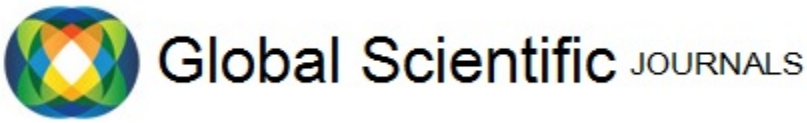
GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186



GSJ: Volume 7, Issue 6, June 2019, Online: ISSN 2320-9186
www.globalscientificjournal.com

A STUDY ON USING DEEP LEARNING TECHNOLOGIES IN CONVOLUTIONAL NEURAL NETWORKS FOR MULTIPLE OBJECTS IDENTIFICATION

VAQ-Nguyen
Vietnam Korea friendship IT College

KeyWords
Deep learning, Convolutional neural networks, CNNs, RCNN, Fast RCNN, Object dentifying, Object positioning, Multiple object positioning, Image processing

ABSTRACT

In this paper, the author will use deep learning technology, specifically here to use techniques in convolutional neural networks (CNN) to identify objects, multiple objects and objects' locations in images. CNN, Region CNN (RCNN), Fast RCNN, Faster RCNN techniques using in convolutional neural networks will be introduced and simulated, citing comparative results to assess the effectiveness of object identification in images. Faster RCNN is considered to be the most appropriate for identifying the objects and their location in the image.

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

I. INTRODUCTION

Artificial Intelligence (AI) is used in many fields in our life. It's used to automatically reply on email, learn how to drive a car, rearrange photos into individual albums, and even help manage house well and so on. AI can simply be interpreted as composed of stacked layers, in which the artificial neural network is located at the bottom, machine learning is located on the next layer and deep learning is located on the top.

In 2011, Google created the Google Brain project with the aim of creating a neural network trained by Deep Learning algorithms. This project later proved Deep Learning's ability to acquire both high-level concepts. Facebook also created AI Research Unit, AI research unit using Deep Learning to create more effective solutions that help identify faces and things on 350 million photos and videos posted to Facebook everyday. Another good example of Deep Learning is in fact the ability to recognize the voice of virtual assistants Google Now and Siri.

Deep Learning is increasingly showing a promising future with applications to drive self-driving cars or robbers. Although these products are still have some limitations, but the things they do today are really hard to imagine just a few years ago. The upgrade rate is also unprecedented high. The ability to analyze big data and using deep learning into computer systems that can self-adapt to what they receive without human's programming. These breakthroughs can be the design of virtual assistants, self-driving car systems, graphic design, music composition, new materials development that help robots understand the world around us. In particular, Deep Learning or general artificial intelligence has a lot of great apps, but we now live in a early stage so the limitations are inevitable.

CNN is one of the Deep Learning algorithms that gives the best results in most machine vision problems such as classification and identification. Since LeNet's success with Yann LeCun and colleagues [3] published in 1998 and ImageNet published by Alex Krizhevsky and his colleagues in 2012 [4], convolutional neural networks (CNNs) have become standard for image classification. From that time, the deviation of CNN has improved to approximately the same level as humans.

If you only need to identify a single object in image, you simply need to use a simple CNN network. But the problem is that when the image has many objects in the image, the problem becomes very complicated. We have to find the objects' position in the image and then proceed to classify. The position of objects may overlap in different contexts, in addition to finding images that require us to define boundaries, differences and relationships with each other. In this section, we will learn through the main techniques used in the detection of classification and location of objects. Specifically, it will introduce the initial techniques of CNN network and later development techniques such as RCNN, Fast RCNN, Faster RCNN. Here are the techniques used to identify locations and objects in the CNN network like using a simple CNN network, R-CNN technology and, Fast R-CNN technology.

II. TECHNICAL USES TO DETERMINE LOCATION AND SUBJECTS IN CNN NETWORK

A. Convolutional neural network

CNN has architecture formed from basic components including Convolution (CONV), Pooling (POOL), ReLU, and Fully-connected (FC). A general CNN architecture is described as Figure 1. CNN is an algorithm with architecture consisting of many layers with different functions in which the main layer operates with a convolution mechanism. During the training process, CNN will automatically learn the parameters for the filters - corresponding to each level. For example in the image classification problem, CNN will try to find the optimal parameters for the corresponding filters in pixels> edges> shapes> facial> high-level features. This is a reason why CNN has superior results compared to previous algorithms.

