ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 14, May 2016

# Object Recognition using Deep Neural Networks

M.Pon Priya
Department of Computer Science
SSN College of Engineering,
Chennai, India
ponpriya1418@cse.ssn.edu.in

Dr.T.T.Mirnalinee
Department of Computer Science
SSN College of Engineering,
Chennai, India
mirnalineett@ssn.edu.in

***Abstract*—Computer vision is a field that helps in analyzing and understanding images from real world. Image understanding can be achieved only if the objects in images are recognized. Object Recognition has been studied for many decades. While humans can recognize different images with little effort (even when they are of different sizes, scales, orientation and occlusion), the task is very complex for computers. Till today, we are unable to understand on what basis human brain recognizes different objects. Different feature extraction methods have been used to bring in some solution to this ongoing basic challenge. However, most of these methods have been hampered by the fact that these extractions are hand-written and hence depends on success of the programmer. Using deep neural networks for recognition brings in a better solution as the network learns the features automatically. Convolution Neural Network has been implemented on CIFAR-10 dataset bringing in 82% success rate.**

***Keywords*—Object Recognition; Deep Neural Network; CNN**

## I. INTRODUCTION

Computer Vision is the science that develops algorithmic basis by which useful information about the real world is extracted and its analyzed from an observed image (or set of video sequences) [1]. This is achieved from computations of general or special computers [2] and involves process of extraction, characterization and interpretation of information. This is a very difficult task as can been inferred from decades of research. Computer vision requires an understanding of human perception. Hence it's treated as a multidisciplinary filed that includes psychology, neurophysiology, cognitive science, artificial intelligence and pattern recognition [1].

Computer vision tasks can be categorized as (i) path planning and obstacle avoidance (ii) grasping and manipulation of objects and (iii) object recognition. But object recognition is the most important task as it is used to achieve the first two tasks. Object recognition is not however a simple task. The basis on which human brain performs recognition is still a mystery. Also modeling all the real world scenarios is highly complex. This is mainly because of its enormous computational challenge. The growth in field of computer processing has lead to advanced techniques to achieve recognition. Although Neural Networks has been in existence for a long period, the recent advancements in computers has made them feasible for implementations. Solutions to many challenging tasks are now being brought in by neural networks.

The remaining of the paper is as follows. Section II gives details about earlier approaches in recognition. Section III is about proposed systems, Section IV is experimental results, Section V discusses emerging applications and Section VI is Conclusion.

## II. RELATED WORK

Object recognition is the science of determining the identity of an object that is under observation. As the backbone of computer vision research, object recognition has been pursued for more than 4 decades [1][3][4]. Humans posses object recognition ability right from their childhood. From a single view, humans can identify and also categorize an object in an image despite changes due to pose, illumination texture, deformation or occlusion. Humans can also generalize a never seen before object from their previous observations. This is a daunting task for vision systems. Some of the major factors contributing to this difficulty are: pose of object to camera, differences in lighting, difficulty in generalizing based on set of exemplars. Due to these complexities, recognition systems, need to adopt representation models that capture these characteristics, so that it helps in developing procedures for identification. The representation or models can be either in 2D or in 3D.

### A. *Geometric Approaches*

The infancy of object recognition techniques took its root as a geometric based approach. The images were treated as pixels and planes and other geometric units of representation. Researches spanning close to three decades were done using these approaches for object recognition of both 2D and 3D images. The principle of geometric description of 3D object is that projected shaped can be accurately predicted in a 2D image using projective projection, which facilitates recognition using edge's boundary information (which is invariant to certain illumination change). Much concentration was given to extract geometric primitives (e.g., lines, circles, etc.) that are invariant to viewpoint change [2][5]. But results



All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 14, May 2016

have shown that such primitives can only be reliably extracted under limited conditions (controlled lighting variations and certain occlusion viewpoint). An excellent review by Mundy on geometry-based object recognition research can also be found in [3][5].

### B. *Appearance based Approaches*

When the geometric approaches were reaching its end of active period, approaches based on appearances were slowly becoming the norm. These methods achieved recognition by finding appearance space that is as close to input image as possible. Appearance based methods based on manifolds [4][6] and invariant intensity features [5][7] were developed. Software Library for Appearance Modeling (SLAM) was used to process images taken over different viewpoints and lightings. SLAM algorithm gave good results with high recognition rates on large set of objects. Appearance based models had serious limitations. They were not robust to clutter, occlusion or geometric transformations. This lead to the development of sliding window approaches. The main idea behind this approach was to use local features for object instance recognition. Large scale image search was done by combining local features, indexing and spatial constraints [8][6].

