Journal of Beijing Institute of Technology, 2020, Vol. 29, No. 2

Object Recognition Algorithm Based on an Improved Convolutional Neural Network

Zheyi Fan✉, Yu Song and Wei Li

(School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: In order to accomplish the task of object recognition in natural scenes, a new object recognition algorithm based on an improved convolutional neural network (CNN) is proposed. First, candidate object windows are extracted from the original image. Then, candidate object windows are input into the improved CNN model to obtain deep features. Finally, the deep features are input into the Softmax and the confidence scores of classes are obtained. The candidate object window with the highest confidence score is selected as the object recognition result. Based on AlexNet, Inception V1 is introduced into the improved CNN and the fully connected layer is replaced by the average pooling layer, which widens the network and deepens the network at the same time. Experimental results show that the improved object recognition algorithm can obtain better recognition results in multiple natural scene images, and has a higher degree of accuracy than the classical algorithms in the field of object recognition.

Key words: object recognition; selective search algorithm; improved convolutional neural network (CNN)

CLC number: TP391.4 **Document code:** A **Article ID:** 1004-0579(2020)02-0139-07

The purpose of object recognition is automatically obtaining multiple objects of interest from the image and identifying the specific classes. Object recognition is the basis of realizing scene semantic understanding and video multi-object tracking. However, object recognition is a challenging task because of occlusion, illumination changes, object deformation and background complexity in real-world scenes.

The traditional object recognition algorithms adopt the procedure of object location, feature extraction and classification. However, it is difficult to extract deep features of images. They rely too much on the manual adjustment of parameters, resulting in a low degree of classification accuracy.

In recent years, many deep learning networks, such as convolutional neural network (CNN) and long short term memory (LSTM), have shown their superior advantages in the field of object recognition. The deep learning methods can extract more complex features, and can automatically adjust the parameters. Besides, it is easier for deep learning methods to find the global or the local optimal solution than traditional methods. Therefore, many researchers introduced CNNs to extract deep features of objects to perform multi-object recognition.

Many researchers improve the performance of deep learning networks by deepening the network and increasing the number of the network parameters. In this way, the deep learning model is more accurate to exhibit information of objects. However, deep learning networks with large depths have many shortcomings. The number of

Received 2019-12-05
Supported by the National Natural Science Foundation of China (61701029); Basic Research Foundation of Beijing Institute of Technology (20170542008); Industry-University-Research Innovation Foundation of the Science and Technology Development Center of the Ministry of Education (2018A02012)
✉ Author for correspondence, assistant professor, Ph.D.
E-mail: funye@bit.edu.cn
DOI: 10.15918/j.jbit1004-0579.19116

training parameters is large, thus, the training process takes up a lot of memory and the trained model may be over-fitting. Some researchers solve the above problem by expanding the image dataset and changing the deep learning network structure. Masi I et al.[1] effectively solve the problem of over-fitting by augmenting the image dataset and improving the generalization ability of CNN. GoogLeNet[2] replaces the fully connected layer with the average pooling layer. In this way, GoogLeNet greatly speeds up the convergence of the training process. However, GoogLeNet is so deep that it has a large network structure and requires high computation resources.

Aiming at solving the problems that the parameters of deep learning models are complex and large computing resources are required, we propose an algorithm based on an improved CNN. First, candidate object windows are extracted by using a selective search algorithm. Then, the network with fewer parameters is designed. We adopt a CNN with moderate depth and width and replace the fully connected layer with the average pooling layer. Afterwards, we extract deep features from the candidate objects using the proposed network. Finally, the Softmax is used to classify the extracted object features. According to the classification result and its confidence score, the multi-object localization and recognition results are determined. Compared with R-CNN[3], which is the classical object recognition algorithm, our algorithm exhibits higher recognition accuracy and comparable localization accuracy. Compared with the Faster R-CNN[4], our algorithm still has certain advantages. Although the recognition accuracy of our algorithm is slightly lower than the Faster R-CNN, our algorithm has higher training efficiency.

1 Object Recognition Algorithm Based on Improved CNN

1.1 Overview

The flow chart of the proposed algorithm is shown in Fig.1. In the training process, first, the candidate object is extracted from the picture A using the selective search algorithm. Then,candidate object windows including the candidate objects are obtained, as shown in the picture B of Fig.1. After that, based on AlexNet[5], Inception V1 is introduced and the fully connected layer is replaced with the average pooling layer to design an improved CNN framework. In this way, the network width is enlarged. Candidate object windows acquired from the train set are subjected to simple preprocessing operation and input into the improved CNN framework. The features and classification results were obtained through the proposed network, and the classification results are compared with the ground truth to calculate the classification errors. According to the errors, the parameters of each layer are updated by backpropagation and gradient descent until obtaining the optimal parameters.

