A Real-time Trademark Detection Method

Yang Leng, Qixiong Fan
School of Information Engineering, Guangdong University of Technology, Guangzhou, 510006, China
fantastic_great@163.com, lengyang_gz@citicbank.com

Abstract. In natural scenes, it is an important and challenging task to improve the accuracy while maintaining the real-time capability of trademark detection. For the needs of real-time trademark detection in natural scenes, a method called LOGO-YOLOv3 based on improved YOLOv3 is proposed in this paper. Channel attention and spatial attention mechanisms are introduced into the Darknet53 feature extraction network. And an additional layer is added to the detection layer, which will be combined with the original three detection layers of YOLOv3 to construct a feature pyramid networks containing four detection layers with different scales. This method can not only greatly reduce the false detection of trademarks in complex scenes, but also deal with the problem of large differences in the size of different types of trademarks. Experiments on the Flickrlogos-32 dataset show that the proposed method outperform the original YOLOv3 in general.

1. Introduction
Trademarks are the commercial signs that used to be distinguished different sources of goods, and are intangible assets of enterprises. It is difficult to find similar trademarks in a large number of trademark databases by manual labor. Real-time trademark detection methods are of great significance to safeguard trademark intellectual property rights.

Trademark detection technology is to find the area where the trademark object is located in an image. Trademark detection is a sub-field of object detection. Many methods of object detection can be applied to trademark detection tasks.

Early researchers mainly improved the detection effect through detection algorithms based on artificial feature operators [1-2]. However, the limitation of traditional trademark detection algorithms based on artificial feature operators is that the design of feature operators is time-consuming and labor-intensive. It is difficult to have good generalization ability, detection accuracy and speed. In recent years, due to the outstanding performance of convolutional neural networks in the field of image processing, most researchers use convolutional neural networks to improve the detection effect. Compared with the detection algorithm based on artificial feature operators, the algorithm based on convolutional neural networks do not require time-consuming and laborious design of feature operators. Therefore, it has replaced the traditional method of manually designing features and become the current mainstream method. Now the detection algorithm based on convolutional neural network can be divided into region-based detection algorithm and regression-based detection algorithm. The region-based detection algorithm is mainly represented by Faster R-CNN [3], which uses the region proposal network (RPN) to extract the detection area. It shares the characteristics of the convolution part with the entire detection network, which solves the problems of slow training and large training space. Faster R-CNN [3] truly implements end-to-end detection, but still cannot achieve real-time...
detection. Moreover, most region-based detection methods are limited by the computational complexity of the region proposal box extraction. To achieve real-time detection still faces huge challenges.

In order to achieve real-time detection while ensuring that the accuracy will not be greatly reduced, some researchers began to study regression-based detection algorithms [4-7]. The typical method is YOLO [6] proposed by Redmon J. YOLO first uses the regression method to predict the bounding box coordinates and classification of objects directly from an image. Secondly, the algorithm also has the characteristics of extracting features for the full image. However, YOLO has the defects with serious positioning error and low detection accuracy. In YOLOv2 [7], a series of methods are used to optimize YOLO's model structure, which significantly improves the detection speed. And the detection accuracy is the same as SSD [4]. However, the basic network of YOLOv2 is relatively simple and does not improve the detection accuracy. In order to achieve higher positioning accuracy while ensuring speed, YOLOv3 [5] was proposed. Compared with the previous network, the improvements of YOLOv3 include multi-scale prediction, a more complex network structure called Darknet53, and cancelling softmax used as the candidate box classification. All these make YOLOv3 faster and more accurate. YOLOv3 makes predictions through a single network evaluation, which makes it very fast. Under the same conditions, it is 1000 times faster than R-CNN [8] and 100 times faster than Fast R-CNN [9]. YOLOv3 has obtained the best balance of detection accuracy and speed at present. But in the trademark detection application in natural scenes, YOLOv3 needs to be improved its performance with large size differences and in complex backgrounds.

To address the above issues, we propose a real-time trademark detection method based on improved YOLOv3. In the feature extraction network, channel attention and spatial attention mechanisms are added. The weighted feature vectors are used to replace the original feature vectors for residual fusion, and second-order terms are added to reduce the information loss in the fusion process and accelerate the convergence of the model. In the detection network, a new detection layer is added to build a feature pyramid [18] with four detection layers of different scales. At the same time, it merges with the features of the low layer to enhance the robustness of the detection model to the size of the trademark.

