Research on Pedestrian Detection based on Faster R-CNN and Hippocampal Neural Network

Biao Hao
Dept. of Electronics Engineering
Dong-A University
Busan, Korea
13315982557@163.com

Su-Bin Park
Dept. of Electronics Engineering
Dong-A University
Busan, Korea
jineus0706@naver.com

Dae-Seong Kang
Dept. of Electronics Engineering
Dong-A University
Busan, Korea
dskang@dau.ac.kr

***Abstract*— This paper use Faster-RCNN and hippocampal neural network algorithms to research. Firstly use convolutional neural network to extract the features of the input image, and then use Region Proposal Networks to extract the standard frame. Here we can judge whether there are objects in the standard frame and know the location of the standard box, then use the Non-Maximum Suppression to select the standard box, finally perform the classification operation and regression operation. The final classification network is the hippocampal neural network. The hippocampal neural network is a spatial structure model that mimics the hippocampus of human brain.**

Keywords—faster-RCNN; hippocampal neural network; convolutional neural network; region proposal networks; classification

I. INTRODUCTION

Artificial intelligence is the mainstream of social development now. In this field, deep learning is used by more and more people. Deep learning is a branch of machine learning. We can understand deep learning as an improvement of algorithms in machine learning. Artificial intelligence is broadly that machines imitate the human brain to build models and then perform related operations. Convolutional neural networks in deep learning have a good effect on image processing. The convolutional neural networks we often use are Lenet-5[1], Alexnet[2], VGG[3], GoogLeNet[4], etc. The original convolutional neural network is Lenet-5, but at that time because of the condition, convolutional neural network development was limited. In recent years, with the development of hardware and software, the convolution neural network has developed very much.

Before the most commonly used image processing algorithm is a machine learning algorithm. The most commonly used machine learning features are HOG (Histogram of Oriented Gradient)[5][6] and LBP (Local Binary Pattern)[7]. Machine learning is that use feature extraction algorithm to extract features and then use the classifier to classify, finally get the purpose of pedestrian detection. However, in the machine learning, the feature extraction algorithm needs human's operation to perform feature extraction, which has great error. In deep learning, the convolutional neural network can autonomously extract features. So now most of the image processing uses convolutional neural networks and achieved good results. The algorithm used in this paper is based on faster-RCNN and hippocampal neural network and then implements pedestrian detection.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors).*

II. RELATED ALGORITHM

A. Convolutional Neural Networks

Convolutional neural network is an important part of deep learning. The original neural network model is a fully connected network model. The network has a large amount of computation and the speed of operation is slow, but with the emergence of convolutional neural networks in deep learning. Its weight sharing structure makes the computational workload between neural networks greatly reduced, and the speed of operation is greatly improved. Neural network is modeled on the human brain to build a model to achieve the function of human brain. The study of the network actually changes the weights. The neural network is composed of neurons. Its equation is as (1). W is the weight, b is the bias and f is the activation function. This can be calculated.

$$y_i = f(\sum_{i=1}^{n} w_{ji} x_i + \theta_i) \tag{1}$$

The process of convolution calculation is as Fig. 1. Input data and kernel calculate, the result is 29.If stride is 2, the next result is 56.

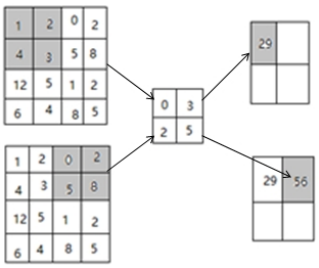

Fig. 1. Convolutional Calculation

There are several basic parts of the training of neural networks. They are activation functions and pooling. Most of the activation functions are sigmoid, ReLU, Tanh, etc. The

most of the neural networks use ReLU function to train. The equation is as (2). The image of function is as Fig. 2.

$$y = \begin{cases} 0, x \le 0 \\ x, x \rangle 0 \end{cases} \tag{2}$$

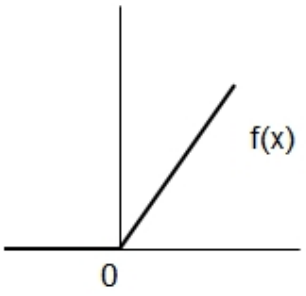

Fig. 2. ReLU Function Image

Most of the activation functions are sigmoid, ReLU, Tanh, etc. Most of the neural networks use the ReLU function.

