ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

Improved YOLOv7-based algorithm for elevator passenger detection

Juesi Xiao

College of Computer and Cyber Security, Fujian Normal University, Fuzhou, 350108, China

Abstract: *The task of identifying the number of passengers in an elevator plays an important role in optimizing the elevator algorithm and ensuring the safety of elevator passengers. In order to enable the elevator passenger-carrying algorithm to be deployed in edge devices, a lightweight elevator passenger-carrying algorithm based on improved yolov7 is designed, which reduces the model size while ensuring high precision. The proposed algorithm adopts a more lightweight attention mechanism structure in the head of the network. And made the lift-person-detection dataset as the experimental dataset. The dataset has nearly 2,000 image samples, including 900 training set images and 600 test set images. Experiments show that the proposed model achieves 98.9% mAP recognition accuracy. Compared with the 71.21MB model size of the original yolov7 network, the model size of the model proposed in this paper is reduced to 63.24MB, and the volume is reduced by 11.13%.*

Keywords: *YOLOv7 network; Attention mechanism; Elevator passenger detection*

1. Introduction

A real-time elevator headcount detection algorithm can provide an important basis for judging the stopping criteria of lifts. This makes the detection of elevator passenger numbers a key research topic. However, the number of lift passenger pictures is complex, the flow of people is dense and there are many types of targets, so how to accurately identify the number of lift passengers is an important challenge for target detection.

With the gradual popularization of artificial intelligence technology and the development of imaging technology, target detection using convolutional neural networks has become the mainstream direction. The current target detection algorithms are divided into single-stage detection algorithms with YOLO (You Only Look Once) and SSD (Single Shot MultiBox Detector) as the mainstream and dual-stage detection algorithms with R-CNN (region-basd CNN), while single-target detection algorithms of the regression class have low accuracy but high speed, and dual-stage The classification detection algorithm proposes to fix on the steps of target classification and regression to accomplish higher accuracy based on the idea of candidate region identification but relatively low execution efficiency and cannot provide a real-time detection means. (This part is replaced with representative works.) In 2019, Chen Jiuhong [1] et al. greatly improved the recognition ability of the target detection algorithm for small targets by improving the R-FCN (Region-based Fully Convolutional Network) network, with a large increase in accuracy. In areas such as intelligent lift systems, in 2021 Dan Tension [2] improved the feature fusion module in the YOLOv3 network and proposed an adaptive spatial feature fusion module-based statistical calculation method for the number of people waiting on the platform optimized the detection capability of the network for small targets. 2014 Xun Zhou et al [3] proposed a HOG feature-based method for identifying the number of people in a statistical lift with high in 2018, Jin Xiaolei et al [4] designed an ARM-based system for detecting the number of people inside a lift car, which improved the accuracy of identifying the number of people in lifts.

Widely paid attention by researchers, lift passenger number detection, as an important module for its front-end data collection, can quickly and accurately identify the number of lift passengers can effectively give an important theoretical basis for the basis of lift stopping and whether the capacity is overloaded. If deployed in edge devices, a more lightweight model is required. It is therefore of profound importance to design lift passenger count detection models with excellent real-time and robustness. This paper proposes an improved YOLOv7 based recognition algorithm model that can effectively reduce the size of the model while ensuring accuracy.

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

2. YOLOv7 network structure [5]

The YOLOv7 network uses an anchor-based method, while adding the most novel E-ELAN layer to the network structure, adding the REP layer for easy deployment in subsequent environments, and using the Aux_detect structure as an aid for detection when training in the head. The network is mainly composed of three parts: input, backbone and head, which are described below. [5]

Figure 1 shows the network structure of YOLOv7.

