Object Detection Using Regions with Convolutional Neural Networks (R-CNN)

Hnin Cherry and Myint Myint Sein
University of Computer Studies, Yangon
hnincherry@ucsy.edu.mm, myintucsy@gmail.com

Abstract

With an importance of artificial intelligence in today's world, deep learning technology has developed very powerful in solving many problems in various fields that is included in speech recognition, natural language processing, computer vision technologies, image processing and video, and different kinds of multimedia. Due to the development of deep learning approach, visual recognition systems have achieved in good performance. With the increase of smart application in visual recognition, powerful object detection systems are necessarily needed. In detecting objects, object classification is as a very important role. Deep Neural Network (DNN) can greatly achieved in classifying objects. In the experiment, object detection system for stop sign is implemented by using Regions with Convolutional Neural Networks (R-CNN) that is used to classify image regions included in an image. The system intended to provide object detection accurately.

Keywords: *Object Detection, Deep Neural Network (DNN), Convolutional Neural Networks (CNN), Regions with Convolutional Neural Networks (R-CNN)*

1. Introduction

Nowadays, due to the efficient results obtained in object detection, image classification and natural language processing, deep learning technology has improved effectively. To do the process of image understanding, different images are needed to classify, show the concepts and locations of objects contained in each image. This task is known as object detection [8]. As it is the problem in computer vision, object detection is developed to provide understanding of images and videos with valuable information for semantic, and is related to various applications, including classification of images [12], human behavior analysis [7], facial recognition [13] and autonomous driving system [2].

With the development of object representation and machine learning approaches, object detection has achieved in advance. The establishment of region proposal methods (RPM) (e.g., [5]) and R-CNN [10] are developed for object detection in advanced technology. Fast R-CNN [9] developed real-time rates by using neural networks [6] with deep learning technology. In this paper, the system provides object detection by using R-CNN with the performance with high accuracy.

This paper organized as follow. The related works on object detection are discussed in section 2, object detection in section 3, object detection using CNN and R-CNN in section 4, experiment in section 5, performance analysis of the system is described in section 6, and conclusion and future work in section 7.

2. Related Works

Badri Narayana Patro and Ganesh Oddupally [1] proposed the system that is used to detect and a certain object in a video and then find out corresponding timing details of that object in the video sequence, i.e, what frames are available in the object or the different time interval is available in the video scene. They proposed an algorithm that will provide object recognition and localization in a video. Object detection and classification are the main part of the object recognition. Object detection system is developed with the use of Gaussian Mixture Model (GMM) and object localization using object proposal calculation method. They have done training data set by using Bag of word model.

Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun [11] described Region Proposal

Network (RPN) which makes for full-image sharing convolutional features with the help of detecting network. RPNs that are simultaneously fully-convolutional networks to predict bounds of object and the scores of objectness for each position and are trained to provide the region proposals with high quality by using Fast R-CNN for detection.

Christian Szegedy, Alexander Toshev and Dumitru Erhan [3] go the use of DNN for the case of detecting objects, that makes not only classifying but also localizing objects in different classes. They describe a powerful and simple formulation of detecting object as a regression problem to bounding box masks for objects. And then they define a multi-scale inference procedure that is used to produce the detections for high-resolution object at a low cost with a few network applications.

3. Object Detection

Today, both images and videos are used in every media such as social networks. The area of research in computer vision has been established by using machine learning approaches and statistics. To understand a real-world scene, it is need to detect by using images and videos, classifying and tracking objects. Object recognition is a process of defining a certain object in an image or a video that is related to computer vision system and image processing applications. It used detecting objects with semantic of a specific class (such as peoples, vehicles, or dogs) in digitalized images and videos.