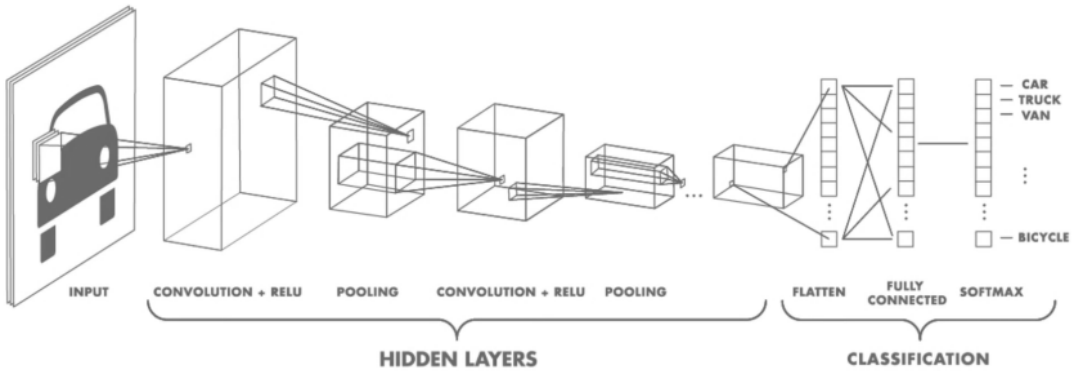


Figure 1. Basic CNN network architecture.

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

B. Determine a single object using a simple CNN network.

Here, the regression algorithm is used to increase the accuracy of bounding boxes to determine the position of the object. The indexes (x0, y0, height, width) of the bounding box are regressed. We train the network with the object image that has been defined with the original index of the bounding box and compare the regression index with the original index for calculating the error of the bounding box. Usually, the locating function is added to the fully connected layer of the convolution network.

C. R-CNN technique

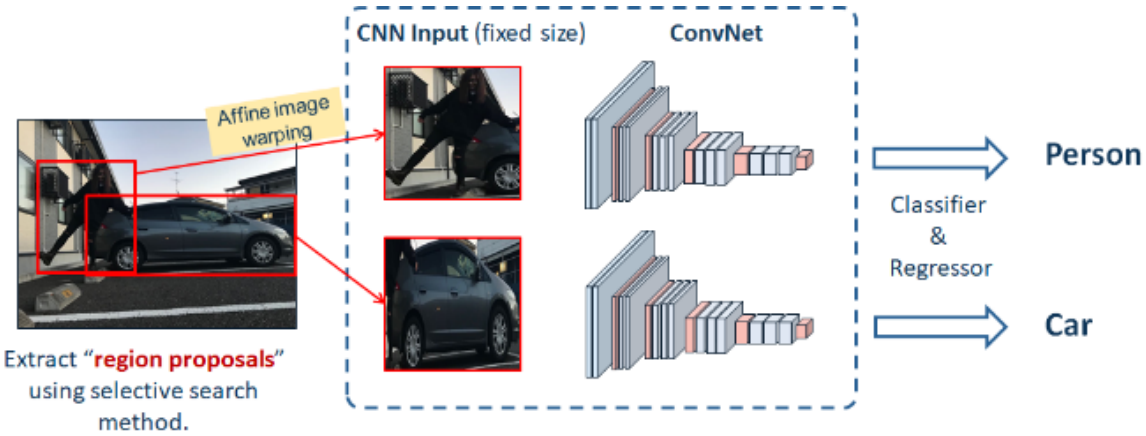


Figure 2. The CNN network using R-CNN technology.

RCNN (Regions + CNN) is a method using selective searching algorithms to predict possible locations of objects in the image, called proposed regions. These proposed regions will be resized by image processing algorithms and included in trained CNN networks. The number of proposed regions can be up to 2000 region per image. After being fed into the input of the CNN network that has been pre-trained for feed forward calculations, it will obtain the convolution characteristics of each proposed region, then continue to train SVM to determine which objects are contained in the proposal region. These proposed regions will be saved to memory. Finally, linear regression algorithms will be used to calibrate the values of the vertices of the proposed regions. Since a large number of proposed regions are to be added to the CNN network and processed in a sequential manner, the speed of RCNN is very slow.

D. Fast R-CNN technique

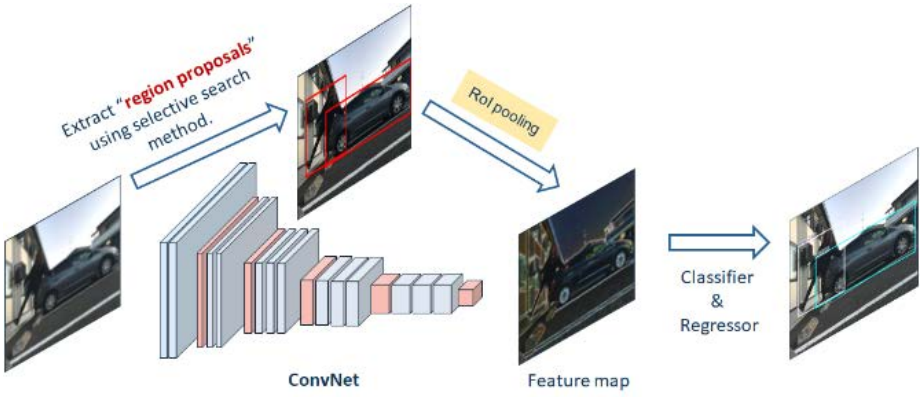


Figure 3. CNN network using Fast R-CNN technology.