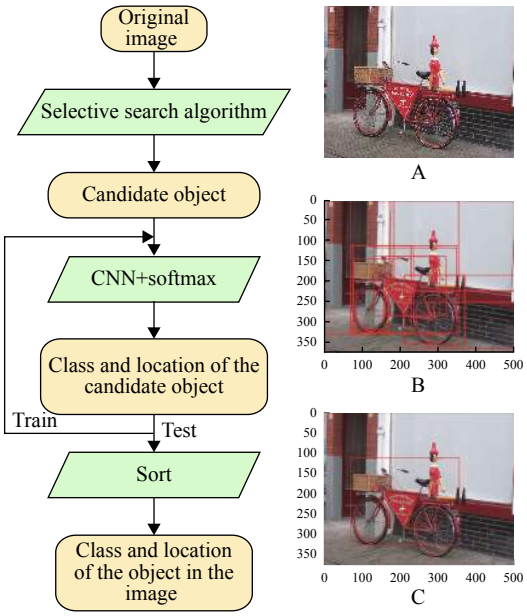

Fig. 1 Workflow of proposed algorithm based on improved CNN

In the testing process, candidate object windows are first extracted from the test set using the selective search algorithm. Then, the candidate object windows are put into the trained network, obtaining the features and recognition results through the forward propagation. Finally,

the candidate object windows are sorted according to the classification and localization accuracy of them. The candidate object window with the highest confidence score is selected as the object recognition result.

1.2 Extraction of the candidate objects

The selective search algorithm[6] is used to extract candidate objects for each image. The selective search algorithm adopts the idea of over-segmentation, and divides the images to acquire multiple initial super-pixels. Then, super-pixels are grouped according to the comprehensive similarity of multiple features of adjacent super-pixels, so that super-pixels with higher similarity are merged into new super-pixels. Finally, all the super-pixel sets that meet the conditions are output.

In order to reduce the computing complexity after grouping, the selective search algorithm utilizes the features which have transferable characteristics. The feature of the new super-pixel can be directly obtained by using only the original super-pixel feature. In this way, the feature of new super-pixel needs not be repeatedly calculated. The selective search algorithm utilizes eight-color features, fast SIFT-Like features and size features. The eight-color features are RGB, the intensity (grey-scale image) I, Lab, the rg channels of normalized RGB plus intensity denoted as rgI, HSV, normalized RGB denoted as rgb,C[7] and the Hue channel H from HSV.

The algorithm defines four different similarity metrics within the range of [0, 1], and finally integrates them to obtain the final similarity. The similarity is calculated as follows:

① Color similarity

For each super-pixel, we obtain a one-dimensional color histogram for each color channel using 25 bins. When three color channels are used, we need to calculate the 75-dimensional color histogram $C_i = \{c_i^1, \cdots, c_i^n\}$ for each super-pixel r_i. Then the color similarity of two super-pixels r_i and r_j is calculated by

$$S_{\text{colour}}(r_i, r_j) = \sum_{k=1}^{n} \min\left(c_i^k, c_j^k\right) \tag{1}$$

② Texture similarity

Gaussian differential is performed on 8 different directions of each channel, and a 10 bins direction histogram is extracted for each channel. In this way, a 240-dimensional texture feature vector $T_i = \{t_i^1, \cdots, t_i^n\}$ can be extracted from r_i. The texture feature similarity is calculated by

$$S_{\text{texture}}(r_i, r_j) = \sum_{k=1}^{n} \min\left(t_i^k, t_j^k\right) \tag{2}$$

③ Size similarity

Size similarity encourages small super-pixels to merge early. This forces the super-pixels that have not been merged, to have similar sizes throughout the algorithm. The size similarity is calculated as

$$S_{\text{size}}(r_i, r_j) = 1 - \frac{\text{size}(r_i) + \text{size}(r_j)}{\text{size(im)}} \tag{3}$$

where $\text{size}(r_i)$ and $\text{size}(r_j)$ represent the size of r_i and r_j, respectively. $\text{size}(\text{im})$ represents the size of the entire image.

④Shape compatibility

Shape compatibility is used to measure the degree of contact between r_i and r_j. The idea is that if r_i are included in r_j, r_i and r_j will be preferentially merged. If r_i and r_j are less in contact, r_i and r_j will not be grouped. The shape compatibility is calculated as

$$S_{\text{fill}}(r_i, r_j) = 1 - \frac{\text{size}(\text{BB}_{ij}) - \text{size}(r_i) - \text{size}(r_j)}{\text{size}(\text{im})} \tag{4}$$

where BB_{ij} is the bounding box that contains exactly r_i and r_j.

