A Lightweight Object Detection Network with Attention Modules

Subjects: Computer Science, Artificial Intelligence

Contributor: Shengying Yang, Linfeng Chen, Junxia Wang, Wuyin Jin, Yunxiang Yu

Object detection methods based on deep learning typically require devices with ample computing capabilities, which limits their deployment in restricted environments such as those with embedded devices.

Keywords: object detection ; embedded platform ; attention model ; feature pyramid

1. Introduction

AS one of the fundamental tasks in computer vision, object detection is widely used in face detection, object tracking, image segmentation, and autonomous driving [1]. The objective is to localize and classify specific objects in an image, accurately find all the objects of interest, and locate the position with a rectangular bounding box [2][3]. In recent years, in the field of computer vision, there has been a growing focus on designing deeper networks to extract valuable feature information, resulting in improved performance [4][5][6][7][8]. However, due to the vast number of parameters in these models, they often consume a significant amount of computational resources. For example, Khan [9] proposed an end-to-end scale-invariant head detection framework by modeling a set of specialized scale-specific convolutional neural networks with different receptive fields to handle scale variations. Wang [10] introduced a pyramid structure into the transformer framework, using a progressive shrinking strategy to control the scale of feature maps. While these models demonstrate outstanding detection accuracy, they heavily rely on powerful GPUs to achieve rapid detection speed [11]. This poses a significant challenge in achieving a balance between accuracy and inference speed on mobile devices with limited computational resources [12][13][14]. Currently, detection models based on deep learning often use complex network architectures to extract valuable feature information. Although such models have a high detection accuracy, they usually rely on powerful graphics processing units (GPUs) to achieve a fast detection speed [15]. With the rapid development of technologies such as smartphones, drones, and unmanned vehicles, implementing neural networks in parallel on devices with limited storage and computing power is becoming an urgent need. Under computing power and storage space constraints, lightweight real-time networks have become popular research topics related to the application of deep learning in embedded applications [16].

Recently, some researchers have reduced the number of parameters and model size of the network by optimizing the network structure, such as SqueezeNet, MobileNetv1-v3 [17][18][19], ShuffleNetv1-v2 [20][21], Xception [22], MixNet [23], EfficientNet [24], etc. The MobileNet series methods replace the traditional convolution by using depth-wise separable convolutions, thus achieving a result similar to that of standard convolution but greatly reducing the number of model calculations and parameters. The ShuffleNet series networks use group convolution to reduce the number of model parameters and apply channel shuffling to reorganize the feature maps generated by group convolution. Other researchers have proposed regression-based one-stage object detectors, such as SSD [25], YOLOv1-v4 [26][27][28][29], RetinaNet [30], MimicDet [31], etc. Instead of taking two shots, as in the RCNN series, one-stage detectors predict the target location and category information directly from a network without region propositions. Based on the regression concept of YOLOv1, SSD uses predefined boxes of different scales and aspect ratios for prediction and extracts feature maps of different scales for detection. Although the SSD accuracy is better than that of YOLOv1, SSD does not perform well in small object detection. YOLOv3 uses the Darknet backbone network to mine high-level semantic information, which greatly improves the classification performance. A similar feature pyramid network is used for feature fusion to enhance the accuracy of small target detection. Since a large number of easily classified negative samples in the training phase can lead to model degradation, RetinaNet proposes focal loss based on standard cross-entropy loss to eliminate category imbalance effectively, similar to a kind of hard sample mining. To improve the accuracy of the one-stage detector, MimicDet uses the features generated by a two-stage detector to train the one-stage detector in the training phase. However, in the inference phase.

MimicDet uses a one-stage method directly for prediction to ensure that the detection speed is relatively fast. The YOLO series methods achieve an excellent balance between accuracy and speed and have become widely used for target detection in actual scenarios. Nevertheless, YOLO models have a complex network structure and a large number of network parameters, so they require vast computing resources and considerable storage space when used in embedded devices. However, the high computational cost limits the ability of YOLO models to perform multiple tasks that require real-time performance on computationally limited platforms [32]. To reduce the occupation of computing resources, lightweight YOLO methods require fewer parameters and improve the detection speed by applying a smaller feature extraction network, such as the latest YOLOv4-Tiny [33]. Therefore, when performing object detection on embedded devices, improving the detection accuracy while achieving real-time performance is a significant problem to be solved.