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

Adopting available training networks to feed forward for predicted regions, it will take a long time because with each image the selected search algorithm will produce thousands of predicted regions. We will only feed-forward once for the original image, and receiving the convolution characteristics of the image. Based on the size and position of the predicted regions, we will calculate the position of the predicted region in the convolution region. Then we predict the position of the edges of contour as well as what object is in the contour. As we can see, due to sharing calculations between regions in the image, the execution time of the algorithm has been reduced from 120s to 2s per image. The calculation that causes congestion is the predicted input regions, which can only be executed sequentially on the CPU. Faster RCNN solved this problem by applying the Region Proposal Network to calculate the predicted regions of this subject.

E. Faster R-CNN technique

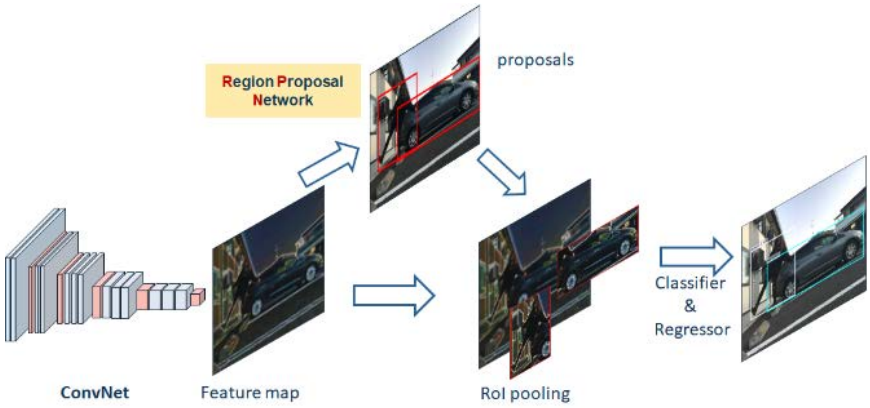


Figure 4. CNN network using Faster R-CNN technology.

The main difference between faster RCNN and fast RCNN is that the RCNN employs selective searching algorithms to initiate the proposed regions while faster RCNN uses RPN to initiate the proposed regions. A time to initialize the proposed regions is greatly shortened when using RPN method versus using selective searching algorithms. RPN adopts an algorithm to sort the cell called anchors and check whether these anchor are able to contain objects in it. Depending on the object you need to define in the network, you can customize the size as well as the amount of anchor to increase the effectiveness of this technique. RPN usually has two main steps: feed-forward images over the network and obtain convolution characteristic and use the sliding windows on the images that have obtained convolution characteristics.

III. NETWORK MODEL AND DATA SET

VGGNet network model was developed by Karen Simonyan and Andrew Zisserman [1]. VGGNet has shown that network performance will depend on the depth of the network. The best network model consists of 16 CONV/FC layers and especially it is uniform in terms of architecture with 3x3 convolutions and 2x2 pooling. One weakness of the network model is that it employs a lot of memory and parameters, with over 138 million weights and needs 24Mb for each image to be processed. Detailed parameters are shown in Table 1.

Table 1. Network parameters CNN uses to identify multiple objects.

Class	Size	Memory	Weight
Input	224x224x3	224 * 224 * 3	0
Conv	224x224x64	224 * 224 * 64	1728
Conv	224x224x64	224 * 224 * 64	36864
Pool	112x112x64	112 * 112 * 64	0
Conv	112x112x128	112 * 112 * 128	73728
Conv	112x112x128	112 * 112 * 128	147456

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

Pool	56x56x128	56 * 56 * 128	0
Conv	56x56x256	56 * 56 * 256	294912
Conv	56x56x256	56 * 56 * 256	589824
Conv	56x56x256	56 * 56 * 256	589824
Pool	28x28x256	28 * 28 * 256	0
Conv	28x28x512	28 * 28 * 512	1179648
Conv	28x28x512	28 * 28 * 512	2359296
Conv	28x28x512	28 * 28 * 512	2359296
Pool	14x14x512	14 * 14 * 512	0
Conv	14x14x512	14 * 14 * 512	2359296
Conv	14x14x512	14 * 14 * 512	2359296
Conv	14x14x512	14 * 14 * 512	2359296
Pool	7x7x512	7 * 7 * 512	0
FC	1x1x4096	4096	102760448
FC	1x1x4096	4096	16777216
FC	1x1x1000	1000	4096000
Total weight:		138,344,128	

In this section we will use the MS COCO data set [7], the properties of this dataset are as follows: 80 objects with 330000 images. The data set is divided into 3 parts with 83 thousand images for training, 41 thousand images for cross-checking and 41 thousand images for accuracy assessment.

