Research status of damage identification algorithm based on deep learning

Zhu Denghui[1], Song Lizhong[1], Feng yuan[1], and Yang Quanshun[1]

[1]School of electrical engineering, Naval University of engineering, 430033Wuhan, China

Abstract. One of the core tasks of computer vision is target detection. With the rapid development of deep learning, target detection technology based on deep learning has become the mainstream algorithm in this field. As one of the main application fields, damage identification has achieved important development in the past decade. This paper systematically summarizes the research progress of damage identification algorithm based on deep learning, analyzes the advantages and disadvantages of each algorithm and its detection results on voc2007, voc2012 and coco data sets. Finally, the main contents of this paper are summarized, and the research prospect of deep learning based damage identification algorithm is prospect.

1 Introduction

Damage identification is to locate and identify the damaged target in the image. However, due to the influence of background noise, target morphological diversity, target occlusion, illumination intensity and image resolution, the successful completion of the target detection task is a great challenge [1]. In order to overcome the difficulties in the current target detection task, many excellent scholars have devoted themselves to the research in this field.

The traditional target recognition algorithm is mainly divided into four stages: 1) selecting candidate regions and generating candidate frames on the target image through a fixed size sliding window; 2) feature extraction using SIFT (scale invariant feature transform) [2-3], hog (histogram of oriented gradient) [4-5] and other traditional feature extraction methods to extract features of candidate regions; 3) feature classification, using DPM (deformable part model) [6], AdaBoost [7], SVM (support vector machine) [8] and other classifiers are used to classify features [9]; 4) NMS)(non maximum suppression)[10] and other methods are used to modify and optimize the results. However, the traditional target detection algorithm does not select the sliding window according to the characteristics of the target, resulting in a large amount of time cost.; and the manually designed feature extraction method is only good for the recognition task with small changes in the background and light, and is not suitable for the target recognition task with large changes in background and light; the traditional classification algorithm is not suitable for the target recognition task with large changes in background and light It is easy to be disturbed by noise and has poor robustness, so it is difficult to meet the requirements of social target detection.

In 2012, the alexanet [11] model proposed by Hinton et al In (ImageNet Large Scale Visual Recognition Competition)competition, it won the championship by more than 10% of the second place. Since then, deep learning has been paid more and more attention in the field of computer vision [12], and has promoted the rapid development of computer vision, and various improved algorithm models have been endless. Deep learning transforms input data into higher-level features through nonlinear model, which has strong learning ability and generalization ability. The research of deep learning also promotes the development of object detection, and object detection based on deep learning has become one of the key research directions in the field of computer vision. According to the different implementation methods, the damage target detection algorithm based on deep learning can be divided into two schools: two stage detection and one stage detection, which correspond to the damage target detection algorithm based on candidate region classification and the damage target detection algorithm based on regression. In this paper, we will systematically introduce the research progress of these two kinds of damage target detection algorithms, analyze the related algorithms systematically, and summarize and prospect the research on damage target detection based on deep learning.

2 DAMAGE TARGET DETECTION ALGORITHM BASED ON CANDIDATE REGION CLASSIFICATION

Firstly, the candidate region is selected from the input image, and then the feature extraction and position calibration of the candidate region are carried out by using the convolutional neural network. The main

representatives are R-CNN,SPPNet,Fast R-CNN,Faster R-CNN,R-FCN,Mask R-CNN,IOU-NET,and D2Det.

2.1 R-CNN

Ross 2014 Girshick et al. Proposed R-CNN model [13]. In order to solve the problem of window redundancy, the model uses selective search algorithm [14] to replace sliding window, and then scales the candidate region into a fixed size image, then extracts features by convolutional neural network, classifies the target class information of candidate area by SVM classifier, and sends it to full connection layer network for position calibration. The algorithm uses convolutional neural network to replace the traditional manual design of feature extraction part, which can extract image features more effectively. However, because R-CNN needs to extract features from each candidate region, the cost of detection is high. In addition, the scaling process of candidate regions will lead to the loss of some useful information in the image, which will affect the final detection results.

