Research on Target Detection Methods under the Concept of Deep Learning

Hang Xu¹, Zhigong Huang¹

¹College of Electronic Engineering, Guangxi Normal University, Guilin, China

Abstract: As a classic subject in the field of image processing and computer vision, target detection has a wide range of applications in traffic monitoring, image retrieval, human-computer interaction and so on. It aims at detecting objects of interest in a static image. In view of the strong expressive ability of convolutional neural networks in deep learning, this paper presents the classical detection framework R-CNN of deep learning. Based on the above detection framework, the functional requirements, such as data pre-processing, training model and image prediction, as well as the non-functional requirements of the target detection system are analysed. According to the above requirements, a target detection system based on deep learning is developed. Practice has proved that the system has good performance in terms of hardware and performance.

1. Introduction

In the current stage, image has become the most important data information in people's lives, and image processing has become a hot topic in recent years. As the core research direction in the field of image processing and computer vision, target detection exists in all walks of life for its wide application needs. We not only pay attention to the simple classification of the image, but also hope to obtain the target and location of the interest in the image accurately and apply the information to a series of realistic tasks such as video surveillance and autonomous driving, so the target vision detection technology has been widely used. Target vision detection has great practical value and application prospect. The application fields include intelligent video surveillance, robot navigation, automatic positioning and focusing in digital cameras, aircraft aerial photograph or satellite image detection, and obstacle detection in vehicle camera images. At the same time, target vision detection is also an important prerequisite for many high-level visual processing and analysis tasks. The behaviour analysis, event detection, scene semantic understanding and so on require the use of image processing and pattern recognition technology to detect the objects in the image, determine the semantic types of these objects to the image, and mark the target pairs.

Traditional target visual inspection often meets the following challenges. The discrepancy between class and class is very different for many objects. There may be great differences in colour, material, shape and so on. It is difficult to train a feature description model that can contain all kinds of changes in the class. In addition, there may be great similarities between different types of objects, and even non-professionals can hardly distinguish them from appearance. Intra class differences may be large, and inter class differences may be very small, which poses a challenge to target vision detection. In the process of image acquisition, in the process of image acquisition, due to the difference of environment, light, weather, shooting angle and distance, the non-rigid body shape of the object itself and the possibility of being partially blocked by other objects, the apparent features of the object in the image are very diverse, and the robustness of the vision algorithm is very high. The computational complexity
of computational complexity and adaptive target vision detection mainly comes from the number of target types to be detected, the dimension of feature descriptors and the acquisition of large scale tagged data sets. Because there are many target types in the real world, each type contains many images, and the recognition of each type requires a lot of visual features, which leads to the feature description of high dimensional space sparsity. In addition, the target model is often learned from large scale tagged data sets. In many cases, data acquisition and annotation are difficult, and a large amount of manpower and material resources are needed. These situations lead to very high computational complexity of target detection, and efficient target detection algorithms need to be designed. At the same time, in the dynamic environment, to improve the accuracy of target detection, it is necessary to explore the appropriate mechanism to automatically update the visual model and improve the adaptive ability of the model to the complex environment. To overcome these challenges, this paper improves the traditional target detection algorithm, and gives the target detection method under the concept of deep learning.

2. Theory exploration of target detection methods under the concept of deep learning

2.1. Artificial Neural Network (ANN)
The artificial neural network abstracts the neural network of human brain from the angle of information processing, establishes a simple model, and makes up different networks according to the different connection modes. In engineering and academic circles, it is often referred to as neural network or artificial neural network. Neural network is an operation model consisting of many nodes connected to each other. Each node represents a specific output function, called the excitation function. The connection between every two nodes represents a weighted value for the connection signal, called weight. The output of the network varies according to the way of network connection, the weight value and the excitation function. The network itself is usually an approximation to some algorithm or function in nature, or it may be an expression of a logical strategy. According to different learning environment, the learning mode of neural network can be divided into supervised learning and unsupervised learning. In supervised learning, the data of the training samples are added to the network input, and the corresponding expected output is compared with the network output, and the error signals are obtained. To control the adjustment of the strength of the weight connection, it converges to a determined weight after many trainings. When the sample situation changes, learning can modify the weight to adapt to the new environment. The neural network model of supervised learning includes backpropagation network, perceptron and so on. In unsupervised learning, a standard sample is not given in advance, and the network is placed directly in the environment, and the learning stage and the working stage are integrated. At this point, the change of learning rule obeys the evolution equation of connection weight. The simplest example of unsupervised learning is the learning rule. The competitive learning rule is a more complicated example of unsupervised learning. It adjusts weights according to the established clustering.

