Advances in Computer Science Research, volume 94
International Conference on Big Data, Electronics and Communication Engineering (BDECE 2019)

An Improved Neural Network Model Based on Visual Attention Mechanism for Object Detection

Zeren Jiang*
College of Software, Beihang University, Beijing, China
*Corresponding author

***Abstract*—The general object detection methods include one-stage and two-stage object detection algorithm. The two-stage approach, such as R-CNN family, is composed by the RPN network and object classification network with a better accuracy. The one-stage object detection algorithm represented by YOLO and CornerNet, which are end-to-end structure. This paper proposes an improved CornerNet structure with soft-attention mechanism, which increases the attention weight in the corresponding corner prediction parts of the hourglass model to compensate visually under occlusion or weak light condition. Experiments based on MS COCO dataset show that the proposed structure can lower the inference time further with basically unchanged mAP under the same conditions.**

Keywords—object detection; cornernet; visual attention mechanism; inference time

I. INTRODUCTION

One type of image processing algorithm that looks for specific objects (such as airplanes, boats, tables, chairs, etc.) in an image is called object detection, whose purpose is to obtain the objects' classification (such as an airplane, boat or table and chair) and the position of the object, which is represented by the coordinates of the rectangular detection frame (or some corner position of the frame). Usually in the object detection, it is necessary to distinguish the different areas of the image, such as front or back and extract where the interest is. The deep learning network structure has replaced the traditional object detection and has become the mainstream method in this field.

Since object detection requires precise positioning of the object position in the form of a rectangular boundary, there are two distinct problems: first, multiple candidate problems, that is, multiple rectangular boundary regions need to be reconfirmed and selected; in addition, the selected candidates are still not precise and the second correction is required. The existence of these two problems still requires the speed and correctness of object detection to be improved [1].

Prior to the advent of the R-CNN algorithm, object detection mainly used feature extraction and machine learning classification methods. Early features include the Histogram of Gradient feature H and the HOG Pyramid feature, which were sent to the SVM classifier for classification; then the Deformable Part Model [2] was introduced, which is a HOG feature with component combinations, optimized by calculating the elasticity score between components. Deformable parameters; Selective Search [3] uses layered merging based on color texture and similarity degree merging after over-segmentation, and gives suggested region sorting; Combining Selective Search with DPM/HoG and using SVM classification, for a period of time A good result has been achieved.

Ross Girshick et al. proposed R-CNN [4], which increased the mAP of target detection by more than 30% on VOC. The element of this approach is introducting CNN into implement regional candidates, and the proposed pre-training approach to small sample cases.

Fast R-CNN [1] is the fast algorithm of R-CNN family, which promotes the detection accuracy of R-CNN and decreases the operating cost. Fast R-CNN training is very faster than traditional testing, 10 times faster than SPPnet testing, and has a better mAP than both systems.

Further, Faster R-CNN [5] adds RPN to Fast R-CNN. RPN is to predict object range and object scores at each location and sharing the full convolutional part of the detectional function to achieve near-cost-free calculations. Faster R-CNN achieved the latest object detection accuracy of 70.4% mAP on the VOC.

R-fcn [6] is a region-based full convolutional network that is more efficient and accurate than the R-CNN family because R-fcn is fully convolved, sharing almost all calculations across the entire image. In order to cope with the contradiction between translation invariance and translational difference, position-sensitive score maps in R-fcn are proposed. R-fcn got a good result of 83.6% mA on the PASCAL VOC 2007 dataset, faster than any previous network structure.

All of the above are two-stage object detection algorithms: when training the network, the RPN network is first trained, and then the object area detection network is trained. The two-stage method is highly accurate, but at a slower rate. The one-stage object detection algorithm (also known as one-shot object detection) is characterized by only one phase, so the speed is relatively fast.

YOLO V1 [7] is the first one-stage and end to end object detection method. By changing the two-stage structure to its goal of the bounding box definition and its classification regression, it is possible to model directly from the original data to obtain both bounding boxes and classes. The speed of YOLO V1 is 45 frames per second. A better Fast YOLO version can handle 155 frames, and the mAP indicator is also

very good. YOLO V1 is also very good at object detection in other image types.