2.2 SPP-Net

In order to solve the defects of R-CNN network, he et al. Proposed spp net network [15] in 2015. This network only runs convolution layer once for the whole image to obtain feature map, which avoids repeated feature extraction operation in R-CNN network and greatly saves operation time. Then, fixed length feature vector is extracted from feature map by spp layer. Compared with R-CNN, the detection speed of the algorithm is 24-64 times faster than that of R-CNN; at the same time, due to the existence of SPP layer, there is no need to zoom the image, which reduces the loss of image information. But in spp net, each training step is still independent and the training cycle is long.

2.3 Fast R-CNN

In 2015, girshick et al. Proposed fast R-CNN model [16], improved spp to ROI (region of interest) pooling layer, and introduced multi task learning mode to unify multiple steps into a convolutional neural network, which greatly improved detection speed and accuracy.

Compared with spp net, fast R-CNN has significantly improved the speed and accuracy, but the algorithm still uses the selection search algorithm when extracting candidate frames, and the detection speed still can not meet the real-time requirements.

2.4 Faster R-CNN

In view of the waste of time caused by the use of search algorithm to select candidate regions, Ren et al. Proposed the faster R-CNN [17] model in 2016. The model combines candidate region feature selection, target frame fine-tuning, classification and position regression, and realizes the end-to-end training completely.The number of candidate frames is reduced by nearly nine tenths by using hand-designed anchor frames. In this way, the quality of candidate frames is improved, and the speed and accuracy of damage target detection have also been improved. However, fast R-CNN will lose a lot of detail information after multiple down sampling, which leads to the general detection effect of the model for small targets.

2.5 R-FCN

In 2016, in order to solve the problems in fast R-CNN, Dai et al. Proposed r-fcn [18]. Fast In R-CNN, the full connection layer is replaced by convolutional neural network for classification and regression, which greatly reduces the parameters: at the same time, Dai et al. Constructed a set of position sensitive score maps to overcome the translation sensitivity problem in damage target detection. The experiment shows that the network can achieve 79.5% map on the voc2007 data set, but the network still needs a lot of computation and detection speed It can not meet the real-time requirements.

2.6 Mask R-CNN

In 2017, he et al. Proposed Mask R-CNN [19] on the basis of fast R-CNN, and added mask prediction branch in Mask R-CNN network for instance segmentation. In addition, in terms of feature extraction, the architecture of feature pyramid network (FPN) [20] was adopted, and ROI align pooling layer was used instead of ROI pooling layer. The network improves the detection accuracy and the recognition of small objects is more accurate, but the detection speed is still not significantly improved.

2.7 TridentNet

In 2019, in view of the traditional multi-scale detection problem, Li et al. Proposed a three branch network tridentnet [21]. In this network, the influence of different receptive fields on the detection results was verified for the first time, and the conclusion was that the detection of large objects by large receptive fields is more accurate, while that of small receptive fields is more accurate. Tridentnet uses RESNET as the basic network. There is no change in the first three stages. In the fourth stage, the receptive field network will be parallelized. The three parallel branches use hole convolution with different number of holes. The receptive field corresponds to the detection of small, medium and large targets from small to large. Compared with the previous algorithm, the network solves the problem of multi-scale detection better, and at the same time, three branches share the weight, which effectively reduces the risk of parameter over fitting. However, with the increase of network structure, the amount of computation is still very large and cannot be detected in real time.

2.8 D2Det

In 2020, Cao et al. Proposed a new two-stage detection algorithm d2det [22]. Its main advantage is that it can

solve the problems of accurate positioning and accurate classification at the same time. In this model, a dense local regression is introduced, which can predict multiple dense frame offsets of an object, which is different from the traditional regression location and key point based location used in traditional two-stage detector. It is not limited to a set of quantized key points in the region, and can regress the position sensitive real dense offset, so as to achieve more accurate positioning In order to reduce the influence of the background area on the target, the binary overlap prediction strategy is applied to improve the precision of dense local regression; the ROI pooling layer for discrimination is introduced in the model, and the samples extracted from each sub region are weighted adaptively to obtain the discriminant features, which greatly improves the performance.