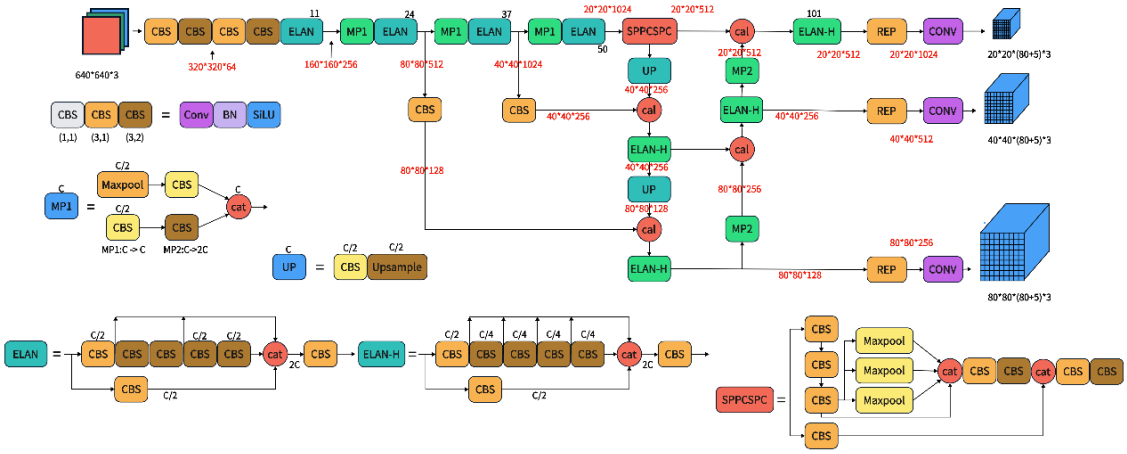


Figure 1: YOLOv7 network structure

2.1. Backbone layer

The Backbone layer consists of a total of 51 layers. First, the input feature map is changed to a size of 640 * 640 * 3 through four layers of CBS convolution layers, and then through an ELAN module, and finally through three MP and ELAN modules alternately to get the final Output, corresponding to C3/C4/C5, the size is 320*320*64, 160*160*256, 80*80*512 respectively.

Figure 2 shows the Backbone model structure.

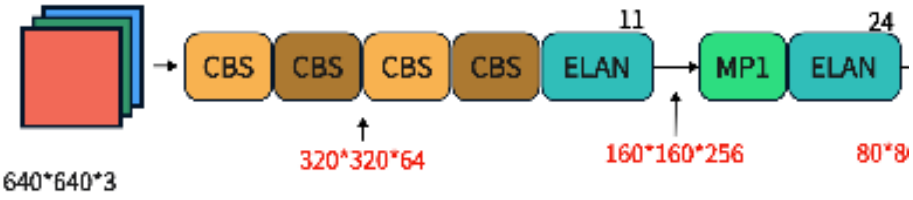


Figure 2: Backbone model structure

2.1.1. CBS model

The CBS module is composed of three different models, consisting of a convolutional layer, a Batch normalization layer, and a Silu layer (composed of an activation function). The CBS module has three different step size settings depending on the color The first CBS structure uses a 1x1 convolution with a stride of 1 to vary the number of channels. The second CBS structure uses 3x3 convolution with stride 1 for feature extraction. The third CBS structure uses a 3x3 convolution with a stride of 2, which is mainly used for feature sampling.

Figure 3 shows the CBS model structure.

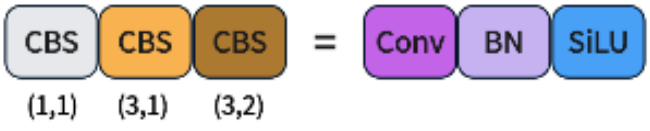


Figure 3: CBS model structure

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

2.1.2. ELAN model

The ELAN module is an efficient network structure composed of multiple CBS modules in parallel. By controlling the length of the gradient path, the network can learn a variety of features, thereby improving the robustness. It consists of two parallel branches, one through 1x1 convolution to complete the change of the number of channels, the other branch first through the 1x1 convolution module to change the number of channels, and then through four 3x3 convolution modules to change the number of channels After the feature extraction operation is completed, the feature extraction result can be obtained by superimposing and summing the four features.

Figure 4 shows the ELAN model structure.

