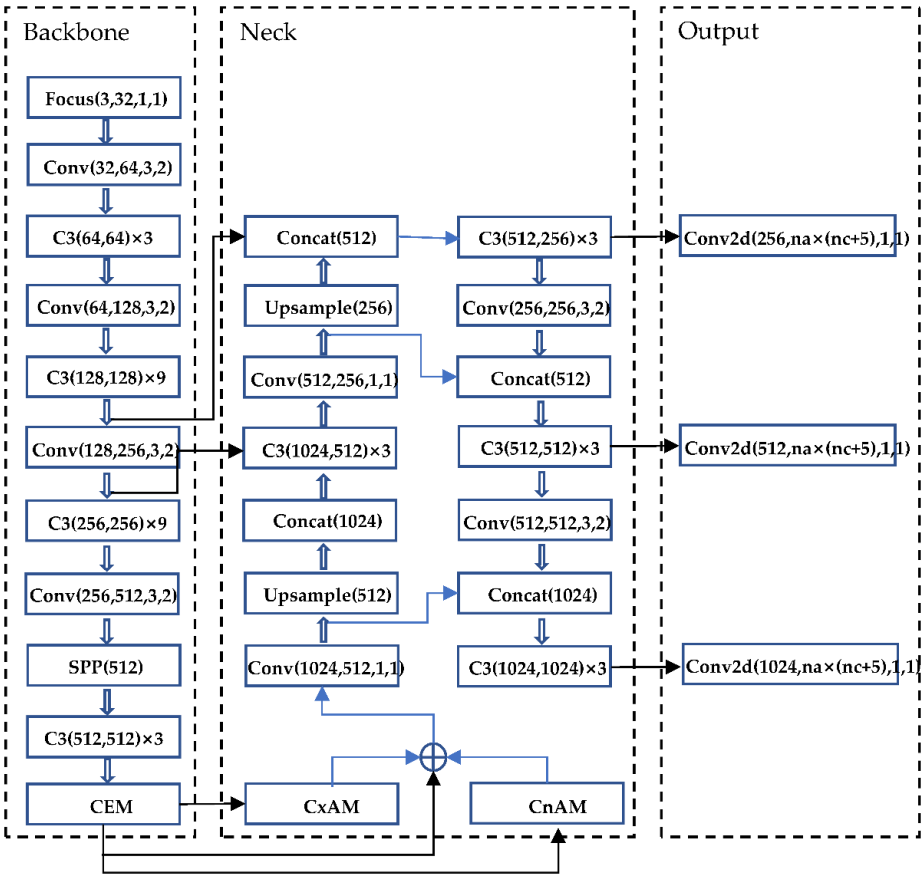

Figure 5. YOLOv5-AC structure diagram.

3. Experiment

3.1. Training Process

This experiment will proceed as shown in **Table 1**.

Table 1. Procedure of the experiment.

Procedure of the Experiment
Step1: Carrying out pruning experiments to select the L1 parameter that makes the pruning rate, P and R optimal.
Step2: Training YOLOv5s to get training metrics.
Step3: Training YOLOv5-AC to get training metrics.
Step4: Contrasting the results of step2 and step3.
Step5: Comprehensive comparison of mainstream models such as YOLOv5-AC and Faster R-CNN, YOLOv4, Efficientdet-B3 [37].
Step6: A series of ablation trial are designed to verify the validity of every contribution.

3.2. Training Metrics

Accuracy and Recall are selected as metrics to compare the quality of the original model and the improved model of the test results. The calculation formulas of Accuracy and Recall are as follows:

$$Recall = \frac{TP}{TP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP is the number of people on the track that were correctly detected. FP is the number of people on the track that were incorrectly detected as people. TN is the number of people on the track that were not detected. FN represents no one on the track and no one detected at the same time. The relationship can be intuitively understood through the following confusion matrix **Table 2**.

Table 2. Confusion matrix.

	Real Situation Somebody	Real Situation Nobody
Predicted somebody	*TP*	*FP*
Predicted nobody	*TN*	*FN*

References

1. Kang, G.; Dong, X.; Zheng, L.; Yang, Y. Patchshuffle regularization. arXiv 2017. preprint.
2. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6 July 2015; pp. 448–456.
3. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 9 June 2020).
4. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13 June 2020; pp. 390–391.
5. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv 2014. preprint.
6. Zheng, S.; Meng, Q.; Wang, T.; Chen, W.; Yu, N.; Ma, Z.M.; Liu, T.Y. Asynchronous stochastic gradient descent with delay compensation. In Proceedings of the International Conference on Machine Learning, Sydney, NSW, Australia, 6 August 2017; pp. 4120–4129.
7. Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22 October 2017; pp. 2736–2744.
8. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-guided context feature pyramid network for object detection. arXiv 2020. preprint.
9. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06),Washington, DC, USA, 20 August 2006; Volume 3, pp. 850–855.