2.9 Summary

Table 1 shows the relevant evaluation indexes and their meanings, table 2 shows the performance comparison of some damage target detection algorithms based on candidate regions, and "-" indicates that there is no relevant data. As can be seen from table 2, with the continuous improvement of the algorithm, the accuracy of the damage target detection algorithm based on candidate region classification is continuously improved, but the detection speed does not improve correspondingly. In order to improve the detection speed, target based detection algorithm has been widely concerned.

Table 1. Main evaluation indicators of COCO data.

Andicators	Significance
AP_{50}	AP value when IOU = 0.5
AP_{75}	AP value when IOU = 0.75
APs	AP value of small objects
AP_M	AP value of mesium objects
AP_L	AP value of large objects

Table 2. Performance comparison of some damage target detection algorithms based on candidate regions

Model	Backbone network	Detection speed/(f.s⁻)	VOC2007(mAP@IOU=0.5)	VOC2012(mAP@IOU=0.5)	COCO(mAP@IOU=0.5:0.05:0.95)	AP_{50}	AP_{75}	A_S	A_M	A_L
R-CNN	AlexNet	0.03	58.5	-	-	-	-	-	-	-
SPP-Net	ZF-5	2.0	59.2	-	-	-	-	-	-	-
Fast R-CNN	VGG-16	3.0	70	68.4	19.7	35.9	-	-	-	-
Faster R-CNN	ResNet-101	5.0	76.4	73.8	34.9	59.1	39.0	18.2	39.0	48.2
R-FCN	ResNet	6.0	79.5	77.6	29.9	51.9	-	10.8	32.8	45.0
Mask R-CNN	ResNet-101	11.0	-	-	39.8	62.3	43.4	22.1	43.2	51.2
TridentNet	ResNet-101	0.7	-	-	48.4	69.7	35.5	31.8	51.3	60.3
D2Det	ResNet-101	-	-	-	50.1	69.4	54.9	32.7	52.7	62.1

3 DAMAGE TARGET DETECTION ALGORITHM BASED ON REGRESSION

3.1 Yolo series

In 2016, Redmon et al. Proposed yolov1 [23], The network overcomes the problem that feature extraction in two-stage damage target detection needs a lot of time. The whole image is divided into grid by grid prediction, and then the grid with the center of the target is taken as the prediction grid. In this way, the complex operation of generating candidate regions is avoided, and the detection speed is significantly improved. However, since the detection area is the whole image, it will increase Adding the number of background classes reduces the detection accuracy, especially when there are multiple target centers in a grid at the same time, it will cause positioning error, and only one target will be detected.

In 2017, Redmon and others proposed yolov2 [24] on the basis of yolov1, using the batch normalization (BP) [25] operation, through which the input of each layer of network can be guaranteed to obey the same distribution, It replaces the dropout method [26]; on the other hand, it uses the high-resolution images as the training set, and uses the anchor method in fast R-CNN for prediction.

In 2018, Redmon and others proposed an improved version of yolov3 [27]. This model takes darknet-53 as the network backbone. Compared with resnet-101, darknet-53 can achieve considerable accuracy, but its speed is greatly accelerated. The model consists of 53 convolution layers, and can be detected on three different scale feature graphs by using feature pyramid structure, which effectively improves the detection effect of neural network for small targets.

In 2019, Choi et al. Improved the y0l0v3 network and proposed Guassian yolov3 [28], adding the dimension of yolov3 boundary box coordinate output, which can output eight dimensions of coordinate information, and optimize the network loss function. Finally, when the network is detected on the Kitti dataset, the accuracy rate is improved by 3% compared with the yolov3 model.