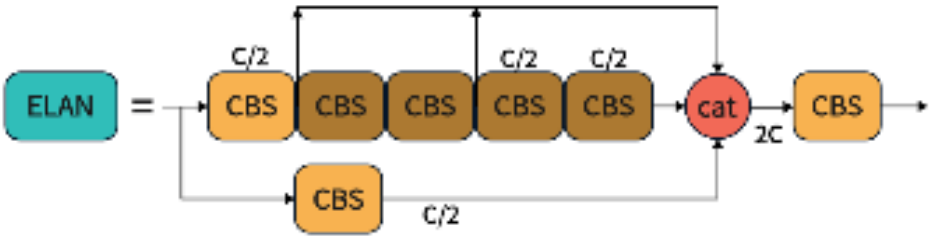


Figure 4: ELAN model structure

2.1.3. MP model

The MP module is also designed by adopting the final merging of the two branches, and its function is to perform downsampling.

The first branch first goes through a maxpool layer for downsampling, and then goes through a 1 x 1 convolutional layer to change the number of channels.

The second branch first goes through a 1x1 convolution to change the number of channels, and then goes through a convolution block with a 3x3 convolution kernel and a stride of 2, and performs the same downsampling operation. Finally, we superimpose the results of the two branches together, which is the final output of the module.

Figure 5 shows the MP model structure.

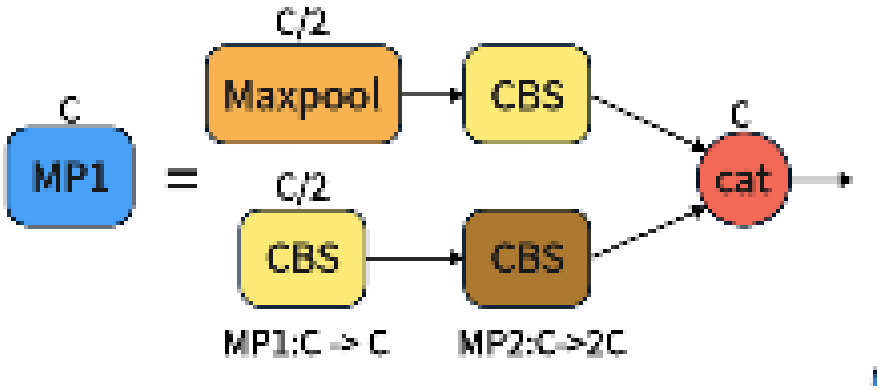


Figure 5: MP model structure

2.2. Head layer

The Head layer is similar to the pafpn structure. First, the C5 dimension reduction feature map transmitted from the backbone layer is passed through the SPPCSP module to increase the value of the receptive field, and the number of channels is changed from 1024 to 512. Then through top down and C4, C3 fusion, get P3, P4 and P5; and then through the UP layer to fuse it in P4, P5.

2.2.1. ELAN-H model

This module is similar to the ELAN module, except that the selection of the number of outputs of the second branch is different. The ELAN module uses three outputs for accumulation, while the

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

ELAN-H module uses five outputs for accumulation.

Figure 6 shows the ELAN-H model structure.

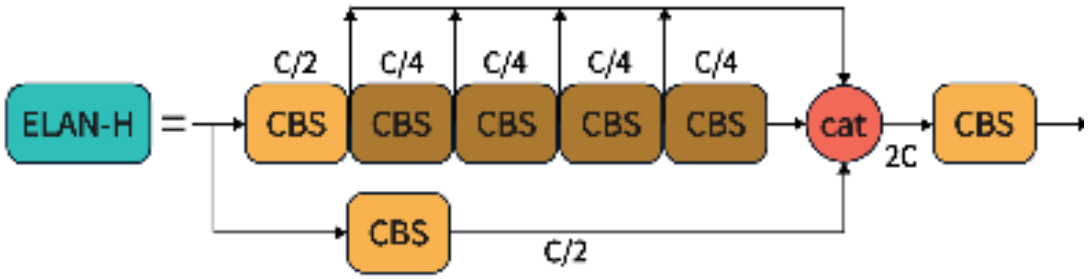


Figure 6: ELAN-H model structure

2.2.2. UP model

The UP module is performing an upsampling operation: using nearest neighbour interpolation.

Figure 7 shows the UP model structure.