### C. *Feature based Approaches*

As more efforts was put to use to extract the local features for image search, feature-based approaches slowly became the norm. The main idea behind feature-based object recognition algorithms is to find interest points that occur at intensity discontinuity, that are invariant to changes in scale, illumination and affine transformation. Scale-invariant feature transform (SIFT) algorithm was proposed in early 2000s by Lowe [9]. This approach used difference of Gaussian for edge sharpening, gradient histogram to achieve a invariance due to rotation. For each descriptor, local image gradients are sampled and represented using orientation histograms. Many applications have been developed using SIFT descriptors [11][12]. Another major contributor for object recognition, parts-and-shape models also came into existence by early 2000s. In these types of algorithm, various parts of images are used separately to find out if and where an object of interest exists. The original idea was given by Fischler and Elschlager [13]. Pietro and Andrew [14] presented a method to recognize object's class from unlabeled and un-segmented cluttered scenes. The idea is to model objects as collection of flexible parts. An entropy-based feature detector was used to select regions within image. The model was then used in a Bayesian manner to classify images. By mid 2000s, bag-of-features models were becoming popular. These models extract features from images and learn the "visual vocabulary". These visual vocabularies are used to quantize the features. For image classification, either discriminative or generative learning methods are used depending on the vocabulary size. The success of these feature based approaches depends upon the strength of the feature extracted. With manual feature extraction, the success mainly depends on the programmer's expertise. Hence a system which automatically learns the features by itself would overcome this bias.

### D. *Deep Neural Networks*

Though humans can recognize objects with little or no efforts, the understanding of human brains works behind the scene to do this is yet a mystery to scientists. Without this knowledge, its difficult to make computers "see". Due to this difficulty, it was thought that neural networks was sought after to bring in some solutions. Although almost 4 decades of effort has been put into Artificial Neural Networks (ANN), their performance is yet to be standardized. The main reason was found to be the need for faster computations of operations and lack of proper training methodologies. In early 1980s, Yann Le Cunn [15] developed an algorithm to train neural networks. There were changes brought into existing ANN architectures so that back-propagation algorithm was used to train the networks. This was a breakthrough strategy that made it possible to use neural networks for training. Extensive research on usage of neural networks made Deep neural networks reality.

## III. PROPOSED SYSTEM

A major stumbling block of deep architecture was the large number of trainable parameters needed to train a network for an application. Modifications proposed to deep architecture to overcome this issue lead to the birth of Convolutional Neural Networks (CNN). CNNs take its inspiration from biological processes and variations of multilayer perceptrons. CNNs are slowly being used in wide applications like video recognition systems and natural language processing.

This paper proposes to use the CNN as a classifier. The major difference from other classification algorithms is that the CNN extracts the features by itself and this removes the problems of hand engineering of features. This lack of dependency on prior knowledge of the system and human effort in feature designing is the major advantage of using CNN. Also, compared to other techniques, CNN uses relatively less amount of image pre-processing.

### A. *Convolutional Neural Network*

Convolutional Neural Network took its inspiration from the findings of Hubel and Wiesels who worked on cat's primary visual cortex [16]. This paper identified the concept of cells with local receptive fields and complex cells which has been used in the CNN architecture. The CNN has three



All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 14, May 2016

design concepts that differentiate it from standard neural networks. The design concepts are:

- local receptive field
- weight sharing
- sub-sampling in spatial domain

These concepts are realized in CNN using three different types of layers, namely Convolution layer, pooling layer and fully-connected layer.

*1) Convolution Layer:* The Convolution layer or CONV layer as it is sometimes referred is the first layer after the input. The layer is made of set of learnable filters. During the forward pass, the filers are moved across the width and height of input volume. The 2D output from these filters are called activation maps. The activation maps are stacked along the depth voume to form the output. The neurons in this layer is connected only to a smaller region in input called the "Receptive Fields". The depth axis of the connectivity of these neurons is equal to the input volume. A value in the output volume is an output of neuron that looks at small region of input and shares paramters with neurons in same activation map. During back propagation, these filters are learnt intuitively. The major design factor of this layer includes designing set of filters/kernels. The Convolutional filter is a shared weight matrix, and is learned in a similar way to other weights. It is initialized with small random values. The output volume for a given input is calculated as follows:

$$L_2 = (L_1 - F + 2P)/S + 1$$

where, $L_1$ is the input width, $L_2$ is the output width, F is the filter dimension, P is number of padding units and S is the stride.

$$H_2 = (H_1 - F + 2P)/S + 1$$

where, $H_2$ is the output height, $H_1$ is the input height, F is filter dimension, P is number of padding units and S is stride.