The subsequent YOLO9000 [8] was able to detect up to 9,000 different objects. By introducing Batch Normalization and High Resolution Classifier, as well as using Dimension Cluster, Direct Location Prediction, Fine-Gained Features and other technologies, YOLO9000 can control speed and accuracy. The structure yielded 76.8 mAP and 78.6 mAP results on VOC.

The effectiveness of Mask R-CNN [9] is the simultaneous output detection, classification, and pixel-level object segmentation. Mask R-CNN is an extension of Faster RCNN. The difference is that RolAlign is used instead of RoIPooling. The reason is that RoIPooling uses rounding, which causes the space misalignment of the feature map RoI to be mapped back to the original image. Instead of quantization, bilinear interpolation is used to achieve pixel-level alignment. In addition, Mask R-CNN adds a split mask, that is, a small FCN (Fully Convolutional Network) applied to each ROI to predict the split mask in pix2pix. Compared to Faster R-CNN, Mask R-CNN's computational cost of Mask R-CNN is the least. Another big benefit is that it will easily be extended to other fields. Mask R-CNN has achieved good results in the COCO challenge data.

Compared to the YOLO family, YOLO v3 [10] has a larger network structure and further improved accuracy while maintaining a lower computational cost. E.g. YOLO v3 runs in 22 ms at 28.2 mAP, achieving 57.9 mAP on the Titan X GPU and 3.8 times faster than RetinaNet.

RefineDet [11] has both the efficiency and accuracy of one-stage and two-stage, and it trains the single-shot object detection method. RefineDet's improvements include optimization of anchor points and detection objects. The combination of SD and RPN and FPN is mainly implemented in three parts. The role of the anchor refinement module is equal to the RPN network. It is mainly used to get the bbox and remove some negative samples. The transfer connection block performs the feature conversion. The object detection module fuses the characteristics of different layers, and finally outputs the category and boundary area. In real-time, on the standard datasets, RefineDet achieves the best detection accuracy.

In this paper, an improved CornerNet structure is proposed. The main method is to increase the weighting of the stack based on the soft-attention in the Hourglass Network structure to decrease the quantization error and further reduce the volume of the parameters and lower the computing cost. Test results on the MS COCO dataset show that the improved network structure has a further speed improvement over CornerNet.

The organization of this paper is as follows: The second part introduces the basic principles of Hourglass Network, the third part introduces CornerNet and its implementation principle, the fourth part gives the results of VOC data and the final part is the conclusions.

II. HOURGLASS NETWORK

A. *The principle of Hourglass*

Human motion understanding, human-computer interaction and other applications require precise gesture recognition. The stacked Hourglass Networks [12], which is shown in Fig.1, is an important network structure for computer vision for attitude detection. By using a full convolutional neural network and using multi-scale features, image data containing human poses can be processed to output precise pixel locations of key points in the ankle, arm, face, etc. of the human body.

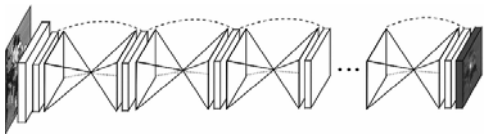

FIGURE I. HOURGLASS NETWORKS FOR POSE ESTIMATION.

Hourglass refers to the network structure shaped like an hourglass, which is shown in Fig.2. The goal of Stacked Hourglass Networks is to find and integrate information on all dimensions of the image. In human body posture detection, local information and global information play different roles for different human body parts. For example, when detecting the hand, local information is more important, and joint point information requires global information. Hourglass Networks reuses the top-down and bottom-up methods. Each top-down and bottom-up structure is a stacked hourglass. This repeated two-way calculation contributed to the performance of Stacked Hourglass Networks.