2. Attention Mechanism

In recent years, attention mechanisms have been widely used in various fields of computer vision to enhance important features and suppress irrelevant noise. These mechanisms provide excellent performance in improving model accuracy, such as SENet [34], CBAM [35], No-local [36], SKNet [37], GCNet [38], NAM [39], ECANet [40], SA-Net [41], SimAM [42], GAM [43], etc. SENet explicitly models the correlation between feature channels and automatically learns the channel-wise weights for feature selection. CBAM focuses on spatial and channel attention information and concatenates the feature maps after average and maximum pooling operations to reduce feature loss, thus making the model focus on the target itself rather than the background. SKNet uses convolution kernels of different sizes to extract semantic features and dynamically adjusts the receptive field by aggregating feature information from multiple branches. Based on SENet and No-local, GCNet proposes a simple global context modeling framework to mine long-distance dependencies and reduce computational pressure. NAM applies a weight sparsity penalty to the attention module, thereby improving computational efficiency while maintaining similar performance. ECANet has mainly made some improvements to the SENet module, proposing a non-dimensional reduction local cross-channel interaction strategy and an adaptive method for selecting the size of one-dimensional convolutional kernels, thereby achieving performance improvement. Although CBAM brings performance improvements, it increases the computational complexity to a certain extent. SA-Net introduces the channel shuffle method, which parallelizes the use of spatial and channel attention mechanisms in blocks, enabling efficient integration of the two types of attention mechanisms. Different from the common channel and spatial attention modules, SimAM introduces an attention mechanism without any trainable parameters, proposed based on neuroscience theory and the linear separability principle. GAM proposes a global attention mechanism that introduces channel attention and multi-layer perceptrons to reduce information diffusion and amplify global interactive representations, thereby improving the performance of deep neural networks.

3. YOLOv4, YOLOv4-CSP and YOLOv4-Tiny Networks

YOLOv4 is an evolution from YOLOv3, and the purpose is to design a real-time object detection network that can be applied in the actual working environment. YOLOv4 proposes a CSPDarknet53 backbone network to reduce repeated gradient learning effectively and improve the learning ability of the network. In terms of data augmentation, YOLOv4 uses a mosaic to combine four images into one, which is equivalent to increasing the minibatch size and adds self-adversarial training (SAT), which allows the neural network to update images in the reverse direction before normal training. In addition, YOLOv4 uses modules such as ASFF [44], ASPP [45], and RFB [46] to expand the receptive field and introduce attention mechanisms to emphasize important features. Based on YOLOv4, YOLOv4-CSP is compressed in terms of network width, network depth, and image resolution to achieve the optimal trade-off between speed and accuracy. Compared with YOLOv4, YOLOv4-CSP converts the first CSP stage into the original Darknet residual layer and modifies the PAN architecture in YOLOv4 according to the CSP approach. Moreover, YOLOv4-CSP inserts an SPP module in the middle position of the modified PAN structure. To reduce the computational complexity for embedded devices, YOLOv4-Tiny is a simplified structure of YOLOv4 and YOLOv4-CSP. YOLOv4-Tiny uses a lightweight backbone network called CSPDarknet-Tiny while directly applying a feature pyramid network (FPN) [47] instead of a path aggregation network (PANet) [48] to reduce computational complexity. In the inference stage, multiscale feature maps are first fused via the FPN. Then, the category scores and offsets of each predefined anchor are predicted by a 1 × 1 convolution kernel, and the predicted bounding boxes are postprocessed using non-maximal suppression (NMS) [49] to obtain the final detection results. Although YOLOv4-Tiny provides a certain accuracy rate and fast detection speed, the regression accuracy for small and medium targets is relatively low, which will be improved in the proposed Mini-YOLOv4 network.

