Real-Time Object Detection based on Divergent Activation

Hui Liang¹,a, Yueyuan Wang¹,b, Yuan Ma²,c and Kewen Liu²,*

¹Xi'an Satellite Control Center, Xian, China
²School of Information Engineering, Wuhan University of Technology, Wuhan, China

*Corresponding author e-mail: liukewen@whut.edu.cn, alianghwui@163.com, b576622378@qq.com, cyuan.ma@whut.edu.cn

Abstract. Object detection is a research hotspot in the field of computer vision and has significant research value in the fields e.g. human-computer interaction, video surveillance, automatic driving, face recognition, medical imaging and etc. This paper proposes a novel real-time object detection algorithm based on YOLOv2 and divergent activation which aims to extract semantically complementary and discriminative features. The introduced divergent activation mechanism can be divided into differential divergent activation and hierarchical divergent activation that obtains the spatial complementary features by fusing different spatial distribution information of objects at the same level and activates a complete range of objects by fusing complementary semantics of objects from different levels, respectively. Experimental results on PASCAL VOC2007 and PASCAL VOC2012 show that our proposed algorithm reached 81.0% on mean average precision (mAP), indicating the effectiveness of the proposed algorithm.

1. Introduction

Object detection refers to locating and identifying the region of interest efficiently and accurately in images. With the booming evolution of computer science, it has become a research hotspot in the field of computer vision and has significant research value in the fields e.g. human-computer interaction, video surveillance, automatic driving, face recognition, medical imaging and etc. [1]

In recent years, the convolutional neural network (CNN) has achieved success in the field of object detection, and CNN-based algorithms have become mainstream ideas for designing object detection algorithms. Compared with the traditional algorithms, CNN-based algorithms have better feature representation ability and higher accuracy. It can learn favourable high-level feature representations, which can be effectively used for object location and recognition, with sufficient training data. [2]

Based on our observation, the feature extraction part of the YOLOv2[3] algorithm network tends to extract only local features, ignoring the integrity of the object, resulting in a decrease in locating accuracy, which in turn affects detection accuracy. To solve the above problem, this paper proposes a novel real-time object detection algorithm based on YOLOv2 and divergent activation which aims to extract semantically complementary and discriminative features.

2. Related Work

Due to the speed limitation of algorithms based on region proposals, this paper mainly focuses on algorithms based on regression.
In 2016, Redmon et al. proposed YOLOv1 [4], which divides the input image into $S \times S$ cells. Each cell is responsible for regressing B boxes (including the location and classification) whose center points fall within the cell. The B boxes share the same category probability. YOLOv1 can predict at once whether all boxes contain the objects' confidence and category probability with a single glimpse, and output detection results, with the sacrifice of a little accuracy, whose detection speed deployed on Titan X can reach 45 FPS.

Although YOLO greatly improves the detection speed, there still exist many problems that should be solved. Combining the advantages of YOLO and Faster R-CNN [5] algorithm, Liu et al. adopted multi-scale feature extraction for location regression and object classification to adapt to multi-scale detection tasks, and proposed SSD [6]. Liu et al. proposed that low-level large-scale feature maps contain more low-level texture details, suitable for learning small-scale targets, and high-level small-scale features contain more semantic details, suitable for learning large-scale targets. Furthermore, Liu et al. replaced a convolutional layer instead of a fully connected layer at the output of the YOLO to regress the detection results and set several prior boxes with different scales and aspect ratios to adapt to different shapes.

To solve the aforementioned problems of YOLOv1, Redmon et al. proposed an improved YOLOv2 [3] who has about six innovations as following. Firstly, batch normalization layers were added after each convolutional layer to accelerate network convergence and reduce over-fitting. Secondly, before training on detection datasets, high-resolution input images are used to train the classification network to help the network better adapt. Thirdly, prior boxes proposed in Faster R-CNN and SSD were introduced to YOLOv2, and K-means was used to select the number and aspect ratio of prior boxes, which improves the recall rate. Fourthly, the offset of prior boxes is constrained to $[0, 1]$ to ameliorate the difficulty of convergence. Fifthly, a pass-through layer was proposed to rearrange high-resolution feature maps and aggregate them with low-resolution feature maps in channel dimension, which effectively improves the accuracy of small object detection. Moreover, a multi-scale training strategy is adopted, and the size of input images continuously change during training, so that the algorithm can adapt to different inputs, making the algorithm more robust.