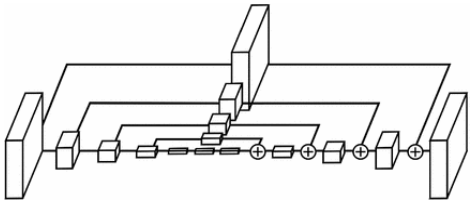

FIGURE II. AN ILLUSTRATION OF HOURGLASS MODULE

B. *Other Hourglass Model*

Another use of the Hourglass model is to perform 3D face alignment [13]. Due to differences in posture, camera, and resolution, facial information is varied, and self-occlusion makes the boundary features not obvious. In the past, occluded landmarks were predicted from contextual information, and other application using the model preserved the mapping between different poses. The CMH Model uses two models, the first to predict both semi-frontal and contour landmarks, and the second Hourglass to estimate 3D faces. Another improvement is to change Hourglass's residual bottleneck block to a multi-scale inception-resnet block. Cascade Multi-View Hourglass performs well on standard 3D face datasets.

In addition, the classic deep CNN is difficult to return to an accurate heat map for occluded backgrounds areas seems like body parts. The Adversarial PoseNet structure [14] mainly draws on the method of human eye prediction posture to

impose constraints on network design. The generative adversarial network is introduced into the pose estimation, and the real-life posture and the false posture are subjected to the generative learning. The structure is a stacked multi-task network structure similar to the Hourglass Network. The good gesture prediction result was achieved on the standard datasets.

III. PRINCIPLE OF CORNERNET

Since the previous method is anchor-based, there are some defects, such as the dense sampling of anchor boxes, which leads to a large number of sample imbalance problems; and anchor boxes bring a lot of hyper-parameters, whose number will decrease the training speed. The obvious advantage of the one-stage approach is that the speed is faster than the two-stage, but the category imbalance problem is very serious, resulting in lower accuracy. In the recent high-precision two-stage target detection series algorithm, Region Proposal needs to extract Anchor boxes, because there are a lot of anchor boxes, such as RetinaNet's anchor boxes, there are about 10^5, and the number of anchor boxes is too large and have many negative samples hyper-parameters.

CornerNet abandoned the anchor boxes and changed the target detection problem [15]. The prediction box is obtained by detecting two key-points in the corresponding corner of the target frame; With this path, the training is started from the beginning, and is not based on the pre-trained target detection.

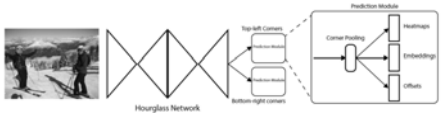

FIGURE III. PRINCIPLE OF CORNERNET

It can be seen from Fig.3, CornerNet has three steps: Hourglass Network, Top-left Corner Prediction Module and Bottom-right Prediction Module. The Hourglass Network is described above: one convolutional neural network predicts a top-left corner, another predicts a bottom-right corner, which called a heatmap. The up-sampling operation increases the resolution of the image and is more capable of predicting the exact position of the object. There are two sets of feature maps, one for the pooling value is the maximum horizontal to the right of the pixel, and one for the vertical downward maximum, after which the two sets of feature maps are summed. he first output of the Prediction Module is C-class channels heatmaps, each of which is a binary mask to represent the corner's position. During the training process, the model reduces the negative samples and sets a positive sample in the radius r region at each ground-truth corner. This is because the vertices falling within the radius r region can still generate valid boundary positioning frames. For the corner point location loss function, the form of focal loss is used.

Compared with the detection frame center or the regional proposal approach, the reason for CornerNet's effective detecting corner points is that the object frame center is difficult to determine for the four corners and their points only depend on the two sides of the object, whose position is easier to be located. Another reason is corner points provide a more efficient discrete boundary space. The formula of focal loss is:

$$L_{\det} = \frac{-1}{N}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\begin{cases}(1-p_{cij})^{\alpha}\log(p_{cij}) & if(y_{cij}=1)\\ (1-y_{cij})^{\beta}(p_{cij})^{\alpha}\log(1-p_{cij}) & other\end{cases} \quad (1)$$

Where N is the number. y_{cij} means the Gaussian bumps is encoded. α and β are the parameters. The mAP of CornerNet on the COCO dataset is further improved compared to RefineNet and its performance is close to the two-stage algorithms, but the computing cost of training and testing has not improved much.

