Research on Target Detection Technology Based on Convolutional Neural Network

Liu Yang* and Huaguang Zhang

Department of Information Science and Engineering, Northeastern University, 1-31, Building 1, 4th Phase, Capital Guanghe city, 26-101, Shenzhen Street, Hunan New District, Shenyang City, Liaoning Province, China
E-mail: yangliu860818@126.com

Abstract. The research and application of neural network has been a research topic in recent years, and the efficient recognition of objects is undoubtedly an important part of the technology application. Convolution of the neural network is in the process of object recognition classification, has carried on the improvement of cognitive machine, thus avoiding the error back propagation in time loss problem, based on the biological neural network structure, convolution weights of neural network are more inclined to study sharing network structure, which shows good performance in the process of a number of image recognition, for translation, proportion to gather scaling and deformation and so on the many kinds of complex operation has certain robustness.

Keywords. Convolutional neural network; target detection; cognitive machine.

1. Instruction

In dealing with images of high dimension input, make different neurons and various types of neurons in a layer of complete before full connection is impossible, convolution, on the basis of neural network to reduce the number of parameters, can fully use of local awareness to sparse connection, to imitate biological neural network solves the target image at the same time the work of testing to identify complex problems, in this application, only need to the local image convolution neural network sensory processing, through iteration synthetical consolidation and rehabilitation in a higher level and local information, so as to realize global information gathering effect, in practice to strengthen the practicability of target detection.

Overview of Convolutional Neural Network: As a network framework for deep learning, convolutional neural network takes images as input elements and omits the tedious feature design links in traditional detection algorithms. Figure 1 shows a traditional target detection process. Convolutional neural networks with fewer pretreatment to obtain the effective raw data, said its passage value sharing, on the basis of the guarantee convolution kernel weight constant convolution, makes an image at different positions of the same characteristics is a convolution kernel extracted as possible, while reduce the complexity of the algorithm, makes the quantity and the network connection number was lower, the capacity of the model also effectively promoted, let CNNs closer to biological neural networks [1].
1.1. Basic Network Structure
The special structure of the convolutional neural network is that it adds the volume base and pooling layer. The other input layer, full connection layer and classifier are the same as the traditional neural network. Its basic structure is shown in figure 2.

The input image convolutional layer C1 pooling layer S2 convolutional layer C3 pooling layer S4 output layer [2].

1.2. The Core of the Training of Convolutional Neural Network
As the core of the training of convolutional neural network, the learnable parameters are divided into several groups existing in the basic level of the volume. One group of images performs the convolution operation through the low-dimensional convolution kernel matrix at the input layer, and then USES the activation mapping function to check the input features and perform the convolution to obtain the new feature image:

\[
x_{i}^{l-1} = f(u_{i})\\
\hat{u}_{i} = \sum_{j=0}^{M_{i}^{l-1}} x_{i}^{j-1} w_{ij} + b_{i}
\]

Equation (1) to activate the new pixel point calculation, in which \(f(.)\) is the activation function, it usually choose ReLU function, such as * as the convolution operation, \(x_{i}^{j}\) represents the convolution layer l features in the image in the number j a pixel values of the output channel, \(W_{ij}\) as the convolution kernel, \(M_{i}\) represents the characteristics of a layer image of \(u_{j}^{l-1}\) as a subset of the calculation, \(b_{i}\) is the offset of the l layer [3].

After pooling connected in general to convolution layer by reducing the amount of data to achieve the purpose of dimension reduction, pooling layer in highly abstract image feature extracting have the effect of lower resolution, the characteristics of accessing more spatial invariance, full differentiation
characteristics category at the end of the link layer will integrate the local information, has the characteristics of two-dimensional sexual images in classification layer assignment and converted into a one-dimensional array, and then classified the characteristics of access.

1.3. Basic Operating Principles
The convolutional neural network USES the gradient back propagation algorithm to learn and train the convolution kernel parameters, and propagates the output error layer by layer to minimize the error. To control the overfitting phenomenon, L2 norm formation cost function is added in the sample (x, y):

$$J(W, b) = J(W, b; x, y) + \frac{\lambda}{2} W^TW$$  \hspace{1cm} (2)

where, y represents the label truth value of the sample, $h_{W, b}(x)$ represents the result of sample x after network fitting, \(\lambda\) represents the regularization control intensity, that is, the control parameter of overfitting should be avoided. In order to minimize the cost function \([4]\), the weight $W$ and the bias item $b$ should be updated in real time. The weight of any layer of l is updated to:

$$W^l_i = W^l_i - \alpha \frac{\partial}{\partial W^l_i} J(W, b)$$ \hspace{1cm} (3)

$$b^l_i = b^l_i - \alpha \frac{\partial}{\partial b^l_i} J(W, b)$$

2. Application of Convolutional Neural Network in Target Detection

2.1. Target Detection Model Based on Classification
The target detection model of classification is a kind of target detection model based on anchor points. Selective, Search, Edge Boxes and other algorithms are used to construct the region containing the target to be detected, and then the detection results of possibility labels and location information are obtained through classification and positioning.