2.2. Convolutional Neural Networks (CNN)
Convolution neural network model is a very successful algorithm for deep learning, and it can train the algorithm of multilayer network structure. The peculiarity of convolution network is that its connection between elements is not fully connected, and the weights of connections between some neurons are shared. The neural units between two adjacent layers in CNN are partially connected, that is, the perception part of a neural unit is a part of the neural unit of the upper layer. CNN has four very important ideological structures: local area perception, weight sharing, down sampling and multiple volume layers. Local area perception means that some local features of the data can be found, such as a corner and an arc in a picture, which are the basis of animal vision. All pixels display a mess of dots, and the relationship between pixels is not being excavated. Each layer in CNN is made up of a lot of Feature map, each of which is composed of multiple neural units, and a convolution kernel is shared by all neural units of the same map. The convolution kernel is actually a weight value. It does not need to calculate a convolution alone but uses a fixed weight matrix to match the image. This operation is like convolution, so it is called a convolution neural network. In fact, BP can also be regarded as a special convolution
neural network, but the convolution kernel is the ownership value of a certain layer, that is, the perceptual region is the whole image. The weight sharing strategy reduces the parameters needed to be trained, making the generalization ability of the trained model stronger. CNN networks usually use convolution layer and lower sampling level alternately, that is, a convolution layer relates to a lower sampling level, and the next sampling layer is followed by a layer of layer. The convolution process is a training convolution kernel to convolution an input image, and then add a bias.

2.3 Region Convolutional Neural Networks (RCNN)
In 2014, Professor RossB. Girshick designed the R-CNN framework, which made the target detection a great breakthrough, and opened the upsurge of detection based on the deep learning target. The RCNN framework mainly solved two problems: first, the classical target detection algorithm uses the sliding window method to determine all possible region RCNN in advance to extract a series of more available. It can be the candidate area of the target, then only extracts the features on these candidate regions, and the two is the feature extracted by the classical target detection algorithm in the region. RCNN uses the training convolution neural network for feature extraction and extracts regional candidate frames from the original image through the original picture search method. First, we use an over segmentation method to divide the image into a small area, and view the existing small area, and combine the two regions with the smallest area after the close texture of the similar colour. To ensure that the consolidated regional candidate frame scale is uniform and shape rules first subtracts the proposed frame pixels from the average value of the proposed frame pixel, and then each proposed frame is input to the CNN network. The classification and boundary regression contain two sub steps, which are the classification of the output vector of the last step and to get the exact target area through the boundary. Because the actual target will produce several sub regions, the aim is to locate and merge the foreground target of the classification accurately and avoid multiple detection. Each proposed frame is a score of the object category. Some of the best points in the class have the highest score, and regression machines are used to return the remainder of the categories, and the final bounding box is obtained, which has the highest score after each category.

3. System requirements of target detection methods under the concept of deep learning

3.1 Functional requirements
The function of target detection is integrated by data pre-processing, training model module and image prediction module. The functional requirement of image data pre-processing module is to pre-process the image data with good annotation. For example, scaling, averaging, down sampling, whitening, and so on. The pre-treatment of training data is of great help to model training, such as accelerating the convergence of models and improving the accuracy of models. The function requirement of the data collection and annotation module is to manually annotate each frame collected by the camera, mark out the information of the object's category and position in each frame, and use it as the label of the training set, and participate in the training of the model. Since data plays a vital role in the accuracy of training models, this part has high requirements for image annotation quality. This system implements the interface of commonly used image data pre-processing operation. The function requirement of the image data reading module is to read the labelled and pre-processed data into the trained neural network when the model is trained. The module is to complete the fusion of the depth learning framework used in the detection system. The core function of the training module of the detection model is to use a very classic detection framework to train the target detection model suitable for the current scene. Finally, it can predict each frame of the camera, give the object's class and position coordinates of the current frame, and have good performance in the precision and speed. The core function of the target detection module is to load the trained model, and then get each frame collected by the camera by reading the data module, and input it into the current trained detection model, and output the coordinates and categories of the target object in the current frame. By reading the data module, we get each frame collected by the camera.
and input it into the currently trained detection model to output the coordinates and categories of the target object in the current frame.