IV. IMPROVED NETWORK WITH ATTENTION WEIGHT

The attention mechanism in vision has been extensively studied in recent years. The attention mechanism does not compress the entire image into a static representation, but dynamically moves the feature of interest to the front as needed. When the target or content in the image is relatively complicated, the use of the uppermost layer from the neural network enables more efficient extraction of information in the image. A hard/soft stochastic/deterministic attention mechanism was proposed and visualized [16].

A recent work is to use the segmentation object tracker to search for targets, and to use the spatial-temporal attention mechanism to handle occlusion problems that are difficult to handle in target detection in the upper and lower frames [17]. Generate a spatial attention map by learning the visible map and continue to improve the latter. The advantage of this approach is that online updates are available, speed and accuracy are verified on standard datasets.

We propose an improved CornerNet network structure based on the Soft-Attention mechanism shown in Fig.4, which increases the attention weighting in the corresponding corner prediction part of the hourglass. This setting assumes that human visual processing is similar to occlusion or when the light is insufficient. More attention is paid to the mechanism of compensation.

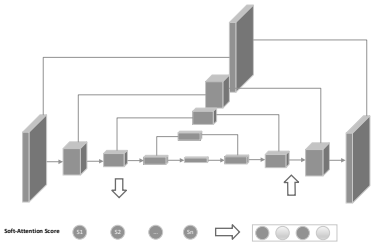

FIGURE IV. THE IMPROVED NETWORK STRUCTURE OF SOFT-ATTENTION.

The attention mechanism of this model is to give different weights to each part of the up-sampling Hourglass model during the training process. By this way, more critical and important information can be extracted, so that the model can

make more accurate judgments. There is no greater overhead in the calculation and storage of the model. Since Soft Attention is parameterized, it can be embedded into the model for direct training. The gradient can be propagated back to the rest of the model via the Attention Mechanism module.

V. EXPERIMENTS

The standard COCO datasets are selected as our basic data to analysis and test the proposed structure [18]. The data in the dataset are divided into three parts with collecting data on various scene types and using classic tools.

TABLE I. THE PERFORMANCE OF THREE DETECTORS

Approach	***backbone***	***Language***	***AP***	***Time***
YOLO2	DarkNet	Python	42.9	193ms
CornerNet	Hourglass	Python	40.1	1.21s
Improved	Hourglass	Python	39.5	136ms

Because YOLO v2 is the advantage method in the two-stage technology with high precision, and CornerNet is an advanced method that is less expensive to calculate in one-stage technology, this paper uses these two network structures to compare with. Tab.1 shows the mAP and inference time for three object detection methods. From the table, the mAP values of these methods are closely and the detection precision of YOLO v2 is better. In terms of inferring time, since most of the computing resources are used for the Hourglass module with a large number of parameters, CornerNet without acceleration has the longest inference time.

VI. CONCLUSION

As an effective end-to-end object detection framework, CornerNet has shown good results on many data sets. However, CornerNet still has some problems with quantization error, large amount of parameters and the speed still needs to be improved. This paper proposes an improved CornerNet based on the attention mechanism, which adds attention weights to the Hourglass to reduce the amount of parameters and further shorten the inference time. Testing on the standard data shows that our improved network structure has a further speed improvement over the two classic networks.

REFERENCES

[1] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[2] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]. 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008: 1-8.

[3] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[5] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems. 2015: 91-99.

[6] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]. Advances in neural information processing systems. 2016: 379-387.

[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] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[10] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[11] Zhang S, Wen L, Bian X, et al. Single-shot refinement neural network for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4203-4212.

[12] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. European conference on computer vision. Springer, Cham, 2016: 483-499.

[13] Deng J, Zhou Y, Cheng S, et al. Cascade multi-view hourglass model for robust 3D face alignment[C]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 399-403.

[14] Chen Y, Shen C, Wei X S, et al. Adversarial posenet: A structure-aware convolutional network for human pose estimation[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 1212-1221.

[15] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]. Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750.

[16] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. International conference on machine learning. 2015: 2048-2057.

[17] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077-6086.

[18] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. European conference on computer vision. Springer, Cham, 2014: 740-755.