International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Volume 7 Issue 4, April-2018, Impact Factor: 3.578

Object Recognition Based on Deep Neural Networks

Dhiraj Khurana

Computer Science & Engineering, UIET, MDU, Rohtak, India

ABSTRACT

Object recognition with the help of neural networks is an active research area. For detecting objects in images or videos and classifying them; Convolution Neural Network (CNN)- a type of neural network is often used. Firstly, an overview of basic neural network structure is covered and then an introduction to the architecture of CNNs is briefed.

SUMMARY

The process of identifying a specific object in a digital image or video is known as object recognition. Object recognition is easy for human beings as we can easily identify objects and their types. We can easily recognise various objects in an image and can identify them. Recognising objects in a video is kind of a challenge but still we can identify objects if asked specifically. The difficulty surfaces when we want a machine to use artificial intelligence and recognise various specified objects. For a machine to recognise objects in an image or a video is sort of a difficult task. Recognising objects is one of the most fundamental tasks in computer vision. Various methodologies can be used to recognise objects. Recognition of objects can be done by matching, applying algorithms for recognising patterns by using feature-based or appearance-based techniques, learning. Objects can be recognised by using various models like machine learning and feature extraction models, deep learning models like CNNs, gradient-based approaches, template matching, blob analysis, image segmentation etc.

Deep learning is a machine learning technique that uses deep neural networks i.e. networks with many layers to make predictions. Deep learning plays a vital role in image recognition. Suppose we are to recognise a handwritten number- 6. First we will have to find the matrix and features of the digit that can be helpful in distinguishing that digit from other digits. Then we calculate those features and train our set and then apply a classifier to recognise it. The goal of Deep learning technique is to provide end to end learning. The images are taken as input and all the feature extraction, training and classification is done on same image. Various deep learning techniques can be used in image recognition like convolution neural networks (Conv Nets, CNNs), regression, directed acyclic graph (DAG) network topologies etc.

Humans have a tendency to learn new things by observing and then seeing various examples of same type. Humans have the efficiency to understand things by observing them. Deep learning is a technique that helps computers learns by example. Deep learning can be used to imitate this quality of human behaviour to recognise objects. Deep learning is applied in many things and is a key technology. With the help of deep learning we can make a computer to learn and perform tasks from images, videos and sounds. In order for computer to recognise our data, first we need to train it very well. For training the models, large datasets are used that contain labelled data. A neural network is made by a large system of interconnected neurons. These neurons can exchange messages among each other. While the network undergoes training, Weights are assigned to connections. Multiple layers of feature detecting neurons are present in a network.

Dataset for image recognition is taken from a large database. For a computer to recognise an object in an image, we need to train it with the best of the data available. This is why dimension and pixels of images required in learning are specified. When we want to teach the machine specific hand writing, we take the high resolution images of the data and then remove noise from the images. When the images are very clear, only then they are used to train the machine. Generally large datasets are available over the net. These data sets are called image nets. Image net is a large database that was proposed for the use in object recognition. Image net contains over 20 thousand ambiguous categories of images. Any category say cats,

pens contain hundreds of related images. Image Net uses trimmed images that have been recognised by the machine. Image Net conducts a challenge every year that lets users participate in a contest in which they have to classify a category such that machine recognises it. Then they reduce the set of their images so that the best quality of images is left. In this way if we take image dataset from image net we can be sure of the quality of images. The actual images that are found in image net are not really owned by them. Image Net has been conducting this challenge since the year 2010 .In ILSVRC the category is object detection and classification. There are three challenges in this competition. First is object localization in which contestants have to pick up the top 5.

This is the original challenge of ILSVRC other two challenges were added recently in the competition. The training dataset is of nearly 1.2 million images and there are about 1000 categories that the contestants can choose from. And the test image set consists of about 150,000 photographs. So each program lists its top 5 labels based on test images. The program with the minimum average error is the winner. Second challenge is of detecting objects. Second challenge is further split into two categories; one is detecting objects in images while other is to detect objects in videos. The participating teams are allowed two submissions per week and there are no regulations on number of neural network layers used and there parameters and learning scheme have to be based only on the training set of the participants. Even though ILSVRC was started in 2010, significant results were noticed when deep learning was applied on the sets.