To further counterpoise accuracy and speed, Redmon et al. proposed an improved YOLOv3 [7] algorithm based on YOLOv2. Compared with YOLOv2, YOLOv3 replaced the Softmax classifier with an independent Logistic classifier and used cross-entropy loss to guide training, so that YOLOv3 has multi-label classification capabilities. At the same time, YOLOv3 expands the pass-through layer first pro-posed in YOLOv2. Multi-scale fusion is also adopted to further enhance the performance of small objects. Furthermore, the basic network is updated, and a large number of $3 \times 3$ and $1 \times 1$ convolutional layers are added to deepen the depth and improve the accuracy.

3. Proposed Method
In this section, the object detection model based on divergent activation was introduced in details.

3.1. Divergent Activation
Xue et al. proposed a divergent activation mechanism [8] in 2019 to solve the problem that CAM (Class Activation Mapping) [9] mainly focuses on local discriminative regions to improve the performance of object classification but ignores the integrity, thereby affects locating accuracy. The introduced divergent activation mechanism can be divided into differential divergent activation (DDA) and hierarchical divergent activation (HDA). HDA mainly focuses on semantic complementary features from the perspective of the hierarchy level, and DDA mainly focuses on mutually exclusive visual features from the perspective of space.

Hierarchical divergent activation (HDA). The purpose of HDA is to activate similar features that are suppressed and locate the full object extent. The basic idea of HDA is to merge two subclasses into the same parent class, and train a classifier for the parent class. The resulting characteristics of the parent class include the similar characteristics of the two subclasses. And continue to merge into a new parent class, will further activate more features. The loss function of HDA is defined as:
\[ \text{Loss}_{\text{HDA}} = -\sum_{h} \frac{1}{C^h} \sum_{c} y_{c}^h \log \left( p_{c}^h \right) \]  

(1)

Where \( L_{\text{A}} \) is the loss of the \( h^{th} \) class hierarchy. \( C^h \) is the number of the classes in \( h^{th} \) class hierarchy. \( y_{c}^h \) is the label of the \( c^{th} \) class where \( c \in C^h \).

Discrepant divergent activation (DDA). The purpose of DDA is to activate spatial complementary information between objects in a single hierarchical category. In order to achieve the above purpose, a single category feature map is expanded into \( K \) feature maps, and the loss function is used to make the \( K \) feature maps as different as possible. The loss function of DDA is defined as:

\[ \text{Loss}_{\text{DDA}} = \sum_{1 \leq k \leq K} C \left( A_{c}^{k}, A_{c}^{k'} \right) \]  

(2)

Where \( A_{c}^{k} \) denotes the \( k^{th} \) activation map for the \( c^{th} \) class. \( C \left( A_{c}^{k}, A_{c}^{k'} \right) \) denotes the cosine similarity between \( A_{c}^{k} \) and \( A_{c}^{k'} \).

3.2. Overall Network Structure

![Figure 1. Overall Network Structure.](image)

The overall network structure of the proposed algorithm based on YOLOv2 and divergent activation is shown in Figure 1, which consists of two parts, namely feature extraction network and detection network. Feature extraction network is used to learn the feature representations of input images, to provide a basis for object classification and locating. Detection network predicts category, location and corresponding confidence of each object, according to input features.

3.3. Structure of Feature Extraction Network

As shown in Figure 2, we introduce a divergent activation on the basis of the feature extraction network based on YOLOv2. The introduced divergent activation aims to extract semantically complementary and discriminative features to solve that the feature extraction part of the YOLOv2 algorithm network tends to extract only local features, ignoring the integrity of the object, resulting in a decrease in locating accuracy, which in turn affects detection accuracy. The network structure of the feature extraction network is shown in Figure 2.
We first extract multi-scale features after the 8th, 13th and 18th convolutional layers of Darknet-19 and fuse by 1×1 convolution to obtain K feature maps belonging to the same category, and then send to HDA and DDA modules. In the DDA module, concatenation operation is used to combine feature maps, and then DDA loss is used to minimize the similarity, thereby achieving spatial complementarity between each feature map. In the HDA module, feature maps of each level are generated by averaging K feature maps of each category. Global average pooling is used to generate un-normalized probabilities supervised by HDA loss, so as to achieve semantically complementary feature maps of each level.