Figure 7: UP model structure

2.2.3. SPPCSPC model

The module is divided into two parts: SPP structure and CSP structure. The function of SPP structure is to improve the value of the receptive field, make the algorithm adapt to different resolution images, and obtain different receptive fields through maximum pooling. First, after three CBS layers, and then in the first branch, different weight vectors are set by maxpool to obtain different objects, and then it is accumulated with the three objects previously processed by maxpool to obtain four different scales of receptive fields. This distinguishes the size of the target. The function of the CSP structure is to first divide the feature into two parts, one part performs conventional processing operations, the other part processes the SPP structure, and finally the two parts of the structure are accumulated together as the output of SPPSPC. This module reduces the amount of computation in half, making it faster and more accurate.

Figure 8 shows the SPPCSPC model structure.

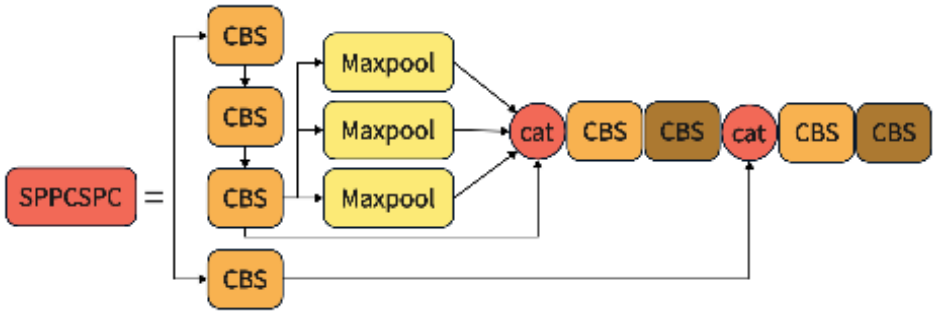


Figure 8: SPPCSPC model structure

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

3. Improvement of elevator passenger detection algorithm based on YOLOv7

3.1. Attention Mechanism

The attention mechanism is more commonly used in computer vision and natural language processing, and is an effective scheme that can improve the weight attributes of feature values [6-9]. The attention mechanism mimics a way in which the human brain processes information, i.e. the brain focuses more on the most important panels while analysing data and temporarily ignores the less important ones, and by doing so increases the usage of the weight vectors of the important panels. As a result, more detailed information about the target to be segmented is obtained, and a greater improvement in accuracy can be obtained.

The attention mechanism has been proposed in the 1990s, and the Google mind team proposed the fusion of RNN recurrent neural network and attention mechanism to achieve a significant improvement in the field of image classification [10]. Meanwhile, Bahdanau et al. used the attention mechanism in 2016 to achieve greater success in the field of natural language processing [11], while the attention mechanism has been applied to a variety of fields, still achieving greater success, and people have been exploring to overcome the limitations of convolutional neural network residuals on the grounds of improving efficiency. 2017 Vaswani et al. published "Attention is all you need" a comprehensive introduction to the transformer model with self-attention as the source model [12], pushing the use of the transformer attention mechanism to a new climax. In recent years, the Google team has used the attention mechanism again, using the transformer model and the self-attention mechanism instead of the traditional Seq2Seq and LSTM model architectures, respectively, to achieve success on translation tasks. The use of attention mechanisms instead of traditional RNN frameworks has now become mainstream in the field of deep learning [13].

3.2. GAMAttention

The GAMA attention mechanism [14] can reduce information dispersion while amplifying the interaction features of the global dimension. We adopted the sequential channel space attention mechanism from CBAM and redesigned the submodule [15].

The model is given an input feature mapping, an intermediate state and an output defined as:

$$F_2 = M_c(F_1) \otimes F_1 \tag{1}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{2}$$

Figure 9 shows the overall GAM network structure, where M_c and M_s correspond to Figure 10 channel attention and Figure 11 spatial attention map respectively; $\otimes$ represents the multiplication operation by element。

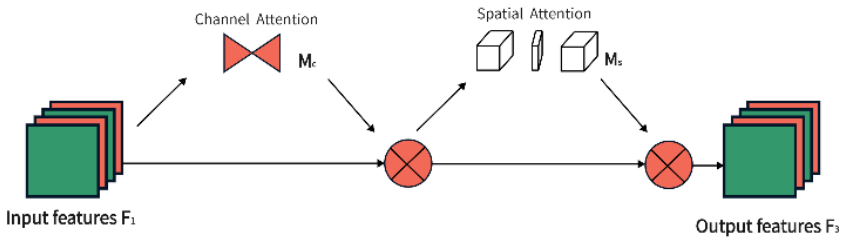


Figure 9: GAM model structure

3.2.1. Channel attention sub-module

The channel attention sub-module uses a three-dimensional arrangement in order to preserve spatial information.It then uses a two laye MLP (Multi-Layer Perceptron) to further amplify the connection between the cross-dimensional channels and space.