2.1.1. R-CNN. R-CNN is the representative of the classification target detection framework, which connects the neural network with the region recommendation algorithm of selective search. Its network structure is shown in figure 3, that is, the region suggestion is generated for the input image by selective search, and then the feature vector in the target region is obtained by means of the convolutional neural network to form the classification \([5]\). Compared with traditional target detection, R-CNN enables the extraction of advanced features on the basis of more universal features, which improves the quality of candidate boundary boxes and effectively reduces the search space.

In practical application, R-CNN also has some problems. First, in terms of time, R-CNN needs to carry out too many repeated operations in image processing, and the useless part of the selective search area still exists, which slows down the processing time and monitoring efficiency of the image. Second, in terms of space, R-CNN USES different measures to process the candidate region, and the data acquired each time needs to be transferred, which leads to excessive occupation of disk space. Thirdly, the fixed scaling of the regional scale results in the loss of local information on the original
image and damages the original information content of the image.

2.1.2. **SPPnet.** Aiming at the problems in the R-CNN, He et al. propose a space pyramid pooling SPPnet testing framework, and it can be connected at the end of the volume at the grass-roots level SPP dimension fixed output [6], as shown in figure 4, the characteristics of any size in a convolution figure input by pooling dimension reduction, mapping feature vector of a fixed size. Due to the reduction of the number of operations, SPPnet improves the operation efficiency and saves the image information completely, but it does not optimize the space occupation.

![Figure 4](image.png)

**Figure 4.** The SPP layer of the base layer and the full connection layer.

2.1.3. **Fast-R-CNN.** Girshick proposed that fast-R-CNN introduces multi-task loss on the basis of SPPnet. As shown in figure 5, the multi-task loss function integrates the target positioning loss into the traditional loss function to modify the location information. It is an end-to-end training method to simplify the pooling layer of space pyramid. In the multi-task learning mode, the operation steps are concentrated in a convolutional neural network. The target classification and the regression integration of bounding box are used to omit the feature conversion steps, and the convolution features of classification and regression are shared, which reduces the space occupation and improves the speed of training and detection. However, as Fast-R-CNN still relies on Selective Search to extract candidate regions in real time, its time loss still exists.

![Figure 5](image.png)

**Figure 5.** Fast-R-CNN network structure.

2.1.4. **Faster-R-CNN.** Aiming at the problems related to fast-R-CNN, BEN proposed the concept of regional recommendation network, that is, integrating RPN and fast-R-CNN to unify the detection structure and realize the weight sharing of convolution features, so as to form the Fay-R-CNN framework. With the help of full convolutional network, RPN slides Windows on the feature graph to make anchor points projected on the original graph, so as to form candidate regions with different proportions. Faster-R-CNN is composed of two parts: one is to generate RPN network of candidate regions, and the other is a fast-R-CNN network used for detecting candidate regions and target recognition. Figure 6 shows the basic structure of faster-R-CNN.
Faster-R-CNN framework, implements the end-to-end integrity for R-CNN in time and space problem to carry on the effective solutions, based on the optimization, but also some deficiencies in the actual operation, specific include: on the one hand, rely too much on the anchor point for the selection of the candidate area, but the anchor because there is a certain problem makes their actual training return in the border areas of performance is not ideal; On the other hand, part of the characteristics of RPN and fast-R-CNN are associated to some extent, which makes it impossible for fast-R-CNN to correct the errors of RPN in time. For example, in the detection of small objects, it is easy to make a mistake to classify the background into the detection area.

![Figure 6. Training process of faster-R-CNN.](image)

### 2.1.5. HyperNet

HyperNet is an improvement based on faster-R-CNN, which fuses generation and detection by combining with convolution features at different scales in the feature extraction network, and forms the concept of superfeature, which reduces the candidate region and optimizes the recall rate, so as to solve the loss of feature resources in faster-R-CNN. Aiming at the problem of small target positioning, HyperNet, which is shown in figure 7, improves the effect of small target detection in the target detection algorithm by fusing it with low-level features.

![Figure 7. HyperNet target detection framework.](image)

### 2.1.6. R-FCN

With Faster-R-CNN all convolution is able to put forward the concept of network, on the basis of through combination with position-sensitive scoring figure, R-FCN model was put forward, and the framework of the relative space position information of the target area for encoding, and makes the translation sensitivity was included in the test network, R-FCN by convolution network, share the testing within the framework of computing, optimized the computing time effectively.

