

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Survey on Convolutional Neural Networks

Priyanka S. Padwal, Dr. Neeta A. Deshpande

Department of Computer Engineering, D.Y. Patil College of Engineering, Akurdi, Pune, India

ABSTRACT: Deep learning has progressed quite fast over the past few years and has become omnipresent in almost every field today. Convolutional neural networks are one of the most extensively used and researched neural networks which have found applications in almost every domain from image processing to natural language processing. They have mastered providing exemplary performances in visual challenges and proved to be better at major recognition tasks than most of their neural network contemporaries, along with being cost-effective and comparatively easier to train and having less number of parameters and lower error rates. This paper discusses the structure and working of convolutional neural networks along with techniques advanced to improve its efficiency. Also, various applications of convolutional neural networks in separate fields and their development over time have been discussed and the effect of depth on their performance is analyzed.

KEYWORDS: Convolutional neural networks, deep convolutions, natural language processing, activation functions, regularization techniques, image recognition.

I. INTRODUCTION

Convolution Neural Network:

Convolutional neural networks (CNNs) are one of the most widely used type of deep artificial neural networks that are used in various fields such as image and video recognition, speech processing as well as natural language processing. These networks have been inspired by biological processes, such as working of the visual cortex in cats and spider monkeys. Hubel and Wiesel, in the year 1969, studied the same to classify the cells in the cortex of these as- simple, complex and hypercomplex. The complex cells were found to have a receptive field, i.e., the area of response to stimulus, approximately twice as large as that of a simple cell. Hence the idea of using translational invariance in order to recognize visual imagery was realized. This property stated that the exact location of an object in an image was of less importance rather than detecting the object. Using convolutional neural networks is better than fully connected networks in many applications since instead of every node in a layer being connected to every other node in the previous layer, the former has every node in the m 'th layer being connected to n nodes in the $(m-1)$ 'th layer, where n is the size of the receptive field of the CNN. This reduces the total number of parameters in the network and hence prevents overfitting and also ensures a built-in invariance. The first convolutional network was introduced by LeCun et al, in 1998, known as Lenet-5. It was a 7 layered network which was used to digitize hand written digits on checks in banks. It introduced 3 basic architectures used in the network - shared weights, local receptive fields and pooling.

II. RELATED WORK

Convolutional neural networks have been discussed along with its various applications where image recognition continues to be the primary field where they are used. One of the major revolutions in the applications of convolutional neural networks was made in 2012 when Alex Krizhevsky et al, applied CNNs to image classification- the most widely used application of CNN in today's scenario [3]. One of the major reasons that this had been achieved was the introduction of GPU computing which ensured that a vast dataset of images could be used to train and test the networks, which allowed a good case of generality and also fast computing due to increased degree of parallel processing. The network proposed by them, called AlexNet, was used to classify a total of 1.2 million images in a 1000 different classes. The error rates recorded were 17.0% (top-5) and 37.5% (top-1) which were a lot better than the previous technologies used in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) for image

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

classification. Regularization techniques such as drop-out were used to prevent over-fitting in the fully connected layers of the network along with data augmentation techniques. In 2013, Matt Zeiler and Rob Fergus, proposed a CNN called ZFNet which had a top-5 error rate of 14.8% and was the winner of ILSVRC [4]. Here, instead of relying on just trial and error, more focus was given on how the processing is exactly done by the layers. This was achieved by using deconvolutional networks which used the components of the convolutional network in reverse, in order to visualize the exact working of the CNN by mapping feature activations to the input pixels. In this network, the output image pixels had a path back to the input. Filter size was reduced from 11x11 (of AlexNet) to 7x7, receptive window was smaller and stride of convolution reduced from 4 to 2.