Convolutional neural network pooling operation mainly has two kinds, one is max pooling, the other is average pooling. The process of pooling operation is as Fig. 3. It is easy to understand this part of the operation

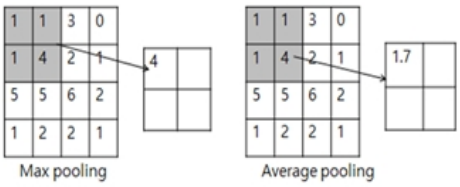

Fig. 3. Pooling Calculation

B. Region-based Convolutional Neural Networks

R-CNN can achieve target detection. R-CNN[8][9] algorithm input the standard frames obtained by selective search[10] and the image into the neural network, first perform convolutional neural network operation and then perform classification and regression. Finally it can achieve the object detection.

C. Fast R-CNN

The change of fast-RCNN[11] is the position that extracted candidate frame by selective search. R-CNN input candidate frames into the neural network along with the images. But in this algorithm first the feature is extracted by CNN and then the input candidate frames extracted by selective search, and then perform a series of other operations, finally perform classification and regression to achieve the task of object detection. The loss function is a multi-loss function. This function is divided into two parts, one is the classification loss function, the other is the regression loss function. It is multi-loss function as (3), i is the number of anchors, p_i is the probability of foreground predict. p_i^* is the probability of ground truth predict. t_i is predicted bounding box, t_i^* stands for foreground anchor corresponding ground truth box.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}\left(t_i, t_i^*\right) \tag{3}$$

III. Proposed Algorithm

The proposed algorithm first uses the VGG convolutional neural network to extract features from the image. However, in this algorithm the region proposal network(RPN) is used to extract the standard frame and then through the Non-Maximum Suppression(NMS) to further reduce the number of candidate frames. Then use the hippocampal neural network to classify these candidate boxes, and finally use the regression algorithm to perform regression operations on the candidate boxes. Thus achieve the task of pedestrian detection. The total algorithm flowsheet is as Fig. 4. The proposed algorithm from Fig. 4 contains VGG convolutional neural network, Region proposal neural network, Non-Maximum Suppression, Fully connection network, classification and regression.

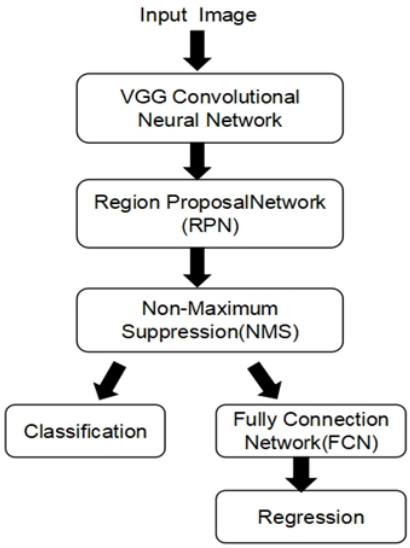

Fig. 4. Total Flowsheet

A. Visual Geometry Group Network

The full name of VGG convolutional neural network is the Visual Geometry Group Convolutional neural network. It has a total of 16 layers but here only apply its convolutional part. It can be seen easily from the Fig. 5, there are 13 convolutional layers. Through the part of the convolutional

neural network it can extract feature. The construe of VGG network is that two convolution layers, two convolution layers, three convolution layers, three convolution layers, three convolution layers and three fully connection layers.

The specific structure of the neural network in turn are 2 convolution layers, 1 pooling layer, then the next 2 convolution layers, 1 pooling layer, 3 convolution layers, 1 pooling layer, 3 convolution layers, 1 layer pooling layer, 3 layer convolution layer, 1 layer pooling layer, and finally 3 full connection layers.

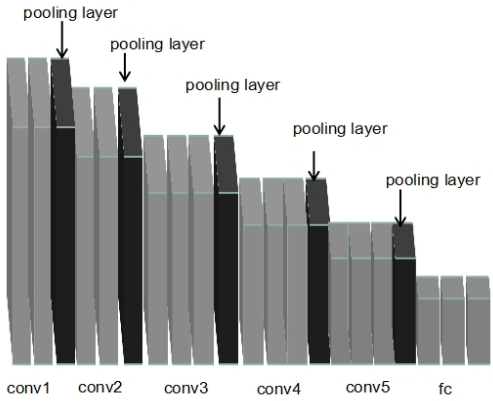

Fig. 5. The Structure of VGG

B. Region Proposal Networks

Region proposal networks extract the standard boxes on the feature map. This network uses 9 different sizes of anchors to extract candidate boxes. In the network, the candidate frame is extracted sliding a 3*3 sliding window on the feature map. The scale of these 9 different anchors is [125, 256, 512], and the ratio of length to width is [1:1, 1:2, 2:1]. It produces 9 different anchors, and then classifies these extracted frames and performs the regression. It is classified as a target frame and a non-target frame, and performs regression operation to get the position of frame. The process of Region proposal network is as Fig. 6.