In 2012 Alex Net (winner of 2012) used neural networks for the first time and got an accuracy of 83.6%. They used 8 layer deep neural network, 5 convolution layers, 3 fully connected layers, 60 million parameters, which showed a significant increase in the accuracy. They trained this set for 6 days and used two nvidia GTX-580 with 3gb memory. In 2014 Google Net (inception V1) won the competition with the accuracy results of 93.3% in object localization. They used 22 layer neural networks with 5 million parameters and trained the set for 1 week in Google DistBelief cluster. In 2015 Microsoft ResNet won this competition with the accuracy of 96.5%. This win was significant because for the first time the machine crossed the human accuracy of detecting objects. The accuracy of human beings to detect an object in the image is 94.9% but in 2015 Microsoft ResNet made a program that surpassed the human accuracy. Microsoft ResNet used 152 layers of Deep learning neural network and trained the set for approximately 3 weeks on 4 nvidia tesla K80 GPUs by using a combined processing capability of 11.3 GFLOPs.

Convolution Neural Networks (ConvNets or CNNs) are the most effective type of neural networks when it comes to image recognition and classification. They are used to power vision in self-driving cars and robots.CNNs are successful in identifying traffic signs, objects and faces. Therefore, CNNs are widely used in machine learning today. The name ConvNets was derived fromconvolution operator. Convolution operator in convNets is used to extract features from image. A convolution network is an order of layers, and each layer modifies one volume of activation function to a further volume by applying differential function. The convolution layers exert operations on inputs and pass the results to next layers. Every image is an exhibit of pixels ordered in a specific manner. The image changes with a minor change in arrangement of pixels or pixel colour.

The machine breaks the image into a matrix of pixels and then it stores the information like the coding of the colours of the pixels in their typical positions. The machine then stores the weights along with the information in a manner that is quite difficult to understand by humans but can be understood and retrieved by machine whenever required. What the machine does is, it simply takes the input image and defines a weight matrix. The input is then convolved so that specific features could be extracted from the image. The information about the pertaining order of the image is maintained. To build Conv-Nets three main layers are used; Convolution layer, pooling layer and fully-connected layer. Each layer takes an input volume and transforms it into output volume by applying a differential function on input. The layers used may or may not have parameters.

Fundamental part of a convolution neural network is its convolution layer. A set of learnable filters are present in a convolution layer. Each filter is made up of some space and it slides across width and height of the input volume and computes the dot product of the entries of filter with the input at any position. These dot products are stored in a matrix form. These outputs are used to make an activation map.

3	0	1	2	7	5
1	5	8	9	3	1
2	7	2	5	1	3
0	1	3	1	7	8
4	2	1	2	4	8
2	4	5	2	3	7

Fig. 1- Image

1	0	-1
1	0	-1
1	0	-1

Fig. 2- Filter Image

-5	-4	0	8
-10	-2	-2	3
0	-2	-4	-7
-3	-2	-3	-16

Fig. 3- 4*4 Image

3	0	1
1	5	8
2	7	2

Fig 4- Convolution 1

0	1	2
5	8	9
7	2	5

Fig 5 - Convolution 2

As the filter is of 3x3 matrixes we take 3x3 matrices from our matrix of input image and compute the dot product. The matrix keeps sliding across until the whole matrix is covered. The matrix slides on pixels and the number of pixels it slides through to cover new pixels is known as **stride.** For example if the weight matrix slides one pixel at a time we call it a stride of 1. The size of image reduces as the stride increases. If stride is big then the loss of data is more so we can apply padding. We can add rows and columns of 0's on each end of the matrix to restore the previous dimensions.

After this process some parts of the image gets cropped so in order to restore the lost padding is done. With padding shrinking problem gets solved. We require padding to maintain the size of the image. Padding helps us to make the output image size same as that of the input image size. In terms of how much we should pad, we have two choices; valid and same convolutions.