Image Database Classification Using Neural Network with CBIR Technique:

Today's scenario image classification and retrieval has become the most challenging and important research work for a wide range of applications like architectural and engineering design, art collections, crime prevention, geographical information and remote sensing systems, intellectual property, medical diagnosis, military applications etc. Locating a desired image in a large and varied collection of database is a considerable impediment for the researchers. Early work on image retrieval can be dated back to the 1970's but they were not based on visual features but textual annotation. Hence it couldn't support task dependent queries. Traditional methods posed such problems which have led to the image retrieving techniques based on content or features such as texture, color and shape. This is called the content based image retrieval (CBIR) technique. Since early 1990's, many systems have been proposed and developed, like QBIC, Virage, Pichunter, Visual SEEK, Chabot, Excalibur, Photo book, Jacob, UC Berkeley Digital Library Project. Most of the above mentioned systems and much of the past research have procured the CBIR from its infancy to the matured stage. Even though in some cases these systems exhibit substantial outcomes but still have limited efficiency. They have concentrated on the extraction of the low-level features. They emphasized on the explicit features of the images. These features are automatic but omit to study the implicit meaning behind the image. The hidden meaning or the semantic idea of the image and video can be interpreted solely by analyzing its contents. This leads to the semantic gap. Until and unless, we study the various implicit meanings of images and videos, which cannot be discernible by the content the semantic gap will not reduce. The performance of these image retrieval systems can be improved if features extracted are used to train artificial neural networks and classify the image based on trained variables. Hence an automatic technique is proposed to classify and categorise images based on supervised learning where collection of trained images are given. But, the problem is to identify a new upcoming unlabeled image. Each instance of the training will contain class specific labeling of the images and their descriptive feature vectors. Hence this paper combines the approach of both the CBIR technique along with the neural network training. The ANN network is trained and the set of image database are labeled using the feedback weight vectors from the users which would distinguish the image as appropriate or inappropriate. The image labeling is done once and the neural network is trained for the concerned database and therefore it can group and classify the images according to the feature vectors. In section 2 theoretical contemplation is elaborated, section 3 explains the feature vector extraction, section 4 brings out the proposed CBIR technique and

III. COMPONENTS OF CONVOLUTIONAL NEURAL NETWORKS

The network architecture of CNNs comprises of various different components which are described below.

3.1 Layers in a Convolutional Neural Network

Artificial neural networks consist of various layers between the input and output layers. These layers known as the hidden layers, majorly consist of the following in the case of convolutional neural networks.

3.1.1 Convolutional Layer

Each convolutional neural network consists of different number of convolution layers depending on network requirements. The first convolutional layers are responsible for learning low level features such as edges, corners, etc. The output of these layers are often fed to other convolutional layers which learn higher level features. Each neuron in

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

this layer is connected to only a limited no. of neurons in the previous layer. The no. of neurons they are connected to is known as the receptive field of the convolutional layer. These layers comprise of a filter which has a typical size of $n \times n$. The filter is slid (or more precisely convolved) during the forward pass across the width and height of the input volume and dot products are computed between the entries of filters and the input position which comprises the final feature map. However, multiple filters are used for convolution since images have multiple features, and this signifies the depth of the feature maps (which is a hyper parameter) which is equal to the no. of filters used.

3.1.2 Softmax Layer

The softmax layer is generally a linear layer with a softmax classifier which converts the activations into values between 0 and 1, such that their sum is 1. It helps determine the final output of the CNN by finding the output of a recognition or classification task by determining the class or category having the highest probability value. It basically represents a probability distribution.

3.1.3 Pooling Layer

The pooling layer is used to decrease the spatial size or the resolution of the image in order to decrease the number of parameters, hence decrease the computational burden. It does so by reducing the no. of connections between the convolutional layers. They are usually alternated between the convolutional layers.

3.1.4 Fully Connected Layer

These layers are present in varying number in a CNN, generally right before the softmax layers. Every neuron in this layer is connected to every neuron in the previous layer. It is used to achieve linearity in the networks but can be replaced or converted into convolutional layers to turn the system into a fully convolutional network (FCN).