References

1. Chen, R.; Liu, Y.; Zhang, M.; Liu, S.; Yu, B.; Tai, Y.-W. Dive deeper into box for object detection. In Proceedings of the 2020 European Conference on Computer Vision (ECCV): 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 412–428.
2. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10186–10195.
3. Qiu, H.; Li, H.; Wu, Q.; Shi, H. Offset bin classification network for accurate object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 13185–13194.
4. Shi, H.; Zhou, Q.; Ni, Y.; Wu, X.; Latecki, L.J. DPNET: Dual-path network for efficient object detection with Lightweight Self-Attention. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 771–775.
5. Termritthikun, C.; Jamtsho, Y.; Ieamsaard, J.; Muneesawang, P.; Lee, I. EEEA-Net: An early exit evolutionary neural architecture search. Eng. Appl. Artif. Intell. 2021, 104, 104397.
6. Sun, Z.; Lin, M.; Sun, X.; Tan, Z.; Li, H.; Jin, R. MAE-DET: Revisiting maximum entropy principle in zero-shot nas for efficient object detection. In Proceedings of the 39th International Conference on Machine Learning (PMLR), Virtual, 17–23 July 2022; pp. 20810–20826.
7. Chen, T.; Saxena, S.; Li, L.; Fleet, D.J.; Hinton, G. Pix2seq: A language modeling framework for object detection. arXiv 2022, arXiv:2109.10852.
8. Du, X.; Zoph, B.; Hung, W.-C.; Lin, T.-Y. Simple training strategies and model scaling for object detection. arXiv 2021, arXiv:2107.00057.
9. Khan, S.D.; Basalamah, S. Disam: Density independent and scale aware model for crowd counting and localization. Vis. Comput. 2021, 37, 2127–2137.
10. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with pyramid vision transformer. Comput. Vis. Media 2022, 8, 415–424.
11. Xin, Y.; Wang, G.; Mao, M.; Feng, Y.; Dang, Q.; Ma, Y.; Ding, E.; Han, S. PAFNet: An efficient anchor-free object detector guidance. arXiv 2021, arXiv:2104.13534.
12. Yang, J.; Liu, J.; Han, R.; Wu, J. Generating and restoring private face images for internet of vehicles based on semantic features and adversarial examples. IEEE Trans. Intell. Transp. Syst. 2022, 23, 16799–16809.
13. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
14. Ke, W.; Zhang, T.; Huang, Z.; Ye, Q.; Liu, J.; Huang, D. Multiple anchor learning for visual object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10206–10215.
15. Wang, J.; Zhang, W.; Cao, Y.; Chen, K.; Pang, J.; Gong, T.; Shi, J.; Loy, C.C.; Lin, D. Side-Aware boundary localization for more precise object detection. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 403–419.
16. Cao, J.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Pang, Y.; Shao, L. D2Det: Towards high quality object detection and Instance Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11485–11494.
17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.
18. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
19. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; Le, Q.V.; Adam, H. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
20. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient Convolutional Neural Network for mobile devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake

City, UT, USA, 18–22 June 2018; pp. 6848–6856.

21. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In Proceedings of the 2018 European Conference on Computer Vision (ECCV): 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 116–131.
22. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
23. Tan, M.; Le, Q.V. MixConv: Mixed depthwise convolutional kernels. arXiv 2019, arXiv:1907.09595.
24. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for Convolutional Neural Networks. arXiv 2019, arXiv:1905.11946.
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 2016 European Conference on Computer Vision(ECCV): 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
27. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
28. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. arXiv 2018, arXiv:804.02767.
29. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.
30. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
31. Lu, X.; Li, Q.; Li, B.; Yan, J. MimicDet: Bridging the gap between one-stage and two-stage object detection. In Proceedings of the 2020 European Conference on Computer Vision (ECCV): 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 541–557.
32. Mao, Q.-C.; Sun, H.-M.; Liu, Y.-B.; Jia, R.-S. Mini-YOLOv3: Real-Time object detector for embedded applications. IEEE Access 2019, 7, 133529–133538.
33. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. Scaled-YOLOv4: Scaling cross stage partial network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13029–13038.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 2018 European Conference on Computer Vision (ECCV): 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
37. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 510–519.
38. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 1971–1980.
39. Liu, Y.; Shao, Z.; Teng, Y.; Hoffmann, N. NAM: Normalization-based attention module. arXiv 2021, arXiv:2111.12419.
40. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.
41. Zhang, Q.-L.; Yang, Y.-B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.
42. Yang, L.; Zhang, R.; Li, L.; Xie, X. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11863–

11874.

43. Liu, Y.; Shao, Z. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions. arXiv 2021, arXiv:2112.05561.
44. Liu, S.; Huang, D.; Wang, Y. Learning spatial fusion for single-shot object detection. arXiv 2019, arXiv:1911.09516.
45. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 40, 834–848.
46. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the 2018 European Conference on Computer Vision (ECCV): 15th European Conference, Munich, Germany, 8–14 September 2018; pp. 385–400.
47. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
48. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for Instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
49. Huang, X.; Ge, Z.; Jie, Z.; Yoshie, O. NMS by representative region: Towards crowded pedestrian detection by proposal pairing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10750–10759.

Retrieved from https://encyclopedia.pub/entry/history/show/117206