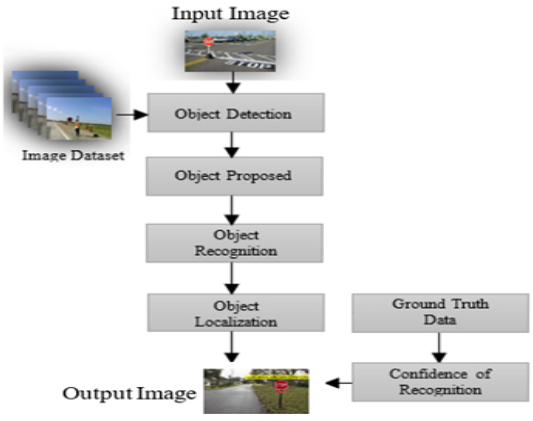


Figure 1. Object detection process

Object detection [14] is one of the domain researches included in face detection and pedestrian detection. It is included as main application in different areas of computer vision like as video surveillance and retrieving image. The major steps for object recognitions and localization as shown in Figure 1.

4. Object Detection Using CNN and R-CNN

The neural networks with deep structures are defined as deep learning models. The following factors are consisted in the deep learning.

• Large-scale for emergence annotated training data, such as ImageNet [4], to fully exhibit its very large learning capacity;

• Parallel computing systems with high performance for the fast development, such as GPU clusters;

• Design of network structures and training strategies are significantly improved in advance.

In the system, R-CNN object detector is used to detect the stop sign in the scene. R-CNN is a framework for object detection, which uses CNN to classify image regions. Instead of classifying every region, R-CNN processes only the region that contains the object in the image.

4.1. CNN

CNN [16] is a class of DNN. CNN are composed of neurons that have learnable weights and biases. CNN contains an input layer, an output layer, and multiple hidden layers. The hidden layers of a CNN have convolutional layers, pooling layers, fully connected layers and normalization layers respectively.

4.1.1. CNN Architecture

The architecture of CNN is described as described in Figure 2.

Input Layer: 'input layer' is the first layer of CNN and it takes images and resizes them to pass ongoing layers that is used to process for feature extraction.

Convolution Layer: Convolution layer is the core building block of CNN. It performs filtering images and finds features from images and also applies to

calculate matching the feature points in testing process.

Rectified Linear Unit (ReLU): ReLU is a non-linear operation. ReLU substitutes every negative pixel value by 0 in the feature map. ReLU is performed after every convolution operation.

Pooling Layer: Pooling layer extracts the sets of features. It takes large images and reduces the spatial size of the image. Pooling is done independently on each depth dimension. So, the depth of images remains unchanged.

Fully Connected Layer: The fully connected layers are the final layer that takes high-resolution features of the images and then utilizes these features for categorizing input images into different classes.

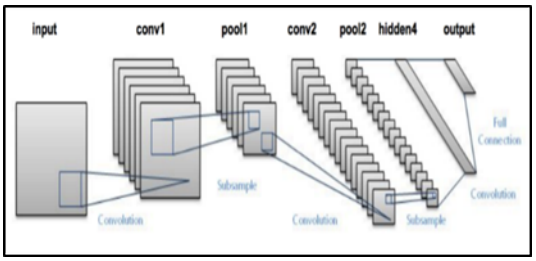


Figure 2. Basic architecture of CNN

The advantages of CNN [14] against traditional methods can be summarised as follows.

• Hierarchical feature representation, which is the multilevel representations from pixel to high-level semantic features learned by a hierarchical multi-stage structure can be learned automatically from data and hidden factors of input data that can be free through multi-level nonlinear mappings.

• When a deeper architecture compared with traditional models, it provides an increased expressive capability.

• The architecture of CNN provides an opportunity to optimize several related tasks together.

• As the large learning capacity of deep CNNs have many benefits, challenges of computer vision system can be reshaped as high-dimensional data that change problems and solved from a different viewpoint.

Due to these advantages, CNN has been applied into many research areas, such as super-resolution reconstruction for image, image classification, image retrieval, face recognition, pedestrian detection and video analysis.

4.2. R-CNN

R-CNN [15] is one of the CNN-based deep learning object detection approaches. R-CNN combines rectangular region proposals with CNN features. R-CNN is a two-stage detection algorithm. The first stage identifies a subset of regions in an image that might contain an object. The second stage classifies the object in each region. R-CNN [17] are based on the following three processes:

- Find regions in the image that might contain an object. These regions are called region proposals.
- Extract CNN features from the region proposals.
- Classify the objects using the extracted features.