The depth is maintained as the same as the input volume.

*2) Pooling Layer:* The pooling layer is inserted between CONV layers. The main purpose of the pooling layer is to reduce the spatial size of it input volume. This helps in reducing the parameters for computation. The layer works in every slice of the depth. Pooling can be done either through max operation or average operation, although max operation is more widely used now. Pooling however does not affect the depth of the input volume. During back propagation, the error is propagated to the highest value in forward pass (if using max pooling) or error is distributed across all input (if using average pooling). Using a max pooling with 2*2 filters with stride 2 down samples the input image by 2. The output volume for a given input is calculated as follows:

$$L_2 = (L_1 - F)/S + 1$$

where, $L_2$ is output width, $L_1$ is input width, F is filter dimension, S is stride.

$$H_2 = (H_1 - F)/S + 1$$

where, $H_2$ is output height, $H_1$ is input height, F is filter dimension, S is stride. The depth dimension is maintained.

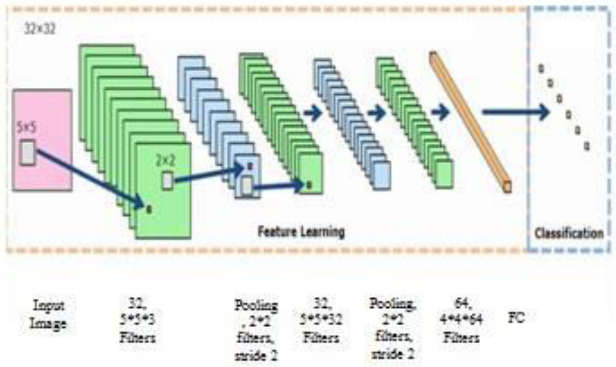


Fig.1. Proposed Architecture for Object Recognition

*3) FC Layer:* The FC layer is like a traditional neural network layer where every neuron is connected to every neuron in the previous layer. This layer has the extracted features from the previous CONV and pooling layers and the class labels from the input. The FC layer begins learning and through series of backward and forward propagations all the weights in the different layers. Once the weights are learnt during the training phase, during testing the images are classified based on the learning done during training.

## IV. Experimentation & Results

### A. Experimental Setup

The CNN implemented consists of 3 CONV, 2 pool layers and 1 FC layer. The first CONV layer consists of 32 filters each of size 5*5*3. The pool layer consists of 2*2 filters with



All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 14, May 2016

stride 2. This down samples the input by 2. The second CONV layer consists of 32 filters each of size 5*5*32. This is followed by a pooling layer similar to the first one. The third CONV layer is made of 64 filters each of size 5*5*32. This is followed by a FC layer. The FC layer learns through forward and back propagations. Stochastic gradient descent algorithm is used to learn the data. Once the training is completed, the class output score of test data is predicted by the FC layer.

### *B. Dataset*

The CIFAR-10 dataset has 60,000 color images. Each image is of size 32*32*3. The images all belong to 10 class of objects namely airplane, automobile, bird, cat, dog, deer, frog, horse, ship and truck. There are 6000 images per class. The dataset has 50,000 training and 10,000 test images.

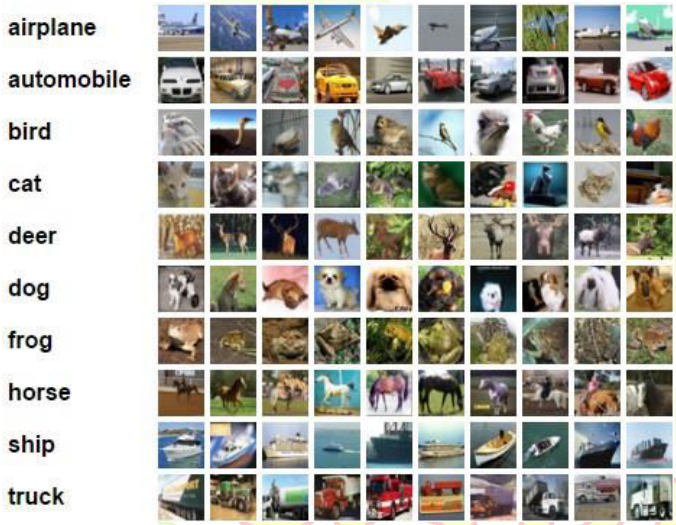


Fig.4. Sample images from CIFAR-10 dataset

### *C. Results*

The use of CNN as classifier for object recognition using CIFAR-10 dataset shows promising results compared to other classifier approaches. The experimentation shows 82% success rate compared to the previous success rate of 79.6%. The architecture implementation has 3 CONV layers and two pooling layers. However, increasing the depth by adding more layers will yield better results. The major advantage of this implementation is there is no prior knowledge of the system required unlike other implemented systems. This advantage will be of great use for implementations in more complex scenarios.

