Recent Deep Learning Approaches for Object Detection

Yanzhe Fu*

Department of Art and Science, Rensselaer Polytechnic Institute, United States

*Corresponding author: fuy9@rpi.edu

Abstract. Object detection, a classic problem in computer vision, has been developed for more than 20 years. From the early traditional methods to today's deep learning methods, the accuracy is getting higher and higher, and the speed is getting faster and faster, which is benefit from deep learning and the continuous development of deep neural networks. Although the research on object detection is constantly developing, there are not many reviews on object detection, so this article will review the object detection after the introduction of deep learning. This article will first introduce the history of object detection, and then focus on a systematic introduction to the development of deep learning object detection in recent years, as well as one-stage detector and two-stage detector in anchor-free and anchor-based. The various methods applied in the stage detector will be sorted out, and the potential problems and future development of object detection will also be analyzed.

Keywords: object detection, deep neural network, convolution neural network.

1. Introduction

Object detection is to find out the objects of interest in an image or video and detect their location and size at the same time. Different from the image classification task, target detection not only needs to solve the classification problem, but also solve the positioning problem, which is a multi-task problem. Conventional object detection is an early object detection algorithm [1]. The traditional target detection algorithm flow is as follows: Artificially construct the characteristics of the detected object; Select the detected area and select the area that may contain objects; Perform feature extraction on areas that may contain objects; Use machine learning methods to classify or regress the extracted features.

There are several classic algorithms for traditional target detection, such as HOG Detector, DPM Detector and so on. HOG (Histogram of Oriented Gradients) detector was proposed in 2005. It is an algorithm for feature histogram extraction based on local pixel blocks. It has good stability under the local deformation of the target and the influence of illumination. HOG has laid an important foundation for many later detection methods, and related technologies are widely used in various computer vision applications. The DPM (Deformable Parts Model) [2], the champion of the VOC 2007-2009 Object Detection Challenge, can be regarded as an extension algorithm of HOG. The DPM algorithm consists of a root filter and multiple part-filters, through hard negative mining, bounding box regression and context priming (Context priming) techniques Improve detection accuracy. However, there are two problems with this type of method: firstly, the construction of task features depends on human background knowledge, so the quality of the features is unstable; secondly, these features are shallow, mostly based on statistical methods, and cannot be very good. to characterize deep-level semantic information. The emergence of deep learning has solved the above problems very well. Compared with machine learning, the advantages of deep learning can be simply divided into three aspects:

1. State-of-the-art performance: Deep networks have achieved accuracies far exceeding classical ML methods in many domains, including speech, natural language, vision, games, and more. In many tasks, classical ML methods are not even comparable to deep learning.

2. Efficient scaling with data: Compared to classical ML algorithms, deep networks can scale better with more data.

Although the deep neural network is applied to the task, it solves many problems and improves the performance of the task [3]. However, there are very few reviews on object detection so far. This

leads to many beginners being unable to know the latest progress of object detection and what new methods have been proposed when they come into contact with object detection. In addition to this, a review can be seen as a staged summary that facilitates the development of follow-up research-based work. Therefore, we have sorted out and summarized the development of this object detection in recent years. We first distinguish the definition of object detection, and then introduce the commonly used datasets and evaluation metrics for tasks. In addition, we also explain the mainstream deep learning methods used for the task. Finally, we introduce some state-of-the-art methods in object detection and potential challenges in the task.

2. Background

2.1. Definition

Object detection has two purposes, one is to identify what several objects are in the detected image, and the other is to detect the specific positions of several objects. As a basic aspect of image classification, image segmentation, etc. in computer vision, object detection has a very broad application.

Pedestrian detection technology has a strong value in use [4]. The main methods of pedestrian detection are two types of schemes using artificial features + classifiers, and deep learning schemes. The classifiers used are Linear Support Vector Machine, AdaBoost, Random Forest. Next, we focus on convolutional network-based schemes.

The first method is Cascade CNN. Angelova et al. [5] proposed a scheme for pedestrian detection with cascaded convolutional networks, which borrows the idea of cascaded AdaBoost classifiers. The previous convolutional network is simple and can quickly exclude most of the background regions. The latter convolutional network is more complex and is used to accurately determine whether a candidate window is a pedestrian. Through this combination, the detection speed is greatly improved while ensuring the detection accuracy. This approach is similar to Cascade CNN in face detection.