The rest of the paper is organized as follows: Section 2 introduces the YOLOv3 model and attention mechanism; In section 3, our method is proposed; our comparative experiments will be showed in Section 4; the conclusion is given in Section 5.

2. Related Work

2.1. YOLOv3 object detection model
The YOLOv3 algorithm does not need to generate the region of interest (ROI) in advance, but directly trains the network in a regression manner. At the same time, the K-means algorithm is used for the training dataset to cluster the training sample bounding boxes. In the end, 3 sets of predefined bounding box sizes are preset on 3 scales respectively, and the subsequent location prediction will be based on these 9 kinds of bounding boxes. First, feature extraction is performed on the original 416 × 416 input image through the feature extraction network, and then the feature vector is sent to the FPN structure to generate a grid area on 3 scales, respectively 13 × 13, 26 × 26, and 52 × 52. Each grid area predicts 3 bounding boxes, and a total of (52 × 52 + 26 × 26 + 13 × 13) × 3 = 10647 bounding boxes is generated. Then predict a vector above each boundary to generate a prediction frame. Finally, non-maximum suppression is performed on the generated prediction frame to obtain the final prediction result [17]. For the loss function of YOLOv3, Redmon J did not directly express it in the paper [5]. After the interpretation of the source code by LYU et al. [14], the loss function of the YOLOv3 network is summarized as follows:
\[
\text{Loss} = \frac{1}{2} \sum_{r} \lambda_{\text{obj}} \left( \left( 2 - \text{truth}_h \cdot \text{truth}_w \right) \sum_{r \in (y,x,k)} (\text{truth}_r - \text{predict}_r)^2 + \sum_{r \in \text{conf}} ((r = \text{truth}_i) ? 0 : \text{predict}_r)^2 \right) \\
+ (\text{truth}_\text{conf} - \text{predict}_\text{conf})^2
\]  

(1)

It is mainly divided into three parts: coordinate loss, loss and classification loss. \( \lambda_{\text{obj}} \) is 1 when there are objects in the grid cell, otherwise it is 0. They are all calculated by SSE, and finally Loss adopts the form of sum [14].

2.2. Attention mechanism

In trademark detection, Attention Mechanism [10-11] is to focus on the trademark in the many information of the image, select the image area that is more critical for trademark detection and ignore the unimportant information other than trademarks.

In the feature transfer process of the convolutional neural network, the channel attention mechanism places different weights on the feature channels and the network is more sensitive to the channels with larger weights to update the parameters. Figuratively speaking, in the process of forward propagation, the feature channels that are important to the detection task will occupy a larger proportion and the parts that the detection network attaches importance to will be more clearly displayed in the output image.

The spatial attention mechanism uses the spatial relationship of features to generate a spatial attention feature map. Unlike channel attention, the spatial attention mechanism focuses on the useful information part, which supplements channel attention. Pooling operations based on channels can effectively high light the target information area [13]. On the connected feature descriptors, we use a convolutional layer to generate a spatial attention feature map \( M_s(F) \in \mathbb{R}^{H \times W} \), and encode it in places where it is emphasized or suppressed. Two pooling operations are used to aggregate the channel information of a feature map to generate two two-dimensional feature maps: \( F_{\text{avg}} \in \mathbb{R}^{1 \times H \times W} \) and \( F_{\text{max}} \in \mathbb{R}^{1 \times H \times W} \), which represent the average pooling feature and maximum pooling feature across channels, respectively. Then connect them and convolve through a standard convolution layer to generate our two-dimensional spatial attention map. In short, spatial attention can be calculated as follows:

\[
M_s(F) = \sigma(f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \\
M_s(F) = \sigma(f^{7 \times 7}([F_{\text{avg}}; F_{\text{max}}]))
\]  

(2)

Where \( \sigma \) is the sigmoid operation, \( 7 \times 7 \) represents the size of the convolution kernel [12].

3. Our Method

To address the problem of background interference with trademark detection, we introduce channel and spatial attention mechanisms [17] into Darknet53, and second-order terms are added to reduce the information loss in the fusion process and accelerate the convergence of the model, which make the detection model pay more attention to the trademark in the foreground area. To deal with the problem of large differences in trademark sizes, we added an additional detection layer to the YOLOv3. Together with the original three detection layers of YOLOv3, a feature pyramid [18] containing four detection layers of different scales is constructed. It is merged with the features of the bottom layer to enhance the robustness of the detection model to the different size of the trademark.