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

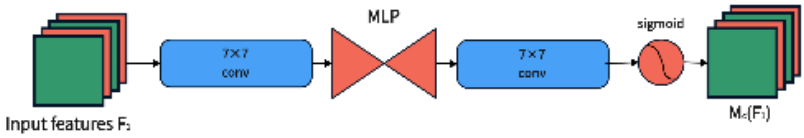


Figure 10: Channel attention sub-module

3.2.2. Spatial attention sub-module

In the spatial attention sub-module, two CNN convolutional layers are used for the fusion of spatial information in order to preserve the 3D information. The same parsimony ratio r is used for the channel attention sub-module, and to reduce the reduction of information due to the operation of the maximum pooling layer. We remove the pooling operation and retain its feature mapping. In addition to prevent a significant increase in model parameters, ResNet50 was used for the convolution operation [16].

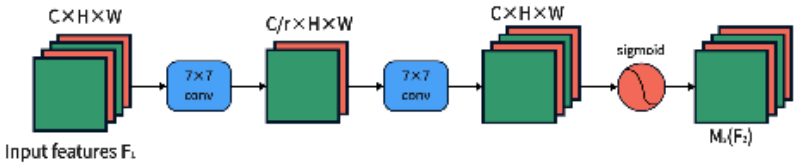


Figure 11: Spatial attention sub-module

4. Experimental procedure and analysis

4.1. Experimental dataset

In this article, we uses a home-made EPD (elevator-people-detection) dataset, consisting of 999 training sets and 603 test sets, satisfying a normal data training ratio of 3:2. Using the deep learning annotation tool labelimg, a dataset representing the elevator-people-detection sample was obtained by manual attainment.

4.2. Experimental environment and evaluation indicators

4.2.1. Experimental environment

The hardware configuration of this experiment is: CPU: 6-core Intel (R) Xeon (R) CPU E5-2650 v4 @ 2.20GHz;

GPU:GTX 1080 Ti * 1； Memory: 15GB; Video memory: 11GB The software environment is ubuntu18.04 Linux operating system, python 3.8, Miniconda conda3, Cuda 10.2.

Table 1 and Table 2 shows the specific experimental environment

Table 1: Software environment

Mirroring	Miniconda conda3
Python	3.8(ubuntu18.04)
Cuda	10.2

Table 2: Hardware environment

GPU	GTX 1080 Ti
CPU	6-core Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
VRAM	11GB
RAM	15GB

4.2.2. Evaluation indicators

In this experiment, we uses recall, precision, mAP (mean average precision), and AP (average precision) of each category to evaluate the performance of the optimized model.

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

The Definition as follows.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

$$precision = \frac{TP}{TP + FP} \tag{4}$$

$$AP = \sum_{k=0}^{N} P(k)\Delta r(k) \tag{5}$$

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP_i \tag{6}$$

Where $\boldsymbol{P}$ and Δr is the P-R (precision-recall) curve, TP (True Positives) is the number of positive samples accurately identified, FN is the number of positive samples not accurately identified, and FP is the number of negative samples not accurately identified [17-22]. The above metrics can be used to accurately judge the strengths and weaknesses of the model network.

To verify the effectiveness of this experiment in improving detection accuracy, we set up ablation experiments to ensure the uniqueness of the variables, and all experiments were tested in the same experimental environment to ensure that no external variables had an impact on the experimental results.

(a)Improved model recognition results *(b)YOLOv7 recognition results*

Figure 12: Network recognition results

According to the statistics in Table 3, we can see that the accuracy of recognition remains high using the optimised model, with the same mAP value of 0.989 as the original YOLOv7 network, but the size of the model has been effectively trimmed. The size of the model was effectively reduced from 71.21MB to 63.24MB, a reduction of 11.13%. As we can see from Figure 12, the network model in this paper can accurately identify passengers in the lift, effectively locate and frame passengers, and optimise the network structure spatially.