Ouyang and Wang [6] used a hybrid strategy to train a convolutional neural network pedestrian classifier on the Caltech pedestrian database. The classifier is used in the last stage of pedestrian detection, that is, the final screening of the final candidate area, because the efficiency of this process is not enough to support exhaustive traversal detection such as sliding windows. The author uses HOG+CSS+SVM as the first-level detector for pre-filtering, and then uses the convolutional neural network to further judge its detection results. This is a coarse-to-fine strategy. The input of the convolutional network is not the image of the RGB channel, but the three channels given by the author's experiment. The first channel is the Y channel in the YUV of the original image, and the second channel is divided into four blocks. The priority is U channel, V channel, Y channel and all 0s; the third channel is the edge of the second channel calculated by the Sobel operator.

In addition, the component detection strategy is adopted. Since each component of the human body has different sizes, the author designs different convolution kernel sizes for different components. As shown in Fig.1, Level1 is for relatively small components, Level2 For medium-sized parts, Level3 is for large parts. Due to the existence of occlusion, the author designs several occlusion modes at the same time.

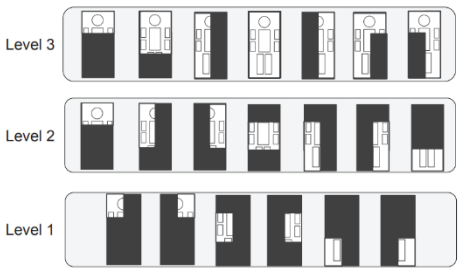

Fig. 1 Three components

Face recognition is essentially the target detection of a single object, which mainly includes four parts, namely, face image acquisition/detection, face image preprocessing, face image feature extraction, and identity matching and recognition. Traditional methods, such as methods based on HOG features, are far less accurate than the current mainstream methods based on deep learning and convolutional neural networks [7]. Among the methods based on DNN, the method based on candidate regions represented by R-CNN, although it has high accuracy, it is still lacking in running speed. The single-shot target detection model based on regression thinking represented by YOLO has better efficiency and can be better by changing the scale of the model. Trade-off speed and accuracy, suitable for the development of real-time systems. There are also three models in face recognition: DeepFace, DeepID and facenet.

DeepFace is the earliest algorithm model that applies CNN to face recognition. It can be seen as the transition of face recognition from traditional methods to deep learning-based methods. DeepFace adopts the process of detection, alignment, extraction, and classification. Deep learning face recognition methods also follow this process. DeepFace adopts a 3D alignment method for face alignment. First, the LBP histogram is used to texture the image, then the corresponding features are extracted, and then the face is 3D modeled according to the facial key points to generate a 3D face. The alignment method adopted by DeepFace is much more complicated than the alignment method used by the deep learning-based face recognition algorithm proposed after it. Experiments show that DeepFace can achieve more than 97% recognition accuracy on the LFW dataset, but due to the increased computational complexity caused by 3D alignment, its speed is about 5 images per second, and due to the classifier structure, the model is different in the input data must be retrained to ensure accuracy, which greatly limits its usability.

As a face recognition model, DeepID also adopts the same process of detection, alignment, feature extraction, and classification as DeepFace. DeepID generates 400 patches when preprocessing the image, and the work of aligning all of them for feature extraction greatly increases the workload of the model, so in DeepID2, the author uses the forward-backward algorithm to select 25 of them. It can be used for feature extraction, which greatly reduces its workload. The author also proposed DeepID2 which structure is basically the same as DeepID, but did not solve the problem of model generalization, and finally did not use Soft-max for classification, but trained multiple Bayesian classifiers, and finally merged into one classification through SVM. device.

FaceNet is a face recognition model proposed by Google. The model does not train a classifier, but extracts feature by calculating the Euclidean distance of feature points, and then sets a threshold to determine whether it is the same face. Its structure is shown in the figure. , the first half is a normal convolutional network, but a triple pair is introduced to construct before calculating the loss function.