In summary, we propose a real-time trademark detection method called LOGO-YOLOv3, which based on YOLOv3. The structural block diagram is shown in Fig.1.

3.1. Feature extraction network with attention mechanism

In natural scenes, trademarks usually exist with the complex background. Objects with similar characteristics to the trademark in the background will be located in YOLOv3 algorithm, which will cause a false detection. We adopt channels and spatial attention mechanisms into the feature extraction network to reduce the interference of complex backgrounds on trademark detection. We only replace
all the residual connections in the Darknet53 feature extraction network and filters the transfer features, so that the information retained during residual fusion is more conducive to reducing the loss in training. Therefore, the accuracy of positioning and classification during detection can be improved.

First, the global maximum pooling operation is added to the channel attention mechanism. After the two pooling operations are completed, they are merged and sent to the MLP for channel information screening. Then, average pooling and maximum pooling are performed along the channel domain and the output of the two is combined to obtain a feature descriptor. Finally, the convolution operation is used to encode and the spatial domain attention map is obtained. The above improvements
can not only help the network to classify more accurately, but also more accurately locate the location of the object. The structure of the attention mechanism in the channel domain and the spatial domain is shown in Fig.2.

In order to shorten the transition time of the inconsistent distribution of the new structure and the original structure during the training process, and ensure the transfer of the effective features in the previous network layer, we add extra second-order term and a small offset to the residual connection of the network to increase the nonlinearity of the entire structure, which refers to the Second-Order Response Transform (SORT) method proposed by Wang et al. [16].

3.2. Multi-scale prediction network

The multiple-scale problem also exists in trademark detection. In this paper, a new detection layer is added, and the original three detection layers of YOLOv3 are used to construct a feature pyramid [18] containing four detection layers of different scales to enhance the robustness of the detection model to the size of the trademark.

In the Flickrlogos-32[15] public trademark dataset, the size distribution of the trademark is shown in Fig.3. It can be seen from Fig.4 that the size of the trademark is mostly distributed in the range of length 100. In YOLOv3, when the input image size is 416 × 416, the size of the feature map generated in the three-layer detection layer is 13 × 13, 26 × 26, 52 × 52, respectively. For the problem that the size of the largest feature map is not large enough, and the prediction of small trademarks is not accurate, we added an additional layer after the original detection layer, as shown in Fig.1. Together with the original three detection layers of YOLOv3, a feature pyramid containing 4 detection layers of different scales is constructed, which are 13 × 13, 26 × 26, 52 × 52, 104 × 104. In the meantime, up sample the feature pyramid in 2 times the stride. In order to enhance the representation ability of the feature pyramid, the up-sampled features are concatenated with the corresponding size feature maps generated by the deep residual network in the previous section. Through this connection, the feature maps used in each layer of prediction are fused with different resolutions and different semantic strength features. These are used for trademark detection corresponding to the resolution size, so that each layer has a suitable resolution and strong semantic features. Therefore, LOGO-YOLOv3 can reduce the impact of large differences in the size of trademarks on detection. This method improves the detection accuracy significantly while maintaining the real-time capability of trademark detection.

4. Experiments

We trained and evaluated the performance of the proposed LOGO-YOLOv3 method on the Flickrlogos-32 [15] public trademark dataset. The experiment was conducted on a computer configured with Ubuntu 16.04, Python 3.5, pytorch0.4.0, cuda10.2, GeForce RTX 2080Ti (11GB) and Intel Core i7-7700K CPU.

4.1. Datasets

Flickrlogos-32[15]: The images in this dataset are photos of brand trademarks in natural scenes, and all trademarks have an approximately planar surface. There are 32 classes with 70 pictures in each class, and 2240 images with annotations in total. Among them, 2016 images and 224 images are randomly selected, respectively as training set and test set.

4.2. Evaluation metrics

The two evaluation metrics used in our experiment are the mAP and the FPS. Both values are larger the better. The mAP is the average of the average precision of all classes in the dataset. The value of mAP must be between 0 and 1.

For the accuracy for all types of objects on all images:

\[
\text{MeanAveragePrecision} = \frac{\sum \text{AveragePrecision}_c}{N(\text{Classes})}
\]  

(3)
Only when the detection rate is fast can real-time detection be achieved, which is extremely important for our trademark detection application scenarios. The common metrics for evaluating the detection rate is Frame Per Second (FPS), which means the number of pictures detected per second.