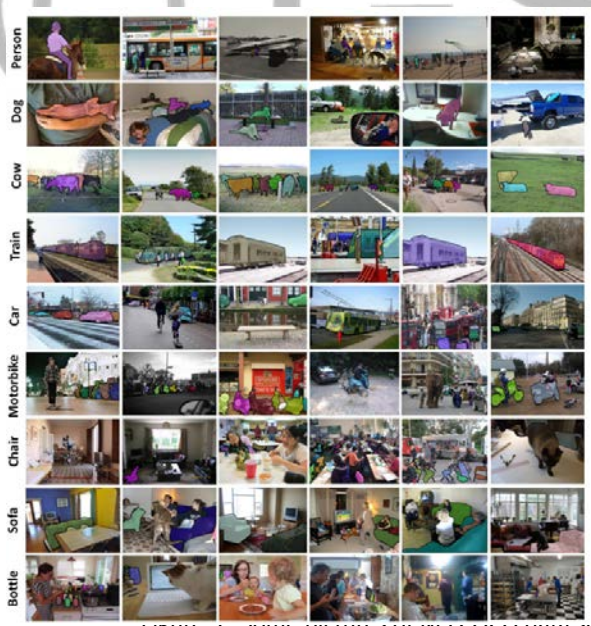


Figure 3 : Some pictures of MS COCO data set

IV. SIMULATION RESULTS

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

(a) (b)

Figure 4. Resulting image
(a) Basic CNN model (b) Faster RCNN technique model .

From 4 images results above, we can remark on the outstanding advantages of CNN using technology Faster RCNN compared to CNN without using Faster RCNN. Images processed via CNN without using Faster RCNN technique almost identify wrong input image with many objects. Based on the simulation results with the CNN technique, Faster RCNN presented above and many researches in [5], [6], [9], [10], we can evaluate techniques as follows:

Table 2. Objects recognition Evaluation

Technology	**Identification ability**	
	Image contains an object	**Image contains many objects**
CNN	✓	NOT

GSJ: VOLUME 7, ISSUE 6, JUNE 2019
ISSN 2320-9186

R-CNN	✓	✓
Fast R-CNN	✓	✓
Faster R-CNN	✓	✓

Table 3 . Locating evaluation

Technology	**Locate**	
	Image contains an object	**Image contains many objects**
CNN	NOT	NOT
R-CNN	✓	✓
Fast R-CNN	✓	✓
Faster R-CNN	✓	✓

Table 4. Processing time Evaluation

Technology	**Processing Time**	
	Image contains an object	**Image contains many objects**
CNN	~ 0.1	NA
R-CNN	~ 50	~ 50
Fast R-CNN	~ 2	~ 2
Faster R-CNN	~ 0.3	~ 0.3

V. CONCLUSION

CNN technique has partly solved the problem of object identification in images. However, the CNN technique only deal with images containing an object. It is almost impossible with multiple objects. But the Faster RCNN technology is born later, not only to overcome this shortcoming, but also to increase processing speed and ability to expand network identity. Thereby also showed the potential of this technique, here is the CNN technology using Faster RCNN to understand what objects in the image are and know the position of the object in the image.

References

[1] Karen Simonyan, Andrew Zisserman, Very Deep Convolutional Networks For Large-Scale Image Recognition , 2015.
[2] Y. LeCun and Y. Bengio, Convolutional networks for images, speech, and time-series, 1995.
[3] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks , 2012.
[4] Vedaldi, Andrea and Karel Lenc, MatConvNet-convolutional neural networks for MATLAB, 2014.
[5] Ross Girshick Jeff Donahue Trevor Darrell Jitendra Malik, Rich feature hierarchies for accurate detection and semantic segmentation Tech report (v5) , 2014.
[6] Ross Girshick, Fast R-CNN , In Microsoft Research, 2015.
[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollar, "Microsoft COCO: Common Objects in Context ", 2015.
[8] Ian Goodfellow and Yoshua Bengio, Deep Learning , 2016.
[9] Ren Shaoqing, Kaiming He, Ross Girshick, and Jian Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks , 2016.