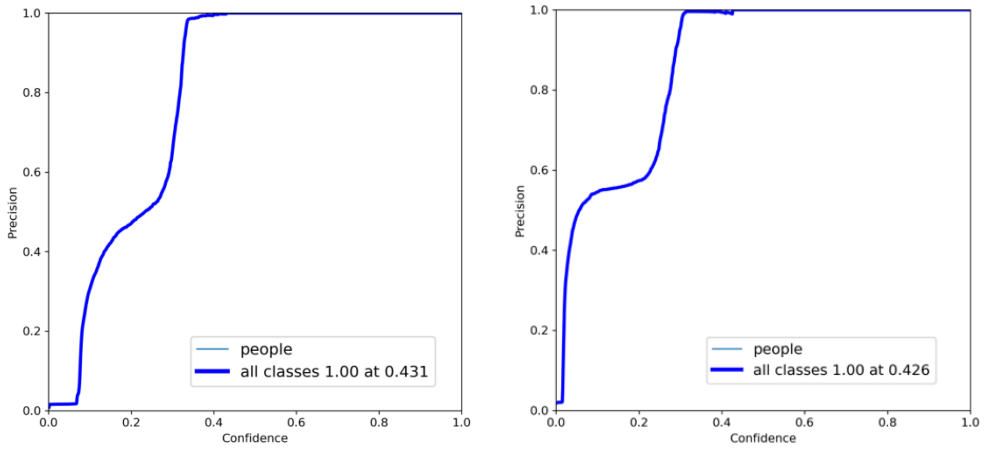
Table 3: Impact of improved modules on network performance on the EPD dataset

Network Model	mAP	Image size	model size
yolov7	0.989	(1280, 720)	63.24MB
yolov7-improvwd	0.989	(1280, 720)	71.21MB

Figure 13 shows the relationship between network precision and confidence and Figure 14 shows the relationship between network precision and recall. From the graphs, we can easily see that the

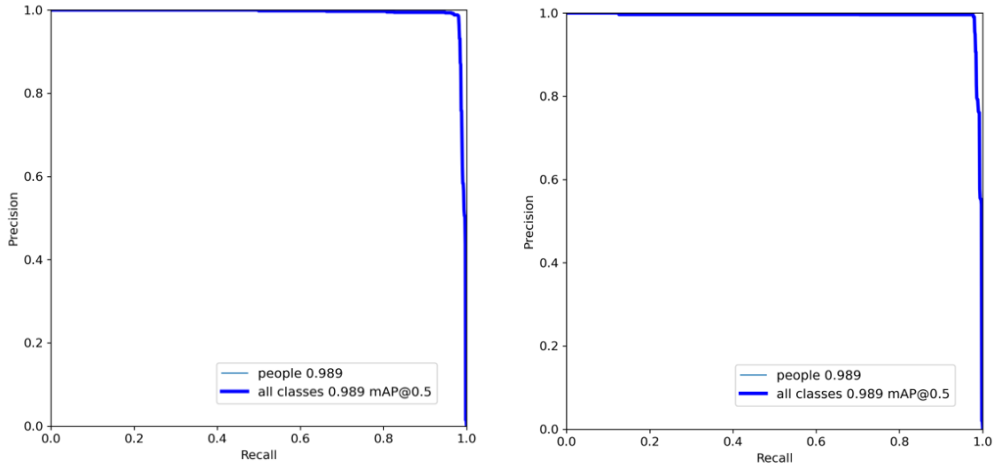
ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

improved GAM attention mechanism network has better results than the YOLOv7 network.



(a) *(b)*

Figure 13: P_cruve



(a) *(b)*

Figure 14: PR_cruve

5. Conclusion

This paper proposes an improved YOLOv7 network with a lighter GAMAttention module based on the C3C2 GAMAttention module to achieve a relatively small model size while maintaining high accuracy. The network achieves a more lightweight network model and improves the detection algorithm of YOLOv7 network in the field of lift passenger carrying.