4.3. Experiment settings
We trained and tested three models on images of the same scale, including the original YOLOv3, YOLOv3-y4 (YOLOv3 with an additional detection layer), and the LOGO-YOLOv3. The input size of the network is fixed at 416×416. The anchor box is generated by k-means clustering.

Throughout the training process, the three models are performed 150 iterations uniformly. The momentum and weight decay are configured as 0.9 and 5e-4, respectively, the parameter update method is Adam, and the batch size is uniformly set to 10. The learning rate is initially 1e-4, and sequentially decreases to 1e-5 and 1e-6. Corresponding to the learning rate at each stage, the network iterates 54, 37 and 59 times respectively. The optimization objective function is the loss function.

4.4. Results and analysis
In order to better measure the detection effect of the LOGO-YOLOv3 model, we use $mAP_{@[0.5:0.95]}$ to comprehensively evaluate the detection performance of the model. Since the new FPN strengthens the robustness of the detection model to trademark size, the mAP of the YOLOv3-y4 algorithm is 3.43% higher than the original YOLOv3 algorithm. According to Table 1, the parameter amount is only increased by 0.4%, and the FPS is only less than 2f/s behind. Based on the YOLOv3-y4 algorithm, due to the superimposed use of channels and spatial attention mechanisms, and improved residual connection, the detection model is more focused on the trademark in the foreground area. The mAP of the LOGO-YOLOv3 algorithm is 3.95% higher than the original YOLOv3 algorithm. According to Table 1, the parameter amount is only increased by 3.2%, and the FPS is only about 2f/s behind.

| Method      | Settings                  | Evaluation Metrics |
|-------------|---------------------------|--------------------|
| YOLOv3      | pre-trained weights       | Total parameters: 6.15237×10⁷, mAP: 47.91, FPS: 36.1 |
| YOLOv3-y4   | pre-trained weights       | Total parameters: 6.17854×10⁷, mAP: 51.34, FPS: 34.4 |
| LOGO-YOLOv3 | pre-trained weights       | Total parameters: 6.35109×10⁷, mAP: 51.86, FPS: 33.9 |

5. Conclusions
In this paper, we proposed a real-time trademark detection method LOGO-YOLOv3. In the feature extraction network, channel attention and spatial attention mechanisms are added. The weighted feature vectors are used to replace the original feature vectors for residual fusion, and second-order terms are used to reduce the information loss in the fusion process and accelerate the convergence of the model, which reduces the interference of complex backgrounds. In the detection network, a new detection layer is added to build a new feature pyramid with four detection layers of different scales, which merges with the features of the low layer to enhance the robustness of the detection model. We verify the effectiveness of our method on public datasets. LOGO-YOLOv3 improves the accuracy while maintaining the real-time capability of trademark detection. In the future, we will focus on the personalized improvement of feature pyramids for different datasets, and the visualization of attention mechanism.

Acknowledgments
This research was supported by the National Natural Science Foundation of China (No. U1701266),
the Natural Science Foundation of Guangdong Province (No. 2018A030313751), the Science and Technology Planning Project of Guangdong Province (No. 2018B030322016) and the Guangdong Higher Education Engineering Technology Research Center for Big Data on Intellectual Property of Manufacturing.

References

[1] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.
[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.
[3] 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.
[4] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
[5] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.
[6] 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.
[7] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
[8] 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.
[9] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[10] Borji A, Itti L. State-of-the-art in visual attention modeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 185-207.
[11] Chikkerur S, Serre T, Tan C, et al. What and where: A Bayesian inference theory of attention[J]. Vision research, 2010, 50(22): 2233-2247.
[12] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
[13] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint arXiv:1612.03928, 2016.
[14] LYU Shuo, CAI Xuan, FENG Rui. YOLOv3 Network Based on Improved Loss Function. Computer Systems Applications, 2019, 28(2):1-7(in Chinese).
[15] Romberg S, Pueyo L G, Lienhart R, et al. Scalable logo recognition in real-world images[C]//Proceedings of the 1st ACM International Conference on Multimedia Retrieval. 2011: 1-8.
[16] Wang Y, Xie L, Liu C, et al. Sort: Second-order response transform for visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1359-1368.
[17] XU Chengji, WANG Xiaofeng, YANG Yadong. Attention-YOLO: YOLO detection algorithm that introduces attention mechanism. Computer Engineering and Applications, 2019, 55(6): 13-23(in Chinese).
[18] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.