Finally, according to Eq.(5), the similarity of the two super-pixels can be obtained by combining the above four similarities, and the value is between 0 and 1.

$$\begin{aligned} S(r_i, r_j) = & a_1 S_{\text{colour}}(r_i, r_j) + a_2 S_{\text{texture}}(r_i, r_j) + \\ & a_3 S_{\text{size}}(r_i, r_j) + a_4 S_{\text{fill}}(r_i, r_j) \end{aligned} \tag{5}$$

For the specific situation of the data used in

our experiments, the initial super-pixel size is set as 200 pixels. The size of the obtained candidate object window is at least 600 pixels. The number of candidate object windows is extracted up to 1 000. So far, by calculating the features of the super-pixels and the similarity of the adjacent super-pixels, candidate object windows of different scales can be acquired. Then the windows can be input into the improved CNN to perform feature extraction.

1.3 Feature extraction

Since the objects in images are of multiple-scales, it is necessary to perform pre-processing on each candidate object window. We resize all images to 256×256.

The proposed network is shown in Fig.2. In the network, the first two layers are cascaded convolutional layers, and then the Inception V1 structure is added. The convolution kernels of 1×1, 3×3, and 5×5 are used to convolve the feature maps to further obtain deep feature maps at different scales. After the integration, the deep feature maps are input into three convolution layers. Finally, the final eigenvector is obtained through two pooling layers.

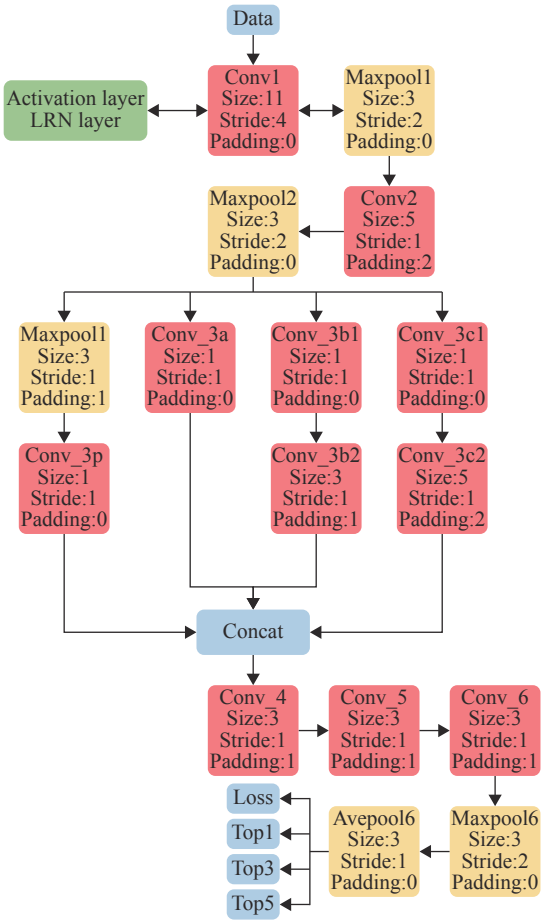

Fig. 2 Improved CNN

The algorithm adds the Inception V1 structure after the second convolutional layer of AlexNet, stacks the convolutional layers of size 1×1, 3×3, and 5×5, and integrates the features of the three scales. The aim is increasing the width of the network structure while deepening the network. Since the convolutional layers of 3×3 and 5×5 require high computing resources, a convolutional layer of 1×1 is added before them to reduce the depth of the feature map, which are input to 3×3 and 5×5 convolutional layers. In addition, the output layer of AlexNet is improved. The fully connected layer is replaced with the average pooling layer. The purpose is to reduce the number of the parameters, thereby improving the training efficiency of CNN and the generalization ability of the network.

1.4 Feature classification

The classifier used in our algorithm is the Softmax, which uses cross-entropy loss for classification. The function is shown as

$$P(i) = \frac{\exp(\underline{\theta}_i^{\mathrm{T}} \underline{x})}{\sum_{k=1}^{K} \exp(\underline{\theta}_k^{\mathrm{T}} \underline{x})} \tag{6}$$

In the testing process, first, calculate the probability $P(i)$ of the input image's feature $\underline{x}$ belonging to the i-th class. Then the class with the highest probability $P(i)$ is selected as the recognition result.