IV. BASIC WORKING OF CONVOLUTIONAL NEURAL NETWORKS

Each neuron in a convolutional neural network in the m th layer is connected to n neurons in the $(m-1)$ 'th layer, where n is known as the receptive field of the CNN. Stride of a convolution basically refers to the step size used in convolution operation. A new feature map is obtained by first convolving the input with a weight/kernel. The convolution occurs as follows-

$$Z_{i,j,k}^m = w_k^m * x_{i,j}^{m-1} + b_k^m$$

For the k^{th} feature map in the m^{th} layer. (i, j) is the location in the m th layer, where the value in the feature map has to be calculated [5].

w_k^m and b_k^m are the weights and bias for the k th kernel in the m th layer and $x_{i,j}$ is the input to be convolved. The convolved output is then subjected to a non-linear activation function which yields the final feature map output. Note that the complete set of feature maps are obtained by convolving the input with several weights/kernel. For e.g, in the case of colored images using an RGB scale, separate kernels representing different intensities of red, green and blue colors are used. However, at a time, the same weights are shared throughout the entire spatial locations of the input. This refers to one of the most central characteristics of CNNs, i.e, the shared weights concept. This concept makes the network less complex and results in fewer parameters. Several convolutional layers are arranged such that the first layer detects low level features such as edges and corners and the latter layers detect high level features such as objects. The output of the convolutional layers after having the pooling and regularization functions performed on them passes on to the Softmax layer which usually has a fully connected layer as its previous layer. The fully connected layer can also be replaced by a 1×1 convolutional layer. The Softmax layer, then classifies the input provided to it into different classes, which becomes the final output. Deconvolutional layers have also been used in convolutional networks in order to visualize the exact way that the CNN works. The layers perform the reverse of the convolution operation by mapping features back to the input pixels by providing a path for the output to be projected back to the input. A sensitivity analysis is also performed to see which parts of the input are more necessary for the recognition or classification process [6]. Any dataset that is used for training the network is usually divided into 3 sets- the training set, the test set

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

and validation set which are used to initially train the network on some images, test with fresh set of images and validate the result obtained using another set of images respectively.

V. APPLICATIONS

CNNs have been proven to be useful in many applications and have achieved better performances in various fields compared to other neural network architectures.

5.1 Image Recognition

The use of CNNs for image detection and recognition has been in place since a long time . Use of CNNs for classification purposes, though introduced later and has been widespread. Used CNNs for generic object detection and recognition while SVM(Support Vector Machines) were trained on the features learned by the CNN to be used for classification purposes.adapted a GPU implementation of image classification task using CNNs , such that the network was flexible as well as fully online,i.e., the weights were updated after each image. This implementation performed 10-60 times faster computationally than a compiler optimized CPU implementation. However, a revolutionary change in the image classification task by CNNs was brought by in 2012.It achieved significantly lower error rates on ILSVRC(2012) than the previous state-of-the-art.It used a very efficient GPU implementation which involved using multiple GPUs in parallel. Use of efficient regularization techniques like dropout was made along with data-augmentation to reduce overfitting. Various improvements on this classification attempt have been made by reducing the filter size and stride of convolution , or increasing the depth as well as width of the network to achieve an even lower error rate on the classification task in the ILSVRC.

Object detection deals with localization of objects as well the ability to deal with the cases where the same object occurs in an image at more than one instances.uses a DCNN for a base feature extraction. The network is used to predict a series of bounding boxes such that a fixed score exists for each box depending on the probability of finding an object of interest. A similar DNN based regression approach was used in , however it could not scale up such that it would include multiple classes as the cost of performing a single regression was high. Another approach is to generate certain region proposals during test timewhich are category independent. A CNN computes the features for each region proposal by using image warping for a fixed size input regardless of the region shape using a linear SVM classifier that is category-specific for classification. It also uses a bounding box regression to reduce errors.

5.2 Text Recognition

CNNs have proved to be useful in many natural language processing (NLP) tasks such as sentence modeling,search query retrieval,various text classification tasks,etc. Sentence classification by CNNs is done using a single convolution layer along with the use of pre-trained word vectors.It consists of different model variations,one of which (referred to as CNN-rad) randomly initializes all the words in the beginning and are then slowly modified as training continues.Other models (one of which is CNN-static) use a set of vectors pre-trained on a 100 billion words from Google News in this case.The classification model can be used for tasks such as finding four similar words to a given word and performed remarkably well.Most sentence classification applications constitute splitting a sentence in sub-sentences and then performing various classification tasks using labels on these sub-sentences, since the granularity of the sub-sentence can positively be used to represent the entire sentence.However, the granularity of any given sub-sentence isn't good enough to be used across all sentences,hence [50] uses a multi-channel variable size convolutional network (MVCNN) which uses variable-sized filters to be used with multigranular phrases.