3.2 Non-functional requirements
Non-functional requirements include performance, modifiable, usability, ease of use, security, environmental requirements, and so on, which are easily ignored in the initial phase of the requirement analysis. The consideration of these content often affects the implementation of the product.

The target may appear in any position of the image, and the size and aspect ratio of the target are also uncertain. The most commonly used method is to use sliding window strategy to traverse the test samples. For accuracy, different scales and different aspect ratios are required. Although the exhaustive algorithm contains all the locations of the targets to be detected in the image, the disadvantages are obvious. Traversing takes too much time and generates redundant candidate regions too, which seriously drag down the speed and performance of subsequent feature extraction and classification. Under the influence of time complexity, the longitudinal and transverse ratio of the sliding window is usually fixed number of fixed values, so for the multi class target detection with larger vertical and horizontal ratio, even if the sliding window is traversed, it cannot get a good area. Since the detection system is based on deep learning, it takes more time to train the model than the traditional machine learning and requires more hardware. It usually takes a day or even longer to make good use of the GPU training model. This is mainly because convolution neural networks contain more parameters than traditional algorithms. Therefore, this requires that in the training model and the use of the model to predict when the precision is available, the speed can be achieved in real time, and can really fit the requirements of the automatic driving scene. The system must be close to the actual application process, as far as possible to meet user's operational habits. At the same time, the computer level of different operators should be taken into consideration, and the usability of the system should be emphasized in the design of the system so that most users can use the system better.

The amount of data to be processed is different, and the design of the technical level will be greatly affected. The requirement analysis phase can help the technical personnel to consider the outline design. Due to the complexity of the scene of automatic driving in real life and the change of outdoor weather, it will directly affect the current road conditions. The same city has different road conditions at different times and different locations, and there are also great differences between different cities. There is even greater difference between the road conditions at home and abroad, which have high requirements for the robustness of the system. The detection system can predict the type and location of objects in the current driving scene under different weather conditions in different cities. The process of input image acquisition will be affected by various noises and reduce the image quality. The filtering algorithm can solve these problems better. The basic idea of filtering algorithm is to process the pixels in the image and the surrounding pixels through the template and adjust the details of the noise to improve the quality of the image. Mean filter is the simplest linear filter. The main method used is neighbourhood average method. The basic principle is to calculate the mean value of all pixel values of each pixel value in its local neighbourhood instead of the original pixel value of the original image. The mean filter algorithm is simple and the calculation speed is fast.

4. System implementation of target detection based on deep learning

4.1 Data pre-processing module
Because of the diversity of various targets, diversity of illumination, and the diversity of background, it is very difficult to design a robust feature by hand in dealing with all applications. Therefore, for different application scenarios, using different features will directly affect the accuracy of detection. Due to the differences between training and testing stages, the system adopts different strategies. During training, Faster R-CNN detects the relevant interfaces provided by the framework, such as scaling and averaging. At the time of testing, the system did not use the Python version of Faster R-CNN. Compared with the pre-processing of the test, the training phase was more redundant. Instead, it is based on Caffe
to implement its own data conversion function, Transform, to read the data and input it into the detection network, which will be explained in detail later. In the training model, the system packages the labeled data into lmdb format, because the data in this format is read more efficiently. Lmdb is an embedded storage engine developed by open ldap project, which is based on file mapping and key-value interface. All of the operations of the lmdb are to load the access text into the address space of the process by map, directly access the address, and reduce the time overhead of the hard disk and user address space. The other is leveldb, which is also based on key-value. Although the memory consumption of lmdb is one point one times that of leveldb, the speed of LMDB is much faster than that of leveldb. Therefore, the system used lmdb format to package data in the training model. Because of the diversity of various targets, diversity of illumination, and the diversity of background, it is very difficult to design a robust feature by hand in dealing with all applications. Therefore, for different application scenarios, using different features will directly affect the accuracy of detection. The GLOH feature description operator is also an extension of the SIFT feature and is transformed into log polar coordinates. The expression of the concentric circle. Therefore, the above three characteristics can be effectively applied to object recognition, robot map perception and navigation, image mosaic and 3D modeling. HoG features are mainly used for pedestrian detection in images, but in the future development, they also apply to pedestrian detection in dynamic video, as well as the detection of various kinds of traffic tools and animals in static images. It transforms image representation into histogram distribution of texture images, and texture images can be obtained through a set of filters, or neighbourhood brightness. Features are generally used for texture classification, that is, image segmentation. SIFT can extract unique key points. This key point will not be changed due to external conditions such as movement, rotation, scaling and affine transformation. RIFT descriptor operator is a description operator with rotation invariance.