1.5 Object localization

Our object localization algorithm is shown in Fig.3. The white rectangular in Fig. 3 represents the object localization result, and the upper left corner marks the class of the object and the recognition confidence score.

After candidate object windows passing through the CNN structure, each window's class and its confidence score can be obtained. The

neighboring windows are sorted by the confidence score of each class. The candidate object window with the highest confidence score is selected as the localization result. The class represented by the highest confidence score is used as the recognition class of the sample object.

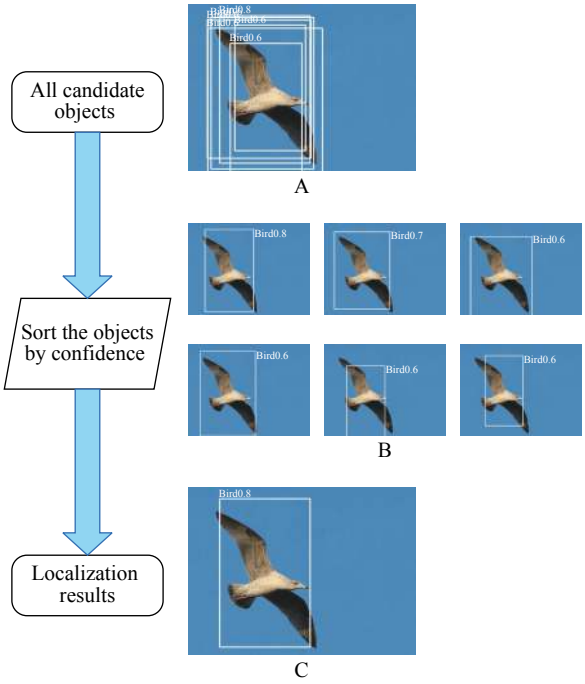

Fig. 3 Object localization algorithm

2 Experiments

The effectiveness and recognition efficiency of our algorithm are verified on the test set of the VOC2007 dataset. The VOC2007 dataset contains 9 963 images in 20 classes. The proposed algorithm differs from R-CNN in network structures. R-CNN selects AlexNet with 8 layers to extract the deep features of the candidate objects. Our algorithm improves object recognition performance on the basis of AlexNet. We add the Inception V1 to widen the network while deepening the network and replacing the fully connected layer with the average pooling layer. In this way, our network greatly reduces the number of parameters and thus reduces the computing cost of the training process.

To verify the effectiveness of the algorithm, Top-1 accuracy, Top-5 accuracy and mAP of our algorithm are calculated. Top-1 and Top-5 accuracy are used to describe the accuracy of object classification, and mAP is used to describe the accuracy of object localization. The network utilized by R-CNN is an 8-layer AlexNet, and the network used by Faster R-CNN is a 19-layer VGG, of which 16 layers are convolutional layers. Therefore, we can compare Top-1 accuracy, Top-5 accuracy of our network with those of AlexNet and VGG for simplicity.

In addition, in order to verify the recognition efficiency of our algorithm, we calculate the time taken to process an image and the average time consumption for the entire VOC2007 test set. In order to verify the performance of our algorithm, it is compared with R-CNN and Faster R-CNN.

To verify the effectiveness of our algorithm based on the improved CNN, we have qualitatively and quantitatively compared the recognition and localization performance with other algorithms.

Experiments are performed on the VOC2007 dataset to qualitatively describe the recognition performance of our algorithm. Fig.4 and Fig.5 show the localization and recognition results of R-CNN and our algorithm. As can be seen from Fig.4, R-CNN has a higher recognition accuracy in recognizing humans because of its higher confidence score. However, confidence scores of objects such as animals and potted plants are low. It can be seen from Fig.5 that our algorithm has strong advantages in recognizing humans and animals. However, when recognizing objects such as bottles and trains, the recognition accuracy of our algorithm is low. Fig.6 shows the results of multi-object recognition based on our method. It can be seen that recognition confidence score of our algorithm is high, and the localization is more accurate.

