Real-time face detection based on YOLO

Wang Yang[1], Zheng Jiachun[2]
(1.Navigation Institute, Jimei University, Xiamen, Fujian, China
2.Information Engineering Institute, Jimei University, Xiamen, Fujian, China
No. 185 Yinjiang Road, Jimei District, Xiamen, Fujian, China
1.86-18860031081 and 1510109901@qq.com
2.86-13806073677 and 415203705@qq.com)

Abstract

As a target detection system, YOLO has a fast detection speed and is suitable for target detection in real-time environment. Compared with other similar target detection systems, it has better detection accuracy and faster detection time. This paper is based on YOLO network and applied to face detection. In this paper, YOLO target detection system is applied to face detection. Experimental results show that the face detection method based on YOLO has stronger robustness and faster detection speed. Still in a complex environment can guarantee the high detection accuracy.At the same time, the detection speed can meet real-time detection requirements

Key words: YOLO, target detection,face detection, dimensional clustering

Introduction

A common target detection algorithm, which selects a region of interest (ROI) on a given image as a candidate region, extracts candidate region features such as a local binary characteristic[1] (LBP) or a directional gradient histogram [2] (HOG) feature, and classifies the area through the training classifier. In the context of large data, however, machine learning methods that rely on preset models have not been able to accurately and fully describe the data characteristics in the application scenario.

In recent years, with the vigorous development of the deep learning, based on the deep study of target detection, has made great development. There are two kinds of object detection based on deep learning: one is the R-CNN series based on regional detection: Fast R-CNN [3], Faster R-CNN [4] and the other is the regression SSD[5] and YOLO[6]. Based on the region target detection method, it contains all sorts of potential regional generating parts and various feature layers, which makes the algorithm's real-time quality not guaranteed.

From network design, the difference between the YOLO and RCNN series network is as follows: (1) training and detection, feature extraction and classification of regression of YOLO, all is done in a single network. It's a separate end-to-end network;(2) YOLO regard object detection as a regression problem. Once the image is input into the network, the position of all objects in the image, their categories and corresponding confidence probabilities can be obtained.The detection results of the RCNN series can be divided into two parts: object category (classification), object location and bounding box (regression problems). In 2018, Redmon et al [7] proposed target detection method of the YOLO V3.When we tested the 320x320 of picture,YOLOv3 runs in 22 ms at 28.2 mAP, as accurate as SSD but three times faster!

Related Working

A. Detection principle of YOLO

The input image is divided into $S * S$ uniform grid, and each cell is composed of (x, y, w, h) and confidence $C(Object)$. The coordinates (x, y) represent the position of the center of the detection boundary box relative to the grid. (w, h) is the width and height of the detection boundary box. Each grid predicts the probability of C categories. The confidence score reflects the probability of the model to include the target object and the accuracy of the prediction detection box. $\Pr(Object)$ stands for whether there is a target object falling into this cell. If there is confidence, it is defined as:

$$C(Object) = \Pr(Object) * IOU(\Pr ed, Truth) \quad (1)$$

If it is determined that the cell does not have a target object, the confidence score should be zero $C(Object) = 0$

IOU is the overlapping rate of the generated candidate bound and ground truth bound, that is, the ratio of their intersection and union.

$$IOU(\Pr ed, Truth) = \frac{area(box_{truth}) \cap area(box_{pred})}{area(box_{truth}) \cup area(box_{pred})} \quad (2)$$

After obtaining the confidence of each prediction box, the low-score prediction box are removed by setting the threshold value, and then non-maximum suppression is performed on the remaining bounding box.

B. Detect network

The size of input image is 416*416. After a series of convolution and batch normalization operations on the input image, the input image is sampled three times, 32 times, 16 times and 8 times, and the multi-scale feature map is obtained. After 32 times down sampling, the feature map is too small, therefore YOLO V3 uses the up sampling with step size of 2 to double the size of the resulting feature map, which becomes 16 times of down sampling. Similarly, the feature map sampled on 16 times is sampled with a step length of 2, and the feature map with a sampling size of 8 times is obtained, so that deep feature can be used for detection.

Detecting targets at different scales is challenging, especially for small targets. Feature pyramid network (FPN) [8] is a feature extractor designed to improve accuracy and speed. It replaces feature extractors in detectors (such as Faster R-CNN) and generates higher-quality feature graph pyramids.