The contributions of this paper are as follows: the proposed home-grown elevator-people-detection EPD dataset (elevator-people-detection dataset) consists of 999 training sets and 603 test sets, satisfying a normal data training ratio of 3:2. Manual attainment completion using the deep learning annotation tool labelimg compensates for the lack of datasets in the elevator-people-detection domain.

In this article we used the latest YOLOv7 network structure, replacing the entire family attention mechanism structure with a lighter one in the network part to ensure that the network can focus on the target features while also further reducing the size of the network in order to ensure the accuracy of the network. 11.13%. This effectively reduces the size of the model and makes it lighter.

References

[1] CHEN Jiuhong, ZHANG Haiyu. Design of Classroom Number Counting System Based on Deep Learning [J]. Software Guide, 2019, 18(10):27-29+35.

ISSN 2616-5775 Vol. 6, Issue 2: 85-93, DOI: 10.25236/AJCIS.2023.060211

[2] Zhang Lidan. Research on Detection Algorithm of Crowding Degree in Bus Carriage and Passenger Count in Platform [D]. Chongqing University.DOI:10.27670/d.cnki.gcqdu.2021.002625.
[3] Zhou Xun, Tao Qingchuan. Research on HOG-based lift headcount counting method[J]. Modern Computer (Professional Edition),2014(03):42-45.
[4] JIN Xiaolei[1], FAN Minghui[1], PAN Peng[2] System of Counting for the Number of People in the Elevator Based on ARM [J]. China Academic Journal Electronic PublishingHouse, 2018, 33(04): 30-33+62. DOI:10.19557/j.cnki.1001-9944.2018.04.007.
[5] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.
[6] HU J, SHEN L, SAMUEL A, et al. Gather-excite: exploiting feature context in convolutional neural networks [EB/OL]. [2021-05-25].
[7] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C] //Proc.of the European Conference on Computer Vision, 2018: 3-19.
[8] LIU F J, TIA H J. Dual attention network for scene segmentation[C] //Proc.of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.
[9] HU J, SHEN L, SUN G, et al. Squeeze-and-excitation networks[C] //Proc.of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.
[10] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems.Cambridge: MIT Press, 2014, 2: 2204-2212.
[11] BAHDANAU D,CHO K,BENGIO Y.Neural machine translation by jointly learning to align and translate[EB/OL].[2016-03-19].https://arxiv.org/pdf/1409.0473.pdf.
[12] VASWANI A, SHAZEER N, PARMAR N, et al.Attention is all you need[EB/OL]. [2017-12-06]. https: //arxiv.org/pdf/1706.03762.pdf.
[13] REN Huan, WANG Xuguang. Review of attention mechanism [J]. Journal of Computer Applications, 2021, Vol. 41 (201): 1-6
[14] Liu Y, Shao Z, Hoffmann N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions[J]. arXiv preprint arXiv:2112.05561, 2021.
[15] Sanghyun Woo, Jongchan Park, Joon-Y oung Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages.3–19, 2018.
[16] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6848–6856, 2018.
[17] Zhang Z, Zheng Y, Hong X, et al. A novel elevator group control algorithm based on binocular-cameras corridor passenger detection and tracking [J]. Multimedia Tools and Applications, 2015, 74(6):1761-1775.
[18] LIU R, LEHMAN J, MOLINO P, et al. An intriguing failing of convolutional neural networks and the coordconv solution. [EB/ OL]. https:/ / arxiv. org/ abs/1807. 03247:arXiv, (2018-12-03), [2021-09-13]
[19] WANG Y, ZHEN P B, HOU J H, et al. Convolutional neural networks with dynamic regularization [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 2299-2304
[20] LIU Z H, ZHANG Y D, CHEN Y Z, et al. Detection of algorithmically generated domain names using the recurrent convolutional neural network with spatial pyramid pooling [J]. Entropy, 2020, 22(9): 1-20
[21] YUN S, HAN D, OH J S, et al. CutMix: regularization strategy to train strong classifiers with localizable features [C] //2019 IEEE/ CVF International Conference on Computer Vision (ICC-V), Piscataway, USA, 2019: 6022-6031
[22] Mikulovich V I. Traffic Lights Detection and Recognition Method Based on the Improved YOLOv4 Algorithm [J]. Sensors, 2021, 22.