To quantitatively demonstrate the recognition accuracy, our algorithm is compared with other object recognition algorithms. Tab.1 illus-

Fig. 4 Object recognition results for each class of R-CNN

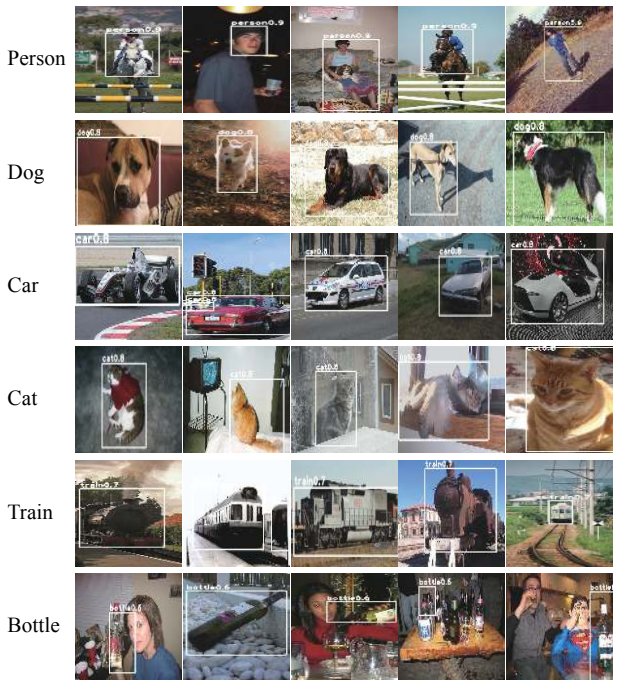

Fig. 5 Object recognition results for each class of our algorithm

trates the recognition accuracy of the proposed network, Alexnet and VGG. As can be seen in Tab.1, Top-1 accuracy and Top-5 accuracy of our algorithm are higher than those of AlexNet, but are slightly inferior to those of VGG. However, our network has a shallower network than VGG, so it requires fewer computing resources and is easier to train. Especially for the case that the dataset is small, our algorithm has higher training efficiency.

In addition, to verify the localization performance of the algorithm, our algorithm is compared with R-CNN and Faster R-CNN under mAP. The comparison results are shown in Tab.2. It can be seen that our algorithm is comparable to other algorithms in object localization, which has certain competitiveness.

Fig. 6 Multi-object recognition based on our method

Tab. 1 Comparison results of recognition accuracy

Algorithm	AlexNet	VGG	Proposed algorithm
Top-1 accuracy/%	62.5	75.2	68.4
Top-5 accuracy/%	83.0	92.5	87.9

Tab. 2 Comparison results of localization accuracy

Algorithm	R-CNN	Faster R-CNN	Proposed algorithm
mAP/%	58.5	59.5	58.7

To verify the efficiency of our algorithm, the processing speeds of R-CNN, Faster R-CNN and our algorithm are calculated respectively under the same hardware condition. The comparison results are shown in Tab.3. It can be seen that the processing speed of our algorithm is faster than R-CNN but slower than Faster R-CNN.

The reason is that Faster R-CNN unifies the four steps of candidate region generation, feature extraction, feature classification and location refinement into one framework, so its processing speed is high. However, compared with R-CNN, our improved CNN has higher object recognition accuracy, thus, the number of objects obtained in the candidate object extraction step is reduced, which reduces the processing time of each image.

Tab. 3 Comparison results of recognition efficiency

Algorithm	R-CNN	Faster R-CNN	Proposed algorithm
Processing time/s	76	0.32	5.34

In summary, our algorithm has higher recognition accuracy, localization accuracy and recognition efficiency than R-CNN. Compared with Faster R-CNN, our improved algorithm is not ideal. However, our algorithm has higher training efficiency and stronger generalization ability. Therefore, compared to Faster R-CNN, the proposed algorithm can be trained in a smaller dataset and lower hardware configuration environment than Faster R-CNN.

3 Conclusion

Aiming at solving the problems of low recognition accuracy of the traditional object recognition algorithm and high complexity of the deep network training, we improve AlexNet and combine the selective search algorithm and Softmax to realize the recognition of multiple classes of objects. First, candidate objects are roughly extracted by the selective search algorithm, and then a plurality of candidate object windows are obtained. Furthermore, the windows are input into the designed CNN. After multi-layer convolution and pooling, deep features are acquired. Finally, Softmax is used to further classify the features extracted from the candidate objects, and the redundant candidate object windows are removed while subdividing the classes of objects. Experiments show that our algorithm has strong advantages in multiple natural scene images. What 's more, It improves training efficiency and recognition efficiency without reducing accuracy.

References:

[1] Masi I, TrÅn A T, Hassner T, et al. Do we really need to collect millions of faces for effective face recognition?[C]//Proceedings of European Conference on Computer Vision. Amsterdam, Holland: Springer, 2016: 579-596.

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems, 2015: 91-99.

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, 2012: 1097-1105.

[6] Uijlings J, Sande K E, Gevers T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[7] Geusebroek J M, Van den Boomgaard R, Smeulders A W M, et al. Color invariance [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(12): 1338-1350.