The following table compares the accuracy of various methods as against the CNN.

TABLE I. Comparison of Different Approaches

| Method | Accuracy |
|---|---|
| Support vector Machines [18] | 39.5 |
| SIFT [9] | 65.6 |
| Fine-tuning GRBM [20] | 64.8 |
| mcRBM-DBN [17] | 71.0 |
| K-means (Triangle, 4K Features) [19] | 79.6 |
| Convolutional Neural Networks (proposed work) | **82.0** |

## V. EMERGING APPLICATIONS

Object recognition is only the first step of computer vision systems. Commercial applications based on these recognition systems are slowly getting implemented in recent times. Some noteworthy applications include implementations by Google and Baidu for personal image search. Singapore based startup company ViSenze, allows users to search e-commerce platforms visually. ZZ Photo is a startup based out of Ukraine and it helps in sorting images stashed in our PCs. NVIDIA is already working to bring in powerful GPUs to increase computation capacity so that medical image analysis systems can be the norm of the future. Google is working on self-driving cars which requires very accurate and faster recognition.

## VI. CONCLUSION

The Convolutional Neural Network was implemented for CIFAR-10 dataset. The results show that CNNs achieving better results compared to other existing techniques. This promising result strengthens the fact that CNN can be used for object recognition tasks and can be further implemented for more complex datasets to achieve recognition.

## *References*

[1] Ana Fred, Terry Caelli, Bob Duin, Dick de Ridder, "Advances in Pattern Recognition", Springer, 2004.



All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 14, May 2016

[2] Martial Hebert and Katsushi Ikeuchi and Hervé Delingette, "A Spherical Representation for Recognition of Free-Form Surfaces", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, pp. 681-690.

[3] D. Marr, "Vision," in W. H. Freeman and Company,1982.

[4] Ullman, S., "High-Level Vision: Object Recognition and Visual Cognition," MIT Press,1996 .

[5] Mundy, Joseph L. and Zisserman, Andrew, "Geometric Invariance in Computer Vision," MIT Press, 1992.

[6] Murase, Hiroshi and Nayar, Shree K., "Visual Learning and Recognition of 3-D Objects from Appearance," Int. J. Comput. Vision, vol. 14, pp. 5-24, January 1995.

[7] Cordelia Schmid and Philippe Bobet and Bart Lamiroy and Roger Mohr, "An Image Oriented CAD Approach", Lecture Notes in Computer Science, Springer, volume 1144, pp. 221-245, September 1996.

[8] J.Philbin, O.Chum, M.Isard, J.Sivic, A.Zisserman, "Object Retrieval with Large Vocaubalries and Fast Spatial Matching", IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[9] Lowe, David G., "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, volume 60, pp.91-110, November 2004.

[10] [11] J.Sivic, A.Zisserman, " Video Google: A Text Retrieval Approach to Object Matching in Videos", Computer Vision, 2003.

[12] Nister, D., Stewenius, H., " Scalable Recognition with a Vocabulary Tree", IEEE Conference on Computer Vision and Pattern Recognition, vol 2, pp.2161- 2168, June 2006.

[13] M.A.Fischler, R.A.Elschlager, "The Representation of Matching Pictorial Structures", IEEE Transactions on Computers, vol. c-22, pp.67-92, January 1973

[14] R.Fergus, P.PErona, A.Zisserman, "Object Class Recognition by Un-supervised scale-invalriant Learning", Computer Vision and Pattern Recognition, vol 2, pp. 264-271, June 2003.

[15] Y.LeCun, " A Theoretical Framework for Back-Propagation", Proceedings of 1988 Connectionist Models Summer School, 1988.

[16] Hubel DH, Wiesel TN, "Receptive fileds of cells in striate cortex of very young, visually inexperienced kittens", J.Neurophysiol,pp.978-993, 1963.

[17] M.Ranzato and G.Hinton., "Modeling pixel means and covariances using factorized third-order boltzmann machines", CVPR, 2010.

[18] L.Bo.X.Ren and D.Fox, "Kernel Descriptors for Visual Objects", NIPS, December 2010.

[19] A.Coates, H Lee, and A.Ng, "An Analysis of single-layer networks in unsupervised feature learning", NIPS, 2010.

[20] M.Ranzato, K.A., and G.Hinton, " Factored 3-way restricted boltzmann machines for modeling natural images", CVPR, 2010.



All Rights Reserved © 2016 IJARBEST