5. Experiment

In the experiment, CIFAR-10 dataset is used to train a CNN. This dataset contains 50, 000 of training images in 10 categories. Each image has dimension with 32 x 32 RGB. Some of the training images are as shown in Figure 3.

Then CNN network for the CIFAR-10 dataset is prepared to do training. As a first step, input layer in CNN is created. Next, the middle layers of this network are defined. Middle layers contain convolutional layer, ReLU, and pooling layers repeatedly.

Figure 3. A few training data

Finally, the final layers that compose of fully connected layers and a softmax loss layer are created. After 3 layers of CNN network have created, these 3 layers are combined to form a layer. And

then, first convolutional layer (layer 2)'s weight is initialized with distributed random number 0.0001 to improve the convergence of training process.

After the CNN network structure has defined, CIFAR-10 dataset is used to train the network. Stochastic Gradient Descent with Momentum (SGDM) with initial learning rate of 0.001 is used in the training algorithm.

After the network is trained, it is need to validate to ensure whether the network is successful or not. First convolutional layer's filter weights can indicate that the network may require or not additional training. In this experiment, the first layer filters have learned edge-like features from the CIFAR-10 training data. The first convolutional layer's filter weights as shown in Figure 4.

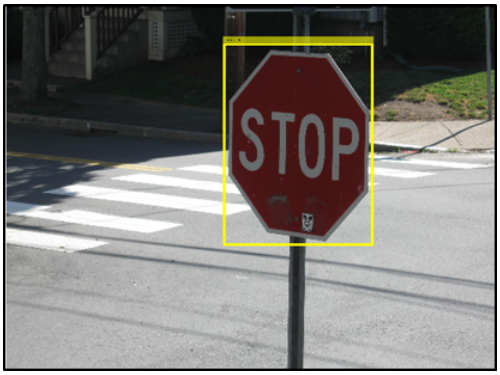


Figure 4. First convolutional layer's filter weights

Next, the classification accuracy of the network is measured to validate the training results by using CIFAR-10 test data. The accuracy of the system is 0.7456 on the test data and it is sufficiently trained network use in training an object detector.

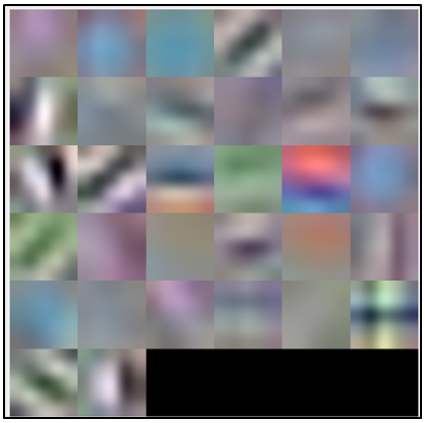


Figure 5. Interest point of an image

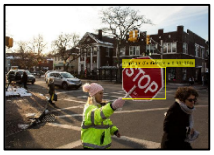
When the network is worked well for the CIFAR-10 dataset, stop sign data is trained on that network for the stop sign detection. In this process, firstly ground truth data for stop sign is loaded. Ground truth table is constructed with the image filename and stop signs attributes as the Region of Interest (ROI) labels. For each of ROI labels, bounding box is represented for the interest of the object as shown in Figure 5.

In final step, R-CNN object detector is trained using the R-CNN training function. The training function automatically modifies the original CIFAR-10 network that classified images into 10 categories, into a network which can classify images into two classes: stop signs and a generic background class. After having trained the R-CNN network, the network is tested to detect stop sign in image.

In the experiment, detection accuracy is described with the confidence whether the successfully detected or not. Detection result in the experiment shows that the stop signs are detected with confidence scores as shown in Figure 6. In the system, CNN is pretrained firstly using the CIFAR-10 dataset. This pretrained CNN is fine-tuned for stop sign detection. And then, R-CNN object detector trained on that pretrained network to detect stop sign. This detection system provided that how to detect stop sign object with the high accuracy by training R-CNN with CIFAR-10 dataset.