In 2020, bochkovskiy et al. Proposed yolov4 [29]. In order to increase the receptive field, spp module was used for reference in csparknet53 backbone network, and a variety of commonly used algorithms were combined. Such as self confrontation training (SAT), cross small batch Standardization (cmbn), cross phase partial connection (CSP) and dropblock regularization. Through these optimization techniques, the model achieves the highest accuracy of damage target detection, and realizes the perfect matching of detection accuracy and speed.

3.2 SSD series

In 2016, Liu et al. Proposed SSD [30], This model refers to yolov1 model, and uses different size prediction frames to detect targets in different feature layers,

predicts small-scale targets in shallow feature maps, and predicts large-scale targets in high-level feature maps by using prediction frames of different sizes, which makes better use of the characteristics of different feature layers with different information emphases and improves the performance of small-scale targets The detection effect of. However, because the features of each layer are input separately, the same target will be detected many times. Due to the lack of semantic information of shallow feature map, SSD still can not meet the requirements for small damage target detection.

In 2017, Cheng et al. Proposed dssd [31], whose biggest improvement is to introduce context information into damage target detection. The backbone network of dssd is replaced by resnet-101, and the residual module is introduced. At the same time, the context information is introduced by anti convolution layer. The model can improve the expression ability of the shallow feature map, and improve the detection accuracy of small targets. However, due to the increase of network layers, the detection speed is not as fast as SSD.

In order to reduce the number of parameters in SSD, Shen et al. Proposed dsod [32] referring to the idea of densenet. The model improves the input of some feature layers and can train data directly from zero without pre training model, and the detection effect of damage target is improved.

In 2017, Li et al. Proposed FSSD [33], which improved the FPN and constructed a network framework for feature fusion. Firstly, feature graphs of different levels were combined, and then new feature map combinations were generated. Finally, the results were obtained through the final prediction network. The model improves the detection accuracy of damaged targets and maintains a good detection speed.

3.3 RetinaNet

In 2017, Lin et al. Compared the one-stage detector with the two-stage detector, and found that the sample class imbalance of the first-stage detector in the training process greatly reduced its accuracy, so they proposed retinanet [34]. In this model, the standard cross entropy loss function is reshaped, and a new loss function named "focus" is introduced. The loss function redistributes the weight of samples to be classified, reduces the weight of easily classified samples, and increases the weight of difficult samples. In this way, more attention can be paid to difficult samples in training, so as to improve the classification accuracy. So that the detector will pay more attention to the difficult samples in the training process. The main feature of RESNET FPN is that it outputs two subnets in each layer of the feature pyramid, which are used for classification and anchor frame positioning and regression respectively. Experiments show that retinanet not only retains the advantage of one-stage detector in speed, but also surpasses the two-stage detector in detection accuracy.

3.4 EfficienDet

In order to solve the problem of fast detection speed and low detection accuracy of one-stage detection method, Tan et al. Proposed efficiendet [35] in 2019. The model takes efficiennet [36] as the backbone network, and bi-directional feature pyramid network (bifpn) is used for feature network, which can repeatedly perform bidirectional feature fusion. At the same time, the method of joint scaling is introduced, in which the idea of weighting is applied. The joint scaling can uniformly scale the depth, width and resolution of backbone network, feature network and frame class prediction network at the same time, so as to achieve the optimal effect. The parameters of eficiendet model combined with these methods are only 1 / 4 of the optimal model parameters at that time, and the detection speed is increased more than 3 times.

3.5 CentripetalNet

In 2020, Dong et al. Proposed centripetalnet [37]. The model can effectively solve the problem of key point matching error when detector is based on key point detection. It can predict the corner position and centripetal displacement of the target, so as to match the corresponding angle. This method can match the angle more accurately than the traditional embedding method. At the same time, a cross star deformable convolution network is proposed innovatively, which improves the feature adaptive ability of the network. After testing, the performance of the model is very good on coco data set. AP surpasses most of the existing non anchor frame detectors.

3.6 Summary

According to table 3, the detection speed of the damage target detection algorithm based on regression is significantly higher than the damage target detection algorithm based on candidate region classification introduced in the previous paper, and the detection accuracy is also constantly improved, and the gap with the latter is smaller. From the current research, the damage target detection algorithm based on regression has a broader development prospect.