3.4. Structure of Detection Network

The network structure of the detection network is shown in Figure 3. The input of detection network is the fusion by pixel average of semantically complementary and discriminative HDA features whose size is 7×7×1024. The detection network is consistent with that of YOLOv2, and is composed of four convolutional layers. The parameters of the first three convolutional layers are set to 3×3×2014, the parameters of the last convolutional layer are set to 1×1 to obtain the final output feature map. The output feature maps can be decoded into classification and location confidence as detection results.

3.5. Loss Functions

The loss function of the proposed algorithm consists three weighted parts is defined as:

\[ L_{\text{all}} = L_{\text{Loss}} + \lambda L_{\text{Loss HDA}} + \lambda \lambda L_{\text{Loss DDA}} \] (3)
Where $\text{Loss}_{HDA}$ represents loss of HDA modules, $\text{Loss}_{DDA}$ represents loss of DDA modules, $\text{Loss}$ represents locating loss and classification loss to guide training, and $\lambda$ represents a balance factor.

4. Experiments

In this section, we mainly describe our experiment implement details and some comparisons of experimental results.

4.1. Implementation details

Training datasets used in this paper are ImageNet [10], PASCAL VOC2007 and PASCAL VOC2012 [11]. For evaluation, we mainly use mean Average Precision (mAP) and frame per second (FPS) to measure the performance. It took several hours to train with one TiTan X.

Training phase is divided into two stages. In stage one, we added cascading convolutional, max-pooling and Softmax layer to train the feature extraction network on ImageNet to equip favorable feature extraction ability. In stage two, we added a detection network to fine-tune on PASCAL VOC2007 and PASCAL VOC2012 to obtain detection results.

For parameter settings, $\lambda$ in loss function is set to 0.6. For optimization, it is optimized by ADAM optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$. Learning rate in stage one is set to 1e-4 and will decay by half when there is no more promotion for 10000 iterations. Learning rate in stage two is set to 5e-5 and will decay by half when there is no more promotion for 5000 iterations.

4.2. Ablation Studies

The ablation studies on PASCAL VOC are used to evaluate the effects of proposed DA modules. Study of K in DDA. As shown in Table 1, we studied the influence of K in DDA on detection precision. Our proposed algorithm achieved the best results when K is set to 8. If K is too small, it is hard to achieve spatial complement without enough spatial difference. If K is too large, the model complexity will increase a lot, and there will be a risk of over fitting, decreasing mAP.

| K  | Training set/Testing set | mAP (%) |
|----|--------------------------|---------|
| 2^8| VOC07+12/VOC07           | 78.8    |
| 2^1| VOC07+12/VOC07           | 79.2    |
| 2^2| VOC07+12/VOC07           | 79.5    |
| 2^3| VOC07+12/VOC07           | 79.7    |
| 2^4| VOC07+12/VOC07           | 79.4    |
| 2^5| VOC07+12/VOC07           | 79.2    |

Study of $\lambda$ in DDA. As shown in Table 2, we studied the influence of $\lambda$ in equation 1 on detection precision. Our proposed algorithm achieved the best results when $\lambda$ is set to 0.01. We set K=8 and $\lambda=0.01$ as default settings, in the following experiments.
Table 2. Study of $\lambda$ in DDA.

| $\lambda$ | Training set/Testing set | mAP(%) |
|-----------|--------------------------|--------|
| 0.005     | VOC07+12/VOC07           | 79.6   |
| 0.010     | VOC07+12/VOC07           | 81.0   |
| 0.025     | VOC07+12/VOC07           | 80.9   |
| 0.050     | VOC07+12/VOC07           | 80.1   |
| 0.100     | VOC07+12/VOC07           | 80.7   |
| 0.200     | VOC07+12/VOC07           | 79.9   |
| 0.500     | VOC07+12/VOC07           | 79.6   |
| 1.000     | VOC07+12/VOC07           | 79.3   |