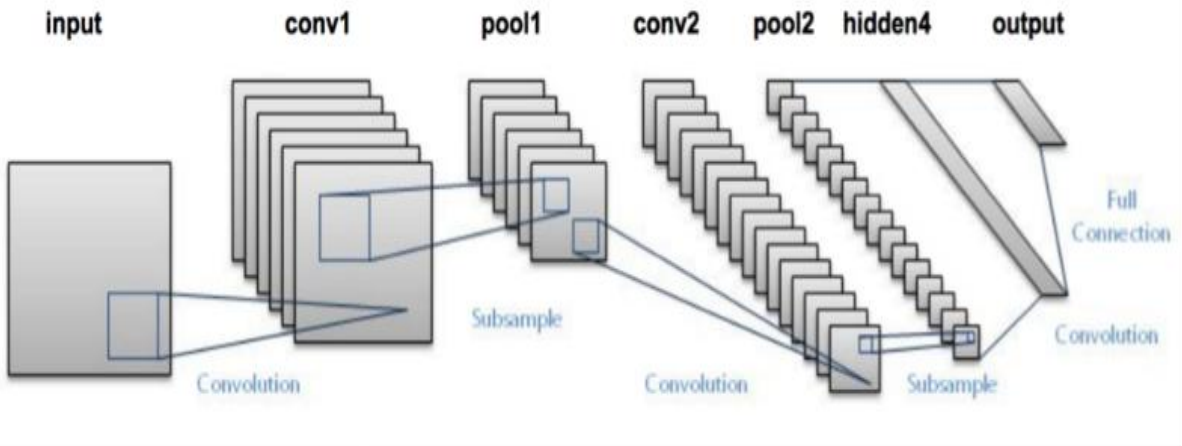
Valid (no padding)
nxn * fxf→ n-f+1
6x6 * 3x3 →4x4

Same (pad so that the output size is same as that of the input size)

n +2p-f+1 xn+2p-f+1
p= (f-1)/2
eg. For 3x3 p= (3-1) /2=1

After the features are detected they are classified. Features can be directly classified by using a classifier like softmax but if image is very large or the feature matrix is very big then detecting features can be cumbersome. Pooling layer works independently on every layer. Usually max operation is applied and a stride of 2 is taken. It discards about 75% of activations. In such case every max operation takes a maxim of four numbers, i.e. 2x2 regions. The dimensions remain unchanged...

Pooling, basically down-samples the volume of the output without changing its dimensions. Pooling layer is included after successive convolutionlayers. Other than max pooling we can also apply average pooling. In average pooling average of the numbers is taken and then a down-sampled matrix is created.



The third type of layer in a CNN is fully connected layer. In fully connected layer the neurons have full connections to all the activations. Their functions can be computed with a matrix multiplication. Then activation functions are applied on

them and later these datasets are used for evaluation. Previously two-dimensional outputs are converted into a one-dimensional feature vector at this stage. This layer uses previously learned features for final classification.

REFERENCES

[1]. R.C. Gonzalez and R.E. Woods, Diptal Image Processing, Addison-Wesley, pp.571-657, 1992.
[2]. K. Fukushlma and S. Miyake, "Neocognitron: A New Algorithm for Pattern Recognition Tolerant ofDeformation and Shifts in Rotation," PatternRecognition, 15(6), pp. 455-469, 1982.
[3]. K. Fukushima, SMyake, I. Takayulu, "Neocognitron: A Neural Network Model for aMechanism of Visual Pattern Recognition," IEEE Trans, Systems, Man and Cybemetics, SMC-13(5), pp. 826-834,1983.
[4]. D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning Internal Representations by ErrorPropagation," Parallel and Distributed Processing: **As47** Explorations in the Mcrostructure of Cognition, Vol. 1,MIT Press, pp.318-362, 1986.
[5]. D.G. Elliman and R.N. Banks, "Shft Invariant Neural Network for Machme Vision," IEEE Vol. 137, **Pt.**1, No. 3, pp. 183-187, June 1990.
[6]. A. Gosh, N.R.Pa1, **S.K. Pal, "Self-Organization** for Object Extraction using a Multilayer Neural Networkand Fuzziness Measure," IEEE Trans on Fuzzy Systems,Vol. 1, No. 1, pp.54-68, FebruaIy 1993.
[7]. P.Im, The Pattem Recognition of Industrial Partsusing a Deterministic Decison Rule Based Strategy, Masters Thesis, Victoria University of Technology, 1993,
[8]. S.K.Pa1, "Fuzzy Sets in Image Processing andRecognition," IEEE International Conference on Fuzzy
[9]. Systems, San Diego, California, USA, March 8-12, 1992.
[10]. M.A. Dalal, N.D. Harale, and U.L. Kulkarni, "An iterative improved k-means clustering,", *ACEEE International Journal on Network Security*, vol. 2(3), pp.45-48, 2011.
[11]. D.T. Pham, S.S. Dimov, and C.D. Nguyen, "Selection of K in K-means clustering", *IMechE 2005*, vol.219, pp.103-119, 2014.