ISBN: 978-1-5386-5267-1

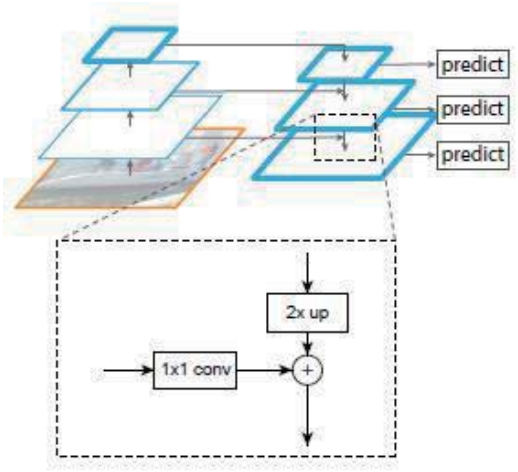


Fig. 1 A feature fusion module

The network connection structure consists of two routes: bottom-up line, stop-down lines, and horizontal connections. The bottom-up process is actually the forward process of the network. Extract the output features of the last layer of each stage to form the feature pyramid. The top-down process is carried out by up sampling. While the horizontal connection is to merge the results of the up sampling and the feature map of the same size generated from the bottom up.

The main function of 1*1 convolution kernel is to reduce the number of convolution kernel without changing the size of feature map. This allows for more valuable feature information from the up sampling layer and fine-grained features from the previous feature map. The top features merge with an up sampling and downward feature, and each layer is independently predicted.

C. loss function

The loss function of YOLO V3 adapots mean square error, which consists of three parts: coordinate error, IOU error and classification error.

$$\begin{aligned}&\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left\|{}_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2]\right.\\&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left\|{}_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2]\right.\\&+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left\|{}_{ij}^{obj}(C_i-\hat{C}_i)^2\right.\\&+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\left\|{}_{ij}^{obj}(C_i-\hat{C}_i)^2\right.\\&+\sum_{i=0}^{S^2}\left\|{}_{i}^{obj}\sum_{C\in classes}(p_i(c)-\hat{p}_i(c))^2\right.\end{aligned}\tag{3}$$

In the back propagation of the training process, it is assumed that the bbox of large and small objects has the same width error ratio $\frac{w_p-w_t}{w_p}$, (w_p is the predicted value and w_t is the real value). The width error w_p-w_t of large size bbox will be significantly smaller than width error w_p-w_t of small size, resulting in the accuracy rate of bbox width prediction of small objects is low. In order to improve this phenomenon, $\sqrt{w},\sqrt{h}$ is used instead of w,h. As shown in the figure 2, for the same width error w_p-w_t, width error $\sqrt{w_p}-\sqrt{w_t}$ of the bbox of small size object is larger than the width error $\sqrt{w_p}-\sqrt{w_t}$ of larger size, thus enhancing the influence of the width error of the bbox of small size object to some extent. So is the height error.

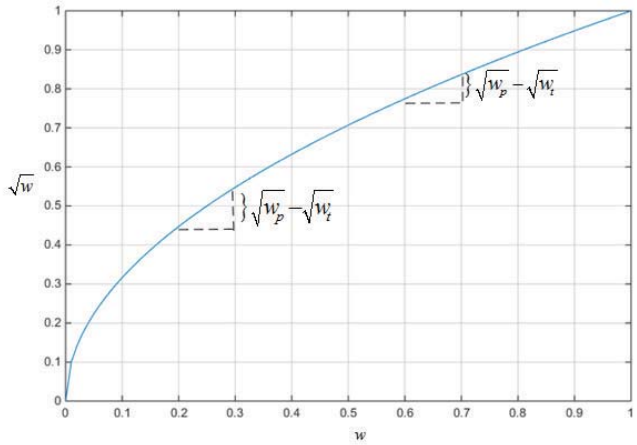


Fig. 2 Object width squared error diagram

D. Dimension clustering

Although YOLO V3 has achieved good detection results, it is not completely suitable for image positioning tasks. Therefore, the corresponding improvements to YOLO V3 are made for specific problems.

When training the network, the initial specification and number of candidate frames are required. With the increasing number of network iterations, the network learns face features, and the parameters of the prediction box are adjusted continuously approaching the real box finally. The anchor of YOLO V3 is identified by VOC2007 and VOC2012 data aggregation classes. These two data sets are rich in categories, and the parameters of the anchor determined are universal, but not suitable for specific detection tasks. Therefore, clustering operations need to be carried out again in the detection data set. In order to speed up the convergence speed and improve the accuracy of face detection, k-means method is used to cluster, and the initial candidate frame parameters closest to the face boundary frame in the image is obtained.