Table 3. Performance comparison of some damage target detection algorithms based on regression.

Model	Backbone network	Detection speed/(f. s^{-})	VOC2007 (mAP@IOU=0.5)	VOC2012 (mAP@IOU=0.5)	COCO(mAP@IOU=0.5:0.05:0.95)	AP_{50}	AP_{75}	A_S	A_M	A_L
YOLO-v2	DarkNet-19	40.0	78. [illegible]	73. [illegible]	33.0	57.9	34.4	18.3	35.4	41.9

YOLO-v3	DarkNet-53	20.0	–	–	21.6	44.0	19.2	5.0	22.4	35.5
YOLO-v4	CSPDarkNet-53	31	–	–	43.0	64.9	46.5	24.3	46.1	55.2
SSD512	VGG-16	22.0	–	–	28.8	48.5	30.3	10.9	31.8	43.5
DSSD321	ResNeXt-101	9.5	–	–	28.0	46.1	29.2	7.4	28.1	47.6
FSSD512	VGG-16	–	–	–	31.8	52.8	33.5	14.2	35.1	45.0
RetinaNet500	ResNet-101	5.4	–	–	34.4	53.1	36.8	14.7	38.5	49.1
EfficienDet-D7	EfficienNet-B7	–	–	–	52.2	71.4	56.3	-	-	-
Centripetal Net	Hourglass-104	–	–	–	48.0	65.1	51.8	29.0	50.4	59.9

4 Summary and prospect

In this paper, we combined with the domestic and foreign development status and studied the development of deep learning based damage target detection algorithm. We introduced from the candidate region classification and regression based two aspects, and compared the advantages and disadvantages of each algorithm . In addition, We analyzed the main characteristics of each model: the algorithm based on candidate region classification can achieve high detection accuracy, but because of the excessive network parameters and the dependence on the generation of candidate regions, the algorithm can not meet the real-time requirements in the detection speed; and the detection algorithm based on regression effectively solves the problem of low detection speed, simplifies the damage target detection to the end-to-end training process, and improves the detection efficiency. Moreover, in the process of continuous optimization of the algorithm, this kind of algorithm detection accuracy The degree of detection is higher and higher, especially for small damage target detection.

Damage target detection has developed rapidly in recent years, and has made some achievements, but there are still many breakthroughs to be made: (1) at present, the model network model is more complex, which puts forward higher requirements on the performance of GPU and the size of data set. In order to reduce or avoid network redundancy, people should optimize the network structure and design a more compact and portable model [38]. (2) When the data set is insufficient, the performance of the damaged target detector will generally decline, and it is prone to under fitting. Therefore, the ability of the detector to learn from a small number of samples should be improved. (3) To realize the automation of damage target detection and optimize the automatic design of backbone network, automatic neural architecture search depth learning method can be used to replace manual feature calibration (4) To realize the intelligent detection of damage targets, explore and design a new network model, so that the detector can intelligently learn new objects, locate and identify new object categories that have not been learned before .