4.2 Training model module
The computational complexity of target visual detection mainly comes from the number of targets to be detected, the dimension of feature descriptors and the acquisition of large scale tagged data sets. Because there are many target types in the real world, each type contains many images, and each type requires a lot of visual features, which leads to high dimension. In addition, the target model is often learned from large scale tagged data sets. In many cases, it is difficult to collect and annotate data, which requires a lot of manpower and material resources. These results lead to high computational complexity of the target detection, and a high efficient target detection algorithm should be set up. To improve the precision of target detection, it is necessary to explore the appropriate mechanism to automatically update the visual model and improve the adaptive ability of the model to the complex environment. To overcome the above challenges, many target vision detection algorithms have been proposed. They are proposed in the target area, image feature representation, candidate region classification and so on. Different processing strategies are adopted. In recent years, with the development of deep learning technology, many target visual detection methods based on deep learning have been put forward in succession. It is obviously better than the traditional method in precision and has become the latest research hotspot. The RBG open source detection framework uses VGG16 and ZF as the feature extraction layer for the detection framework. Because ZF is more streamlined than VGG16, the speed is faster, but the precision is not so accurate as VGG16. In view of this, a network structure is a compromise method to improve the speed as much as possible when the precision is acceptable. An attempt was made to remove a portion of the coiling layer, reducing the number of the convolution cores of the remaining coiling layers to 1/2, and then adding a layer of BN to the front of the last coiling layer. Caffe is a clear and efficient framework for deep learning. Caffe is a pure C++/CUDA architecture that
supports command lines and provides interfaces between Matlab and Python. It can seamlessly switch between CPU and GPU. The Caffe framework is mainly composed of four large components. Then we analyse some of the operations of the Net class. This class mainly realizes the forward and backward propagation of convolution neural network during training. The Forward Backward function calls the Forward and Backward functions to carry forward and reverse transmission of the network. The Forward From To function performs forward forwarding from the start layer to the end layer. The Backward From To function implements the propagation from the start layer to the end layer. The To Proto function mainly implements the serialization of neural networks to disk files.

4.3 Training model module

As the application scene of this system is autopilot, the final test is the traffic video, which is predicted by reading the front of the car to the single camera, which is captured in each frame of the current vehicle video. This part of the video is read by using the Video Capture class, which implements the method of reading video and reading the camera. Because of the test on the computer, we collected the driving video in advance. After loading the model and reading the video, the forward prediction for each frame in the video is followed by an encapsulated function for the forward operation of each frame read, and the object category and location information of each frame is output. Before designing the final user interface, to see the visual results easily, we get the category of each candidate frame based on the detection mode, and then use the function to draw different or different colours of the different object categories. Classifier is an important basis for discriminating the target and non-target. In the traditional target detection, after extracting the feature, the classifier is put into the classifier to get the training model, and the model is used to judge the test image. The full name of the classifier is support vector machine. The classifier is mainly used in the classification problem. Unlike the current popular target detection framework using the candidate region, the framework does not need to extract the candidate region information before training, only the true value of the target position is needed, after sending the sample to the underlying network to output the deep convolution layer, and when this true value is determined, we can lose the loss. Function and backpropagation are applied to the end-to-end framework. During training, we need to establish the correspondence between the position truth information and the candidate frame generated by the convolution feature layer in the multiscale. For the true value information of each target, we choose the candidate frame with the highest coincidence degree by selecting the candidate frame of different location, width to height ratio and the scale. This is to get more accurate candidate areas. By size design, step size and Pooling layer, the size of the convolution neural network is smaller after the convolution layer in the deeper convolution layer. It can not only reduce the amount of computation and memory, but also ensure the translation and scale invariance of the feature, to obtain different targets. Scale, some algorithms transform the samples into different scales, and then use convolution neural network to deal with them. In fact, using the convolution feature mapping of different layers in the same network framework can also achieve the same effect, and can also share parameters in the scale of all targets. Therefore, we use the convolution layer under the same framework to detect. The algorithm is essentially a linear classifier with the largest interval in the feature space. The blue circle and the red rectangle represent different samples respectively. The goal of the classifier is to find an optimal classification line that separates the two samples from the maximum interval. Caffe
provides a complete set of tools that can be used for model training, prediction, fine tuning, and good automated testing.