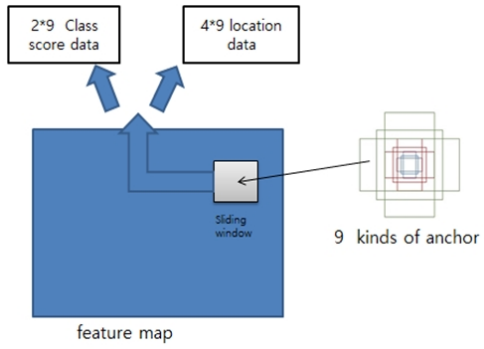

Fig. 6. Process of Region Proposal Network

C. Non-Maximum Suppression

Non-maximal suppression(NMS) is to suppress elements that are not maximal and search for local maxima. In pedestrian detection, the sliding window is used to extract candidate frames. After being classified by the classifier, each window will receive a score. However, the sliding window will cause many frames to overlap with other frames. At this time, it is necessary to use the NMS to select those with the highest score in the neighborhood and suppress those windows with low scores. The frame with highest score its probability of being pedestrians is the highest. It is effect that obtained through NMS as Fig. 7.

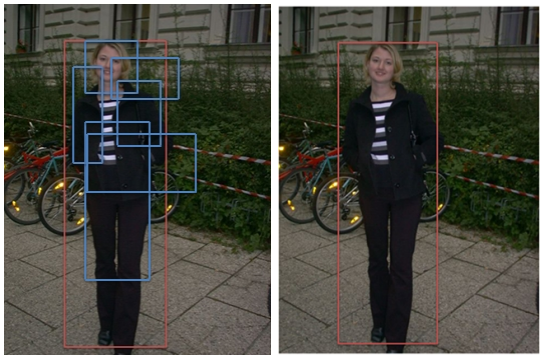

Fig. 7. The Effect of NMS

D. Hippocampus Neural Network

The hippocampal neural network[12][13] trains the extracted feature maps and then classifies the images. The hippocampal neural network mimics the memory function of the hippocampus in the human brain. Its greatest feature is the ability to convert short-term memory into long-term memory. The hippocampus can memorize the stored information and classify the information into required information and unnecessary information.

For those information data that are divided into short-term memory, the number of times of memory exceeds the critical value, and this data will be transferred to the place of long-term memory. As Fig.8, the hippocampal neural network is mainly divided into Entorhinal Cortex, Dentste Gyrus, CA3 and CA1.

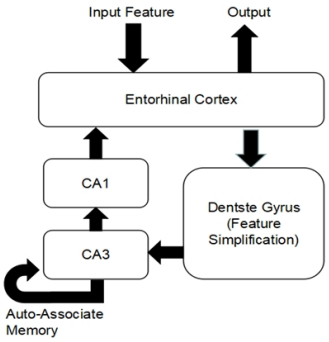

Fig. 8. The Structure of Hippocampus Neural Network

Entorhinal Cortex: This is the neural network input and output interface

Dentste Gyrus: Directly connected to Entorhinal Cortex, Simplification of input features. The average value of the input module is set as a critical value, and the value of the input module that does not exceed the critical value within the allowable range of the deviation is represented by 1 and the threshold value is not exceeded by -1.

CA3: This section is a circular associative memory module that imports loop concepts. Recycled associative memory is a homogeneous association memory that outputs feedback to the input. The CA3 part performs auto-association operation and can better distinguish the noise-containing data. This section is similar to hopfield.

CA1: This part is the final stage of information processing. Decide on long-term memory and short-term memory. This part is similar to perceptron.

E. The Choice of Tensorflow

The software library based on deep learning now has caffe, Caffe, Theano, and MXNet, etc. We are using tensorflow. Tensorflow is an artificial intelligence platform developed by Google. He is a computing framework based on computational graphs. From the name we can see that tensorflow can be divided into two parts, one is tensor and the other is flow. tensor is an N-dimensional array, flow is the meaning of graph.

TensorFlo is an open source software library for numerical computation using data flow graphs.