2.2. Datasets

Fire and Smoke Image Dataset [8]: This dataset consists of a dataset of images of early fires and smoke. The dataset consists of early fire and smoke images captured with mobile phones in real scenes. There are about 7000 image data. Images were taken under various lighting conditions (indoor and outdoor scenes), weather, etc. This dataset is ideal for early fire and smoke detection. The dataset can be used for fire and smoke identification, detection, early fire and smoke, anomaly detection, and more. The dataset also includes typical household scenarios, such as waste incineration, paper and plastic incineration, field crop incineration, home cooking, etc.

AITEX dataset: The database consists of 245 4096 x 256 pixel images of seven different fabric structures. There are 140 defect-free images in the database, 20 for each type of fabric, in addition to 105 images of different types of fabric defects (12 defects) commonly found in the textile industry. The large size of the image allows the user to use different window sizes, thereby increasing the number of samples.

T-LESS dataset: The targets collected in this dataset are industrial applications, targets with few textures, and lack of distinguishing colors, and the targets have symmetry and cross-correlation. The dataset is obtained by three synchronized sensors, one structured light sensor, one RGBD sensor, a

high-resolution RGB sensor, obtained 3.9w training set and 1w test set from each sensor, in addition, 2 3D models were created for each target, one was handcrafted by CAD and the other was semi-automatically reconstructed. The backgrounds of the training set pictures are mostly black, while the test set pictures have a lot of background changes, including different lighting, occlusion, etc.

2.3. Evaluation Metrics

There are three aspects to the general evaluation of target detection: Assuming that the classification target is two classifications, which are positive samples (Positive, abbreviated as P) and negative samples (Negative, abbreviated as N).

Accuracy is the proportion of all predictions that are predicted correctly:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

Precision is defined in terms of predicted outcomes. Precision is also called precision. It should be noted that Precision and Accuracy are not the same:

$$\mathrm{Pre} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{2}$$

Recall refers to the probability of correct identification among all positive samples:

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

3. Recent research

Literature [9] introduced UR3D, a single stage and multi-scale framework that can learn a unified representation of objects within different distance ranges for monocular 3D object detection, which leads to a compact and robust network, which can significantly reduce the cost of network capacity generated by simply learning the feature covering all possible location. Also, the unified object representation reduces the number of learnable parameters and thus prevents overfitting. There are two vital challenges for monocular 3D detection: omissions of candidate boxes with high-quality 3D information prediction and Recovering object physical size.

The solution to the first issue is a new distance-guided NMS, which automatically selects candidate boxes with better distance estimations and thus make UR3D achieve better distance estimate and 3D accuracy. The solution to the second issue is a new fully convolutional cascaded point regression to estimate the projected 2D center points and corner points of 3D boxes accurately and efficiently. By utilizing the new method, the authors used the predicted key points to post-optimize the physical size and orientation by minimizing a projection-consistency loss.

Literature [10] consider that the significant difference between normal nature images and drone captured images makes the object detection extremely challenging because of visual differences and objects occlusion caused by density. By compare the two-stage detector, one-stage detector and anchor-free detector, the author improved the anchor-free detector by mixing it with a re-regression module. The hybrid detector firstly predicts the center point and the width and height of each object instead of using the anchor box with two detectors. Secondly, the authors transform these center points and sizes to coarse bounding boxes. Finally, they feed the deep feature maps and the coarse bounding boxes into a re-regression module to adjust the coarse bounding boxes and generate the final accurate bounding boxes.

Literature [11] aim to transfer label-rich source domains to label-independent target domain, while sufficiently mixing the source and target domain. They created an ATF, containing 3 streams upon the backbone. The first two streams with shared parameters are the Chief net trained by ancillary data guided domain adversarial loss and source label guided detection loss.

Another stream is the ancillary net which is independent of the chief net to better characterize the domain-invariant detector and learn the ancillary target features, enhancing the transferability of Faster-RCNN. The asymmetry originates from that the ancillary net is independent of the parameter shared chief net. Because the independent ancillary net is only trained by the labeled source data, the asymmetry can largely avoid source collapse and feature distortion during transfer. Their model inclines to preserve the discrimination of source features and simultaneously guide the structural transfer of target features.

