Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/19/20231039

Review of pedestrian detection technology based on deep learning

Zhaoyang Zhong

School of Electronic and electrical engineering, Shanghai University of Engineering Science, Shanghai, China, 201620

1871968819@qq.com

Abstract. Pedestrian detection is one of the main research problems in the field of computer vision. With the development of deep learning technology, pedestrian detection algorithms have made great progress and breakthroughs. Currently, pedestrian detection technology using deep learning has become mainstream. The first type of algorithm is called two-stage detection algorithms, which are based on Region of Interest (ROI) and single-stage detection algorithms adopt an end-to-end training method. So this paper conducts extensive research on the basis of classification and comparison, studies the characteristics and development of the above two detection algorithms, respectively, and finally puts forward suggestions for their future development, such as the fusion of image background and target state.

Keywords: target detection, deep learning, convolutional neural networks.

1. Introduction

Pedestrian detection is a target detection task with a clear goal that primarily targets pedestrians on the road [1]. Although it has been continuously improved and developed in the past few years, challenges such as overlapping pedestrians and small pedestrian scales in densely populated scenes have not been well addressed in real-life situations [2]. Therefore, researching an efficient detection algorithm that can perform equally well in densely crowded pedestrian scenes has significant practical significance.

Currently, pedestrian detection technology using deep learning has become mainstream. Deep learning-based pedestrian detection technology can be divided into two categories: two-stage detection algorithms [3] and single-stage detection algorithms [4]. With the development of science and technology, both detection technologies have been updated. Therefore, the purpose of this paper is to sort out and review the development of the above two detection algorithms. This article will be beneficial to people in need of a knowledge update.

2. Two-stage detection algorithm

2.1. R-CNN (Region-CNN)

The R-CNN model can be summarized as follows [5].

1) The input process involves feeding training set images into the system.

2) The detection process utilizes the Selective Search (SS) algorithm to generate 2k region proposals from the input image, which are then scaled down to a fixed size of 227x227.

DOI: 10.54254/2755-2721/19/20231039

3) The CNN backbone network extracts image features from each region proposal, yielding a fixed-length feature vector.

4) The feature vector is then input into both an SVM (SupportVectorMachine) classifier and a fully connected network for classification and position regression, respectively.

The R-CNN method employs the Selective Search algorithm to pre-extract a set of candidate regions (usually between 1,000 to 2,000) that may contain objects. These candidate regions undergo further feature extraction and classification to determine whether they contain the target object.

After extracting deep features, recognition-based deep features are used to recognize each candidate region. Since the candidate regions may not be completely accurate, the R-CNN approach introduces the bounding box regression strategy, which uses non-maximum suppression to eliminate unnecessary bounding boxes. The Intersection-over-Union (IOU) metric between the candidate and calibration boxes is used for object/background classification. A candidate region is classified as an object if the overlap ratio is greater than 0.5, and as background, if the ratio is less than 0.5.

2.2. SPP-NET (Spatial Pyramid Pooling Net)

In 2015, SPP-NET was proposed by Kaiming He etal.as an extension of R-CNN [6]. In order to accommodate images of any size and address the problem of information loss due to normalization in the R-CNN model, the model incorporated a spatial pyramid pooling (SPP) layer before the fully linked layers.As is shown in Figure 1.

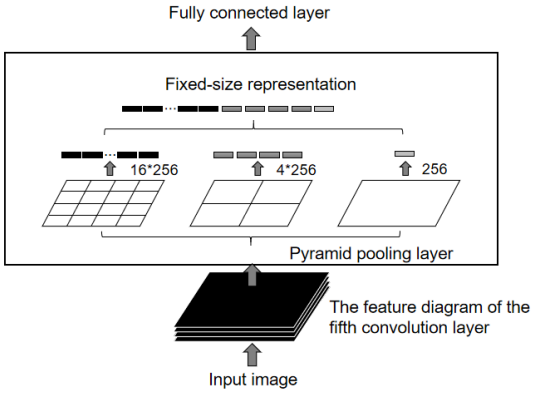

Figure 1. SPP layer model.

Similar to R-CNN, SPP-NET also requires generating candidate regions beforehand. However, instead of inputting approximately 2k candidate regions into the CNN feature extraction network, the entire image containing all candidate regions is input, and the feature representation for the entire image and all candidate regions is obtained through one pass of the convolutional network.

2.3. Fast R-CNN (Fast Region-CNN)

Fast R-CNN similarly uses the Selective Search algorithm (SS), but instead of requiring the input of various area images, Fast R-CNN maps the regions of interest (ROIs) onto the CNN model's feature layer and immediately extracts the appropriate deep features [7]. Using Softmax and a network that has been trained to learn a bounding box regressor, the recovered features are then used directly to predict the area category. One module contains the whole feature extraction, classification, and bounding box regression processes.

During training, Fast R-CNN adjusts the entire network using the VOC (Voice of customer) object detection dataset, and trains the classifier and bounding box regressor at the final part of the network. The training approach in Fast R-CNN uses stochastic gradient descent (SGD), where 2 images and 128 ROIs are selected each time, and the images are set to a fixed size. Although only 2 images are input

DOI: 10.54254/2755-2721/19/20231039

per iteration, up to 128 ROIs can be trained, and feature calculations can be shared between every 64 ROIs.