Study of HDA. As shown in Table 3, it can be noticed that the derivative algorithm DarkNet-19+HDA increased by 1.5%, and DarkNet-19+DDA increased by 1.1% on mAP, compared to YOLOv2. The above experimental results verified the effectiveness of HDA and DDA modules for object detection. However, since the main purpose of HDA is to activate a complete range of objects, improving locating accuracy, it is more efficient than DDA. The algorithm DarkNet19+HDA+DDA increased by 2.4% mAP compared with YOLOv2 with the sacrifice of a little running time. However, it's acceptable and can still meet real-time needs.

Table 3. Study of HAD.

| Configuration         | Training set/Testing set | mAP (%) | FPS |
|-----------------------|--------------------------|---------|-----|
| DarkNet-19            | VOC07+12/VOC07           | 78.6    | 40  |
| DarkNet-19+HDA        | VOC07+12/VOC07           | 80.1    | 34  |
| DarkNet-19+DDA        | VOC07+12/VOC07           | 79.7    | 36  |
| DarkNet19+HDA+DDA     | VOC07+12/VOC07           | 81.0    | 32  |

4.3. Comparison with state-of-the-art algorithms

Table 4. Comparison with state-of-the-art algorithms.

| Algorithm            | Training set/Testing set | mAP (%) | FPS |
|----------------------|--------------------------|---------|-----|
| Faster R-CNN[3]      | VOC07+12/VOC07           | 76.4    | 2.4 |
| MR-CNN[12]           | VOC07+12/VOC07           | 78.2    | 0.03|
| R-FCN[13]            | VOC07+12/VOC07           | 80.5    | 9   |
| SSD513[14]           | VOC07+12/VOC07           | 80.6    | 6.8 |
| DSSD513[15]          | VOC07+12/VOC07           | 81.5    | 5.5 |
| RefineDet512[16]     | VOC07+12/VOC07           | 80     | 24.1|
| YOLOv2[5]            | VOC07+12/VOC07           | 78.6    | 40  |
| Ours                 | VOC07+12/VOC07           | 81.0    | 32  |

Comparison with state-of-the-art algorithms is shown in Table 4 where bold represents the best and underlined represents the second best. Clearly from Table 4 that our proposed algorithm can achieve superior results among all the algorithms except DSSD513. However, its FPS is merely 5.5, it’s also worth noting that our proposed algorithm only lacks 0.5% on mAP.

5. Summary
In this Paper, we propose an improved YOLOv2 target detection algorithm based on divergent activation. The algorithm introduces a divergent activation mechanism based on YOLOv2 which aims to extract semantically complementary and discriminative features. Experiment results based on
PASCAL VOC2007 and PASCAL VOC2012 datasets indicate our proposed algorithm can effectively improve the mAP on the premise of ensuring real-time performance.

References
[1] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection [C] //Advances in neural information processing systems. 2013: 2553-2561.
[2] Yu Jinyong, Ding Pengcheng, Wang Chao. Overview: Application of Convolution Neural Network in Object Detection [J]. Computer Science, 2018, 45 (S2): 17-26.
[3] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[4] 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.
[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] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C] // European conference on computer vision. Springer, Cham, 2016: 21-37.
[7] Redmon J, Farhadi A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv:1804.02767, 2018.
[8] Xue H, Liu C, Wan F, et al. DANet: Divergent Activation for Weakly Supervised Object Localization [C] //Proceedings of the IEEE International Conference on Computer Vision. 2019: 6589-6598.
[9] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.
[10] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge [J]. International journal of computer vision, 2015, 115 (3): 211-252.
[11] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge [J]. International journal of computer vision, 2010, 88 (2): 303-338.
[12] Giradis S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware cnn model [C] //Proceedings of the IEEE international conference on computer vision. 2015: 1134-1142.
[13] 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.
[14] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C] //European conference on computer vision. Springer, Cham, 2016: 21-37.
[15] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector [J]. arXiv preprint arXiv:1701.06659, 2017.
[16] 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.