Confidence=1.000000

Confidence=1.000000



Confidence=0.991380

Confidence=1.000000

Figure 6. Example of tested images

6. Performance Analysis

To evaluate the classification result, the precision and recall were calculated as

$$Precision = \frac{TP}{TP+FP} \qquad \text{Eq (1)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{Eq (2)}$$

where TP, FP and FN are true positive, false positive and false negative respectively. A true positive is a correctly-detected object. A false positive occurs if the object is detected where there is none. A false negative is a case where the object is missed. A true negative describes the cases where non-object regions are correctly identified as non-object regions. Precision is defined as the number of true positives over the number of true positives plus the number of false positives and recall is defined as the number of true positives over the number of true positives plus the number of false negatives. Figure 7 shows the precision and recall curve for stop sign detection for the input test data.

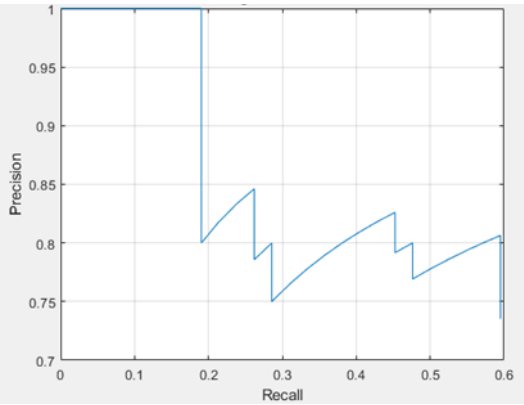


Figure 7. The precision-recall curves

7. Conclusion and Future Work

As the development of technology in deep learning, it is necessary to applied in areas of many researches like as computer vision system, automatic recognition system in speech and various fields with images and videos. R-CNN that makes CNN that is used to classify image regions included in an image. In this paper, the system is developed to detect the object by using R-CNN. It uses CIFAR-10 dataset to train the network.

When the object is detected, the object must be taken by more camera angles. The system is difficult to overcome the problems the overlapping area of image to detect a certain object.

As a future work at high accuracy by using deep learning network to detect objects in real time video surveillance system at low cost. And then the system will be developed to detect the object in the overlapping area with an image.

References

[1]. Badri Narayana Patro, Ganesh Oddupally, "Object recognition and localization", https://pdfs.semanticscholar.org.

[2]. C. Chen, A. Seff, A. L. Kornhauser, and J. Xiao, " Deepdriving: Learning affordance for direct perception in autonomous driving," in ICCV, 2015.

[3]. Christian Szegedy, Alexander Toshev and Dumitru Erhan, "Deep Neural Networks for Object Detection", Google, Inc.

[4]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, " Imageenet: A large-scale hierarchical image database, " in CVPR, 2009.

[5]. J. R. Uijlings, K. E. van de Sande, T. Gevers, and A.W. Smeulders. Selective search for object recognition. IJCV, 2013.

[6]. K.Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[7]. M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of th art analysis." In CVPR, 2014.

[8]. P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, p.1627, 2010.

[9]. R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015.

[10]. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurat object detection and semantic segmentation. In CVPR, 2014.

[11]. Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Microsoft Research.

[12]. Y. Jia, E. Shelhamer, J. Daonahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in ACM MM, 2014.

[13]. Z. Yang and R. Nevatia, "A multi-scale cascade fully convolutional network face detector," in ICPR,2016.

[14]. Zhong-Qiu Zhao, Member, Shou-tao Xu, and Xindong Wu, "Object Detection with Deep Learning: A Review", in JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, MARCH 2017.

[15]. [2014 CVPR] [R-CNN] Rich feature hierarchies for accurate object detection and semantic segmentation

[16]. http://cs231n.github.io/convolutional-networks/

[17]. http://www.mathworks.com/help/vision/ug/faster-r-cnn-basics.html.html.