How Tensorflow works, so you first create a calculation graph in memory using the TensorFlow function, then start an execution session and use session.run to perform the actual training task. It is a calculation as Fig. 8. If we want to do calculations, we first need to define a graph as Fig. 8, then substitute the value and then get the result.

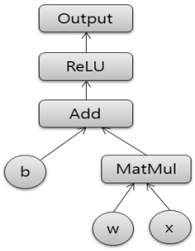

Fig. 8. The Process of Calculation

IV. Experiment and Result

This study uses the INRIA database which is currently the most used static pedestrian detection database.There are 614 positive samples and 1218 negative samples in the training set. There are 288 positive samples and 453 negative samples in the test set. The positive image and negative image is as Fig. 9. The positive image is that has object in image, negative image is that don't have object.

(a)Positive Image (b) Negative Image

Fig. 9. Positive Image and Negative Image

The result of experiment is as Fig. 10. Pedestrian detection's aim is that realize the classification for pedestrian and location of objects in a picture.

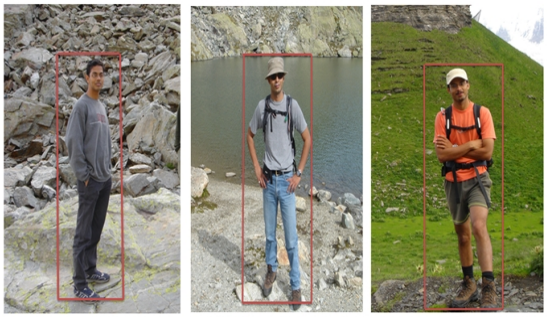

Fig. 10. Experimental Results

V. Conclutions

In the field of artificial intelligence image processing can achieve good results through convolutional neural networks. In the problem of pedestrian detection, faster-RCNN has a good effect, and there is a good performance of using hippocampal neural network in the task of image classification. Therefore, in this paper use faster R-CNN and hippocampal neural network to do pedestrian detection. The used convolutional neural network in this algorithm is a well-performing VGG convolutional neural network, Use it to extract features from input image, and then performs a series of processing on the extracted features to perform classification and regression operations. This network can well implement the task of pedestrian detection. The hippocampal neural network added in this algorithm can increase the recognition rate in the classification task, but reduce the running speed. But the overall effect is still very good. In the future research, I will continued study deep learning related algorithms and related object detection algorithms.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(NO. 2017R1D1A1B04030870).

REFERENCES

[1] Y.L Cun, L. Botton, Y. Bengio and P. Haffner, “Gradient-Based Learning Applied to Document Recognition,” PROC. OF THE IEEE, November.1998.

[2] A. Krizhevsky, Il. Sutskever and G. E. Hinton, “Image Classification with Deep Convolutional Neural Network,” Advances in Neural Information Processing Systems (NIPS 2012). 2012.

[3] K. Simonyan and A. Zisserman, ”Very Deep Convolutional Networks for Large-Scale Image Recognition ,” Computer Science, Computer Vision and Pattern Recognition, Cornell University Library. Sep 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Semannet, S.Reed, D. Anguelov, etal.”Going Deeper with Convolutions,” CVPR, Computer Vision Foundation, 2015

[5] C.Xiong and W. W. Wang, “Research on pedestrian detection based on DPM,” Electronic Design Engineering. Vol.22, 2014.

[6] X. Y. Wang, T. X. Han, and S. C. Yan, “An hog-lbp human detector with partial occlusion handling,” In ICCV, pp. 32–39, 2009

[7] M. A. Rahim, M. N. Hossain, T. Wahid and M.S. Azam, “Face Recognition using Local Binary Patterns (LBP),” Global Journal of Computer Science and Technology Graphics & Vision, USA.2013.

[8] H. K.Lee and H. J. Kim, “Detection of human using R-CNN for vision data,” KSAS, pp.751-752, Spring, 2016.

[9] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” CVPR, 2014.

[10] J. RR. Uijlings, K. EA. Van de Sande, T. Gevers and A. WM. Smeulders, “Selective search for object recognition,” IJCV, 2013.

[11] R. Girshick and M. Research, “Fast R-CNN,” Computer Vidion Foundation.

[12] S. M. Oh and D. S. Kang, “Development of Learning Algorithm using Brain Modeling of Hippocampus for Face Recognition,” Journal of the Institute of Electronics Engineers, Korea, Sep, 2005.

[13] S. M. Oh and D. S. Kang, “Development of the Hippocampal Learning Algorithm Using Associate Memory and Modulator of Neural Weight”, Journal of the Institute of Electronics Engineers, Korea, July, 2005.