Literature [12] directly fine-tuning the model on the data of new classes will severely decrease the performance of old classes, which is known as catastrophic forgetting. It is necessary to avoid catastrophic, aiming to improve the robustness of object detectors in the open world by enhancing the ability of object detector to learn new object classes continuously. To preserve the significant knowledge, the authors propose a novel multi-view correlation distillation with selective features based incremental object detection method (MVCD), which mainly focuses on the design of distillation losses and the feature selection in the feature space of the object detector. To get a good trade-off between the stability and the plasticity of the incremental model, this paper design correlation distillation losses for the sample-specific selective features from three views (channel-wise, point-wise and instance-wise) for regularizing the optimization of the object detector, which is also a new standard to estimate the incremental learning performance of incremental object detector.

4. Challenges

Based on the five detection applications of target detection, this paper will correspondingly propose the difficulties and challenges of application detection.

The first is pedestrian detection in traffic detection, in which the difficulties and challenges can be summarized into 4 points: small-sized pedestrians, difficult negative samples, dense pedestrian occlusion and overtime detection. Small-sized pedestrians refer to pedestrians whose height is less than 30 pixels, which is not easy to detect and features are difficult to extract. Even if the Fast/Faster RCNN algorithm was used to perform SOTA at the time, its detection effect on small targets was not ideal due to the low-resolution features extracted by the detection head. The second problem is difficult negative samples. In this case, because some backgrounds in the scene image are visually very similar to pedestrians, it is difficult for the network to correctly classify negative samples, and some background objects that are very similar to pedestrians are often predicted as pedestrians. pedestrian. The third problem is that there are many dense and occluded pedestrians in the image, which makes it difficult for the network to accurately detect occluded pedestrians. In the Caltech dataset, according to statistics, unconcluded pedestrians account for about 29% of the total number of pedestrians. The fourth problem is timeout detection. Since some applications such as autonomous driving and video surveillance require detection algorithms to provide detection results in real time so that the system can make quick decisions, real-time pedestrian detection in video is critical.

The traffic signal detection in traffic detection has also attracted much attention in recent years. With the development of autonomous driving technology, automatic detection of traffic signs and traffic lights has attracted much attention in recent years [13]. Sign detection in fixed scenes such as traffic lights and traffic signs still have considerable challenges. The problem lies in the following four points: lighting changes, blurred traffic signs, weather changes, and real-time pedestrian monitoring. Many navigation apps will often update the status of road signs in various places. If the existing traffic sign data in the navigation software can be used to enhance the real-time detection of traffic signs, it should be greatly improved.

The second challenge is the problem of face recognition. In China, the payment method for many commodities has adopted face payment, that is, using facial recognition to open an online wallet and then pay. Apple phones also turned-on Face ID in 2017 for unlocking phones and paying in stores. However, there will be a lot of inconvenience due to the huge changes in the facial posture of the human face. Since there may be various changes in the human face, such as changes in expression,

skin color, posture, and motion, this may lead to unsuccessful object detection. Another problem is that the target is occluded, and face recognition cannot quickly and accurately identify a person with only partial features.

The third challenge is text detection. Text detection now has no problem converting standard fonts. But there are still four difficulties that have not been overcome. 1. Large differences in fonts and languages 2. Text rotation and perspective changes 3. Dense text 4. Missing or blurred fonts. The current method to improve text rotation and perspective is in the Anchor box, and introducing additional parameters by rotating the ROI with Arbitrary-oriented scene text detection via rotation proposals, but this heavily relies on manually designed anchors, when dealing with artistic fonts or irregular text difficulties still exist. Most current end-to-end text recognition models use RPN to generate proposals, and there is a disadvantage: proposals are axis-aligned rectangles, and in the case of dense text, multiple adjacent text instances are usually included in one proposal, Recognition will be affected.

The fourth challenge is remote sensing object detection. In recent years, with the improvement of remote sensing image resolution, remote sensing image target detection, such as aircraft and ships, has become a research hotspot. Remote sensing image target detection has a wide range of applications, such as military reconnaissance, disaster rescue, urban traffic management and so on. The problem of remote sensing target detection mainly lies in the large difference in resolution of images and the adaptation of different domains. Domain adaptation has been partially addressed in this paper.

5. Conclusion

This paper introduces the development process of traditional target detection algorithm to target detection algorithm based on deep learning and introduces three technical routes based on CNN target detection algorithm development: one-stage, two-stage and anchor free detection algorithm. In the future, object detection will continue to develop with the improvement of the algorithm in many aspects such as field adaptation and small target detection.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[8] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[9] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[10] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[11] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[12] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[13] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.