References

1. G. P. Sun , X. Y. Hao, Z. J. Zhang, et al. Cooperative target recognition method based on multi feature judgment. JSS, 2377-2383, **30**, 6(2018)
2. D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 91-110, **60**, 2(2004)
3. L. Juan, O. Gwun. A comparison of sift, pca-sift and surf. IJIP, 143-152, **3**, 4(2009)
4. N. Dalal, B. Triggs. Histograms of oriented gradients for human detection. IEEE, 886-893, 1(2005)
5. Q. Zhu, M. C. Yeh, K. T. Cheng, et al. Fast human detection using a cascade of histograms of oriented gradients. IEEE, 1491-1498, 2(2006)
6. P. Felzenszwalb, D. McAllester, D. Ramanan. A discriminatively trained, multiscale, deformable part model. IEEE, 1-8(2008)
7. Y. Freund, R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. JCSS,119-139,**55**, 1(1997)
8. J. A. K. Suykens, J. Vandewalle. Least squares support vector machine classifiers. NPL, 293-300, **9**, 3(1999)
9. B. X. Xu, J. Gong, Z. X. Sun. A review of target detection models based on convolutional neural networks. CTD, 1-8, 11(2019)
10. A. Neubeck, L. Van. Efficient non-maximum suppression. IEEE, 850-855, 3(2006)
11. A.Krizhevsky, I. Sutskever, G. E. Hinton . Imagenet classification with deep convolutional neural networks. ACM, 84-90, **60**, 6(2017)
12. D. Dan. A review of target detection based on deep learning. ISTT, 1-2, **27**, 13(2019)
13. R. GIRSHICK, J. DONAHUE，T. DARRELL，et al. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE, 580-587, (2014)
14. J. R. UIJLING，K. E. VAN，T. GEVERS, et al. Selective search for object recognition. IJCV, 154-171, 104, 2(2013)
15. K. He, X. Zhang, S. Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE, 1904-1916, **37**, 9(2015)
16. R. Girshick. Fast R-CNN.IEEE, 1440-1448 (2015)
17. S Ren, K. He, R Girshick, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE, 1137-1149, **39**, 6(2016)
18. J. Dai, Y. Li, K. H, et al. R-fcn: Object detection via region-based fully convolutional networks. NIPS, 379-387(2016).

19. K. He, G. Gkioxari, P. Dollár, et al. Mask R-CNN. IEEE, 2961-2969(2017)
20. T. Y. Lin, P. Dollár, R. Girshick, et al. Feature pyramid networks for object detection. IEEE, 2117-2125(2017)
21. Y. Li, Y. Chen, N. Wang, et al. Scale-aware trident networks for object detection.IEEE, 6054-6063(2019)
22. J. Cao, H. Cholakkal, R. M. Anwer, et al. D2Det: Towards High Quality Object Detection and Instance Segmentation. IEEE/CVF, 11485-11494(2020)
23. J. Redmon, S. Divvala, R. Girshick, et al. You only look once: Unified, real-time object detection. IEEE, 779-788(2016)
24. J. Redmon, A. Farhadi. YOLO9000: better, faster, stronger. IEEE. 7263-7271(2017)
25. S. Ioffe, C Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv, 1502.03167(2015)
26. N. Srivastava, G. Hinton, A. Krizhevsky, et al. Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research, 1929-1958.**15** (2014)
27. J. Redmon, A. Farhadi. Yolov3: An incremental improvement. ArXiv.1804.02767(2018)
28. J. Choi, D. Chun, H. Kim, et al.Gaussian YOLOv3 ： an accurate and fast object detector using localization uncertainty for autonomous driving.ICCV. 502-511(2019)
29. A. Bochkovskiy, C. Y. Wang, H. Y. MLiao. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv.10934(2020)
30. W. Liu, D. Anguelov , D. Erhan , et al. Ssd: Single shot multibox detector. Cham, 21-37(2016)
31. C. Y. Fu, W. Liu, A. Ranga, et al. Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.(2017)
32. Z. Shen ， Z. Liu，J. Li，et al.DSOD：learning deeply super-vised object detectors from scratch//IEEE Interna-tional Conference on Computer Vision,1937-1945(2017)
33. Z. Li, F. Zhou. FSSD: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960(2017)
34. T .Y. Lin , P. Goyal, R. Girshick, et al. Focal loss for dense object detection. IEEE . 2980-2988.(2017)
35. M. Tan, R. Pang, Q .V.Le . Efficientdet: Scalable and efficient object detection. IEEE/CVF. 10781-10790(2020)
36. M. Tan, Q. V. Le . Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946(2019)
37. Z. Dong, G. Li, Y. Liao, et al. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. IEEE/CVF. 10519-10528(2020)
38. G. Chen, W. Choi, X. Yu, et al. Learning efficient object detection models with knowledge distillation. ANIPS. 742-751(2017)