2.4. Faster R-CNN (Faster Region-CNN)

Ren et al. proposed the Region Proposal Network (RPN) in 2015, which led to the development of the Faster R-CNN model [8]. This algorithm integrates all four basic steps of object detection into one deep network, improving the overall performance of the algorithm, particularly in terms of detection speed.

The Faster R-CNN model no longer relies on the Selective Search algorithm. Instead, the image features are first extracted using a CNN backbone network. The RPN network receives the feature map after which each feature point has an anchor box with a predefined scale and aspect ratio. To determine whether an object is present at each place, the intersection-over-union and offset between the anchor boxes and the ground-truth object boxes are determined. The RPN network is subsequently trained using the deviation loss for position regression after classifying the predefined anchor boxes as foreground or background. This enables the Region of Interest (ROI) position to be modified and then transmitted to succeeding networks. The ROIs of different sizes are then subjected to ROI Pooling to obtain a fixed-size feature vector, which is inputted into a subsequent fully connected network for fine-grained object classification and position regression. The final detection results are then obtained.

2.5. Mask R-CNN (Mask Region-CNN)

Using a tiny Fully Convolutional Network (FCN), Kaiming He et al. implemented a parallel mask branch to Faster R-CNN in 2017 [9]. While Faster R-CNN outputs the class labels and box coordinates, Mask R-CNN adds an additional output, i.e., the object mask. In Mask R-CNN, bilinear interpolation is used to address the pixel misalignment problem, i.e., ROI Align.

After using ROI Align instead of ROI Pooling, Mask R-CNN has achieved outstanding results in the field of object detection, surpassing Faster R-CNN. The Mask R-CNN model is flexible and can be easily adapted to various tasks.

3. Single-stage detection algorithms

3.1. YOLO v1

Different from the two-stage detection algorithms represented by R-CNN, YOLO (*You Only Look Once)* has a simple network structure and is about 10 times faster than Faster R-CNN, with good real-time performance [10]. The model uniformly scales the input image to 448×448×3 and divides it into 7×7 grids, each of which is responsible for predicting the position and confidence of two bounding boxes bbox. These two bbox correspond to the same category, with one predicting a large object and the other predicting a small object. During training, the model adjusts the predicted position of bbox as the network weights are updated. In the training process, the model calculates which grid contains the center of the real object by using the Ground Truth (GT) bounding box, and then detects the object by the grid containing the center point (even when the object spans multiple grids).

3.2. YOLO v2

The YOLO v2 model divides the original image into a grid of 13x13 cells and borrows the idea of anchor boxes from Faster R-CNN [11]. Through clustering analysis, YOLO v2 determines that each cell should be assigned 5 anchor boxes, and each anchor box should predict 1 class. YOLO v2 performs object position regression by predicting the offsets between the anchor boxes and the grid cel1s. This approach is more favorable for neural network training.

In addition, the features of small objects may have been ignored or become indistinguishable after being extracted by multiple layers of the CNN network. To detect fine-grained features, researchers designed a passthrough layer. In YOLO v2, the 26×26×512 feature map before the last pooling layer is divided into 4 pieces of 13×13×512, and then passed through a 1×1 convolution and 2×2 pooling to

DOI: 10.54254/2755-2721/19/20231039

obtain a 13×13×1024 feature map, which is concatenated with the feature map of the same size from the previous layer and outputted. This helps improve the accuracy of detecting small objects.

3.3. YOLO v3

In 2018, Redmon J et al. proposed an improved version of the YOLO, called YOLO v3 , based on YOLO v2. YOLO v3 predefines 3 anchor boxes for each grid through cluster analysis and only uses the first 52 layers of the darknet structure in YOLO v1 and YOLO v2, and extensively utilizes residual structures for skip connections [12]. The basic module of the residual structure consists of convolution (conv), batch normalization (BN), and the Leaky ReLU activation function.

To improve the detection accuracy of small objects,YOLO v3 uses upsampling to extract deep features , making them have the same dimensions as the shallow features tobe fused but with different channel numbers. The features are then concatenated along the channel dimension to achieve feature fusion, and the corresponding detection heads also use fully convolutional structures.This approach can not only increase non-linearity and generalization performance, but also reduce the number of model parameters and improve real-time performance.

YOLO v3 achieves downsampling by setting stride = 2 in 5 convolution operations.

3.4. YOLO v4

YOLO v4 is characterized as an integration of training techniques for object detection algorithms [13]. It builds upon the YOLO object detection framework and adopts the best optimization strategies from the recent developments in the field of convolutional neural networks (CNN) in areas such as data processing, backbone networks, network training, activation functions, and loss functions. It is considered one of the strongest real-time object detection models available.

YOLO v4 is an efficient and powerful model that enables developers to train a super-fast and accurate object detector with only one 1080Ti or 2080Ti GPU, thus lowering the training barrier for the model.