In general, k-means clustering uses Euclidean distance to measure the distance between two points. In this paper, IOU is used to replace the Euclidean distance in k-means. The objective function of clustering is:

$$d(box,centroid)=1-IOU(box,centroid)\tag{4}$$

Experiment

Experimental environment: the processor is Inter Core i7-7770, the CPU is 3.60GHZ*8, 7.7GB memory, GPU is GTX 1080. The operating system is Ubuntu16.04, 64-bit. The depth learning framework is darknet

A.Data Set

This experiment uses WIDER FACE data set. During training, the SGD algorithm is adopted. There are 30,200

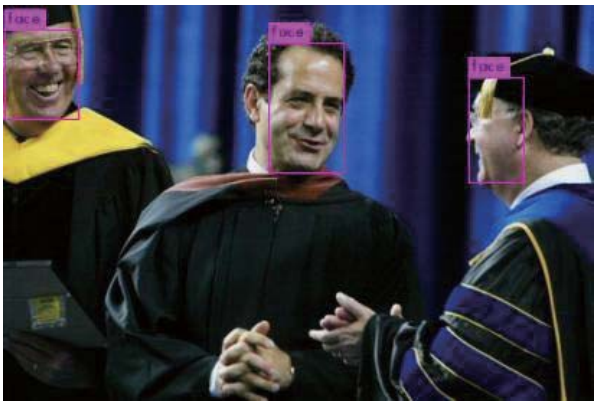
ISBN: 978-1-5386-5267-1

iterations in total during training. Learning rate is 0.001 and the learning rate is reduced at 5000 and 20000 times.

B. Result

Respectively in the three data sets: Celeb Faces[9], FDDB[10],WIDER FACE[8] choose some pictures on the verification test results. (a) and (b) are from Celeb Faces dataset, (c) and (d) are from FDDB dataset, (e) and (f) are measured on WIDER FACE.

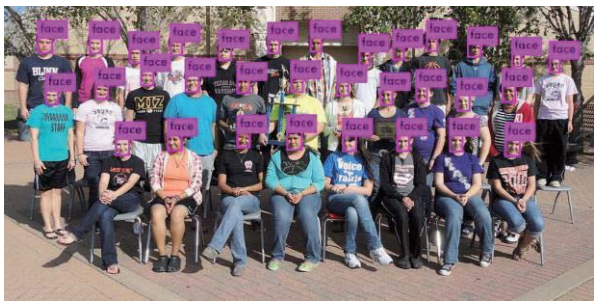
(a) (b)



(c)

(d)

(e)



(f)

Fig. 3 Test results

20 images are selected from the three data sets, and the image sizes are adjusted respectively to obtain the average detection time, as shown in table 1.

TABLE I
AVERAGE DETECTION TIME

Image size	178*218	450*320	2014*680
Detection time(s)	0.027199	0.027246	0.029455

Conclusion

Compared with the traditional algorithm, the face detection method based on yolo v3 has shorter detection time and stronger robustness, which can reduce the miss rate and error rate. It can still guarantee a high test rate in a sophisticated environment, and the speed of detection can meet the real time requirement, and achieve good effect.

Acknowledgment

This work is financially supported by the science and technology key projects of Fujian province（2017H0028）, the natural science foundation of Fujian province (2013J01203), The authors would like to thank once again.

Reference

[1] Li L, Feng X, Xia Z, et al., *Face spoofing detection with local binary pattern network*, 54th ed., Journal of visual communication and image representation, 2018, pp. 182-192.

[2] Kim S, Cho K, *Trade-off between Accuracy and Speed for Pedestrian Detection Using HOG Feature*, IEEE, Berlin,

Germany, pp. 207-209, September 2013.
[3] Girshick R, *Fast R-CNN*, IEEE ICCV. pp. 1440-1448, 2015.
[4] Ren S, He K, Girshick R, et al., *Faster R-CNN: Towards real-time object detection with region proposal networks, Advances in neural information processing systems*. pp. 91-99, 2015.
[5] Liu W, Anguelov D, Erhan D, et al., *Ssd: Single shot multibox detector*. 9905th ed., IEEE ECCV, 2016, pp. 21-37.
[6] Redmon J, Divvala S, Girshick R, et al., *You only look once: Unified, real-time object detection*, IEEE CVPR, 2016, pp. 779-788.
[7] Redmon J, et al., *You only look once: Unified, real-time object detection*, unpublished.
[8] Yang S, Luo P, Loy C C, et al., *Wider face: A face detection benchmark, IEEE CVPR*, 2016, pp. 5525-5533.
[9] Yang S, Luo P, Loy C C, et al., *From facial parts responses to face detection: A deep learning approach*, IEEE ICCV, 2015, pp. 3676-3684.
[10] Jain V, Learned-Miller E, *Fddb: A benchmark for face detection in unconstrained settings*. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

ISBN: 978-1-5386-5267-1