Figure 3. Flow chart of Image prediction

5. Conclusions
Based on the concept of deep learning, the specific methods and system design of target detection are presented in this paper. The main conclusions are as follows:

(1) This paper analyzes the similarities and differences between Artificial Neural Network and Convolution Neural Network and gives the classic framework RCNN of target detection based on deep learning.

(2) This paper explores the system requirements of the target detection based on deep learning, including functional requirements (data pre-processing, training model and image prediction) and non-functional requirements (usability and robustness).

(3) The design and implementation of data pre-processing module, training model module and image prediction module are introduced in detail in this paper.

References
[1] Fraser S, Nikora V, Williamson B J, et al. Automatic active acoustic target detection in turbulent aquatic environments[J]. Limnology and Oceanography: Methods, 2017, 15(2): 184-199.
[2] Chen C L P, Li H, Wei Y, et al. A local contrast method for small infrared target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(1): 574-581.
[3] Nasrabadi N M. Hyperspectral target detection: An overview of current and future challenges[J]. IEEE Signal Processing Magazine, 2014, 31(1): 34-44.
[4] Willett R M, Duarte M F, Davenport M A, et al. Sparsity and structure in hyperspectral imaging: Sensing, reconstruction, and target detection[J]. IEEE signal processing magazine, 2014, 31(1): 116-126.
[5] Chen X, Guan J, Liu N, et al. Maneuvering target detection via Radon-fractional Fourier transform-based long-time coherent integration[J]. IEEE Transactions on Signal Processing, 2014, 62(4): 939-953.
[6] Li X, Cui G, Yi W, et al. Coherent integration for maneuvering target detection based on Radon-Lv’s distribution[J]. IEEE Signal Processing Letters, 2015, 22(9): 1467-1471.
[7] Hu Y, Dong M, Ota K, et al. Mobile target detection in wireless sensor networks with adjustable sensing frequency[J]. IEEE Systems Journal, 2016, 10(3): 1160-1171.
[8] Han J, Ma Y, Zhou B, et al. A robust infrared small target detection algorithm based on human visual system[J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(12): 2168-2172.
[9] Geng X, Ji L, Sun K. Clever eye algorithm for target detection of remote sensing imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 32-39.
[10] Zhang L, Zhang L, Tao D, et al. Sparse transfer manifold embedding for hyperspectral target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(2): 1030-1043.
[11] Du B, Zhang L. Target detection based on a dynamic subspace[J]. Pattern Recognition, 2014, 47(1): 344-358.

[12] Zhang R, Gao M, Bai S, et al. A fluorescent light-up platform with “AIE+ ESIPT” characteristics for multi-target detection both in solution and on paper strip[J]. Journal of Materials Chemistry B, 2015, 3(8): 1590-1596.

[13] Razakarivony S, Jurie F. Vehicle detection in aerial imagery: A small target detection benchmark[J]. Journal of Visual Communication and Image Representation, 2016, 34: 187-203.

[14] Zhang L, Zhang L, Tao D, et al. Hyperspectral remote sensing image subpixel target detection based on supervised metric learning[J]. IEEE transactions on geoscience and remote sensing, 2014, 52(8): 4955-4965.

[15] He Y J, Li M, Zhang J L, et al. Small infrared target detection based on low-rank and sparse representation[J]. Infrared Physics & Technology, 2015, 68: 98-109.

[16] Yu W, Wang Y, Liu H, et al. Superpixel-based CFAR target detection for high-resolution SAR images[J]. IEEE Geoscience and Remote Sensing Letters, 2016, 13(5): 730-734.

[17] Wei Y, You X, Li H. Multiscale patch-based contrast measure for small infrared target detection[J]. Pattern Recognition, 2016, 58: 216-226.

[18] Paulus A S, Melvin W L, Williams D B. Multichannel GMTI techniques to enhance integration of temporal signal energy for improved target detection[J]. IET Radar, Sonar & Navigation, 2016, 11(3): 395-403.