3.5. YOLO v5

The YOLO v5 algorithm, proposed by the Ultralytics team in 2020, is a one-stage object detection method that preserves detection accuracy while providing four different models: V5-S, V5-M, VS-L, and V5-X, based on different network depths and widths [14].

The YOLOv5 network structure consists of three parts: the Backbone feature extraction layer, Neck feature fusion layer, and Head detection layer.

The C3 module, Conv, and SPPF are the three components that make up the backbone network. Its job is to take the input image's features and extract them. After preprocessing, the input image is repeatedly run through C3 and the Conv module for feature extraction and downsampling. Finally, pooling is done to extract backbone features using the SPPF structure.

The Neck module is designed to fully leverage the feature information extracted by the Backbone network. YOLO v5 employs a feature pyramid network (FPN), which can transmit high-level semantic information to the bottom layer, and can also transmit bottom-level localization information to the top layer.

The Head layer of YOLO v5 outputs three feature maps of different sizes.The 80×80 feature map is responsible for detecting small objects in the image, the 40×40 feature map for medium-sized objects, and the 20×20 feature map for large objects. On each layer of feature maps, three prediction boxes with different aspect ratios are predefined, each containing the location information and confidence of the predicted object. Finally, the Non-Maximum Suppression (NMS) algorithm is used to discard overlapping prediction boxes whose intersection over union (IOU) exceeds a certain threshold.

3.6. SSD (Single Shot MultiBox Detector)

The low localization accuracy and difficulty in recognizing small objects in the YOLO algorithm were solved by Liu W et al.'s SSD algorithm, which was proposed in 2016 [15]. The grid division technique

DOI: 10.54254/2755-2721/19/20231039

is still used in SSD, but features are extracted from several convolutional layers of the fundamental network. Each feature map has a varied number and size of anchor boxes, and to boost the detection precision of multi-scale objects, the anchor box size is set from tiny to large as the number of convolutional layers rises.

4. Conclusion

With the rapid development of scientific research technology, updating computer hardware has become increasingly powerful, and the scale of domestic and foreign datasets continues to expand. As a result, the accuracy and speed of deep learning-based object detection algorithms are constantly improving. Generally speaking, two-stage object detection frameworks based on region proposals achieve high accuracy with low false negative rates, but are slow and cannot meet the requirements of real-time or embedded mobile platforms. Single-stage object detection frameworks based on position regression provide another idea, directly conducting classification and position regression to improve detection speed, and the performance in terms of false negative rate and detection accuracy is also continuously improved in subsequent versions. However, there are still many difficulties and challenges with object detection. These include: 1) the fusion of image background and target state; 2) multi-level and multi-dimensional feature extraction methods; 3) feature representation based on deep learning. It is believed that with the continuous development of deep learning, more and more scholars will explore pedestrian detection, and better and more optimal algorithms will emerge.

References

[1] Fu Xie, Dingju Zhu. Summary of deep learning objective detection methods[J]. Computer system applications, 2022, 31(2): 1−12.

[2] Wei Zhou, Research on the pedestrian detection method based on the convolutional neural network[D]. Chengtu: University of Electronic Science and Technology of China, UESTC, 2019.

[3] Zefang Guo, Summary of deep learning algorithms for image object detection[J]. Mechanical engineering and automation, 2019(1): 220-222, 224.

[4] Zhengming Li, Jinlong Zhang. Detection and positioning of grasping objects based on deep learning[J]. Information and control, 2020, 49(2): 147-153.

[5] AGRAWAL P, GIRSHICK R, MALIK J. Analyzing the performance of multilayer neural networks for object recognition [C]// European Conference on Computer Vision. Switzerland: ECCV, 2014: 329-344.

[6] HE K M, ZHANG X Y, RRN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C]// European Conference on Computer Vision, Springer, Cham, 2014: 346-361.

[7] GIRSHICK R. Fast R-CNN [C]//IEEE International Conference on Computer Vision. Santiago : ICCV, 2015: 1440-1448.

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[9] HE K M, GKIOXARRI G, DOLL R P, et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[10] Xiaoyan Zhou, Ke Wang, Lingyan Li. Summary of deep learning-based object detection algorithms [J]. Electronic measurement technology, 2017, 40(11): 89-93.

[11] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]// IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas: CVPR,, 2016: 6517-6525.

[12] HENRIQUES J F, CAREIRA J, RUI C, et al. Beyond hard negative mining: efficient detector learning via block-circulant decomposition[C]//IEEE International Conference on Computer Vision. Sydney: ICCV, 2014: 2760-2767.

DOI: 10.54254/2755-2721/19/20231039

[13] BOCHKOVSKIY A, WANG C Y, LIAO H. YOLOv4: optimal speed and accuracy of object detection [Z/OL]. (2020-04-23) [2021-05-20]. https: / /arxiv.org /abs/2004. 10934.
[14] Zhongmin Zhang, Ze Wu. Dense pedestrian detection method based on an improved YOLOv5[J/OL]. Applied technology. https://kns.cnki.net/kcms/detail/23.1191.u.20221101.1549.002.html
[15] LIU W, AAGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [C] // European Conference on Computer Vision, Springer, Cham, 2016: 21-37.