As shown in figure 8, R-FCN can generate fixed k2 position sensitive score graphs of grid for different categories at the end volume base level, and realize the integration of the full connection layer to realize the target detection of the full convolution architecture. In addition, it has good performance in PASCAL VOC and Microsoft COCO, and realizes the detection of each graph of 170 ms.

### 2.1.7. Transformable Convolutional Network

To enhance target detection network in the modeling of geometric transformation ability, improve because of DNNs convolution core and pooling cell shape is
relatively fixed, and the adaptive ability is low, is a combination of shape variable convolution and ROI networks have been proposed, this method improves the detection network modeling ability in terms of geometric transformation.

According to the geometric configuration, the variable convolutional network can take advantage of the additional volume base to carry out the active learning of offsets, so as to improve the free transformation ability of network sampling. Similarly, the deformability of ROI is formed by assigning an offset to the ROI pooling unit, so that the pooling layer can adaptively locate objects of different shapes.

![Figure 8. R-FCN network structure.](image)

2.2. Target Detection Model Based on Regression

Detection algorithm based on target candidate area is basically achieved the goal of real-time detection and end-to-end training pattern, but as a result of training two convolution network at the same time, still make it exist long training time and too many parameters, complex network structure and so on, therefore, based on the simplified flow concept, return target detection model based on the researches on the detection by a network of convolution.

2.2.1. YOLO and Its Improvements. YOLO will detect the picture divided into several regions, integration of the target class prediction and bounding box, return to a convolutional neural network used for detecting a convolution of the network, will detect images using 7 x 7 grid division, the modification of border regression to detect target, through the connection layer to extract information, YOLO testing framework model as shown in figure 9.

![Figure 9. YOLO network structure model.](image)

The structure of YOLO network has effectively improved the detection speed and made it more
suitable for real-time detection. However, there are also problems such as large positioning error, unsatisfactory recognition effect on small targets, and insufficient utilization of feature images. Therefore, the improved versions of YOLOv2 and YOLOv3 based on the YOLO network structure further improve the detection accuracy and realize more rapid and multi-scale target detection.

2.2.2. SSD and Its Improvement. For YOLO intensive multi-scale detection in network structure, SSD target detection framework is proposed, the method based on VGG16 use pyramid structure feature, combined with Faster-R-CNN anchor theory, by adding additional volumes at the grass-roots level, in the end of the VGG can decrease step by step, with different volume base on the sliding window, realize the scale of target detection, but the SSD on multiple adjacent small target detection problem is still inadequate, although has certain increase its speed, but in accuracy compared with detection method based on the candidate regions are not slightly. In terms of improving target accuracy and small target detection, improved methods of SSD such as FSSD and DSSD are proposed.

3. Prospect of the Development Direction of Convolutional Neural Network

It is a development direction of the convolutional neural network to increase the depth by increasing the number of layers. With the increase of convolutional layer in the long-term development, its ability to acquire information is gradually improved. For example, VGGNet has been extended to the network depth of 19 layers, and 3×3 convolution kernel is used in the primary level of each volume. GoogleNet also further optimized the network depth, so that the network performance could be improved under the same calculation amount. In practical application, the detection of small objects is an alternative approach. Due to the excellent feature extractor of convolutional neural network, it performs well in the training process of reducing manual annotation. Depending on the relatively small weight of convolutional neural network, video and 3D target detection will also be widely used.

4. Conclusion

With the application of the convolutional neural network, it is necessary to use the optimized feature extractor to improve the recognition of the model. The special structure of convolution layer and pooling layer makes the application of YOLO network structure more comprehensive. In the process of face recognition, gender discrimination is further avoided. The improvement of target detection accuracy will make the training results more representative.

References

[1] Han Z S, Wang C P and Fu Q 2020 Arbitrary-oriented target detection in large scene SAR images Defence Technology 16 (4) 933-946.
[2] Ya F, Ren D H and Ji H R 2018 Context-augmented convolutional neural networks for twitter sarcasm detection Neurocomputing 308-309.
[3] Sebastian R 2020 DeepMind technologies limited chapter: action recognition in videos using 3D spatio-temporal convolutional neural networks in patent application approval process Journal of Technology 231-234.
[4] Pierre-E M J, Benois-Pineau R and Pé teri J M 2020 Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks Multimedia Tools and Applications 121-123.
[5] Rolf-Dieter K 2020 Information Technology Chapter: Efficient inception V2 based deep convolutional neural network for real-time hand action recognition Network Weekly News 89-92.
[6] Chen C 2020 Signal processing chapter: Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural Network Weekly News 128-130.