Unlike other tasks like character recognition for scanned documents, recognizing text from an unconstrained image consists of various variations of font,texture,background,etc.As a result, many such applications use hand-crafted features to detect and later recognize these images. uses CNNs to recognize text by implementing unsupervised feature learning algorithms which extract features from the images automatically.Text detection is performed using a detector window which is slid across the entire image to identify texts on which operations like word segmentation are performed and later recognition The network uses a lexicon (which a collection of pre-defined candidate words).The lexicon is used to recognize the text techniques like NMS and beam search along with the application of two separate

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

CNNs with 2 convolutional layers each, for both detection and recognition operation. However, the recognition step in the approach above uses a set of known words (lexicons) and would not generalize well to an image containing unseen text or text structurally different than that used for training, like alpha-numeric values used for addresses are license plates

5.3 Handwriting Recognition

The task of recognizing handwriting was one of the first applications of image analysis by convolutional neural networks. Le Cun et al used CNNs to recognize handwritten digits on checks and hence, relied on automatic learning rather than hand-crafted features which improved the task efficiency in terms of computational cost as well as time. One of the major problems for recognizing handwriting wasn't identifying individual characters, but separating the characters from the ones near them in order to form a word or a sentence. This used to be done with the help of Heuristic Over-Segmentation which consisted of generating a large no. of potential cuts between characters and selecting the best combination by manually labeling all the character hypothesis. This proved to be cost-ineffective as well as time consuming. Handwriting can contain size, position, writing style as well as slant variations. Invariance to these features could in principle be obtained by training fully-connected neural networks but it would require a very large no. of training instances and a large no. of parameters. However CNNs have a built-in shift invariance which can achieve this task easily.

The handwriting recognition task today is not just recognizing the handwriting and converting it into digital text but identifying the identity of the writer. The recognition can be offline (static) or online (dynamic). In the former, the handwriting is obtained after the writing process has been completed by scanning the document containing the handwritten text. In the latter case, the identification process occurs while the handwritten text is being obtained. [54] used CNN to perform signature verification of individuals such that it could differentiate between genuine signatures and skilled forgeries. The system worked by using skilled forgeries as well as genuine signatures for a subset of users to train the network by identifying gestural cues as difference between the two, so that the learned features can perform well in identifying forgeries for unseen users. Handwriting analysis can be Writer Dependent (WD) or Writer Independent (WI). For Writer Dependent classifiers, a training set is constructed separately for an individual user with positive and negative samples being fed to it with respect to a particular task for an individual user. Writer Independent classifiers, on the other hand, use a single model for training for all users by using a dissimilarity vector (which represents the difference between features of the query and template handwriting) as input. Generally, the recognition approach constitutes segmenting handwritten text or words into smaller parts for recognition after which the final recognized result consists a combination of individually recognized parts or character which are then fed into a word recognition module which mostly implements a dictionary search algorithm. [55] used multiple convolutional neural networks in order to identify a large character set. Each CNN recognized a part of the character class (such as alphabets, digits, etc.) and also different languages which can then be combined by programming algorithms to create a rather flexible classifier which could be used to recognize differential big character classes.

5.4 Video Recognition

CNNs generally use 2 spatial axes for the representation of images. However, representation of videos using this configuration becomes difficult since they consist of still images changing with respect to time. Videos also provide natural data augmentation (jittering) for a single image classification. Action recognition in videos primarily consisted of two steps. First, performing motion recognition using certain handcrafted features and second, classifying these features. These approaches also tend to make certain assumptions which would certainly not always exist in a real world environment. Furthermore, the handcrafted features are highly problem dependent and cannot be generalized.

The earlier approaches to using CNNs for action recognition comprised of dividing the video into multiple contiguous frames of images and object detection took place within these still frames. However, there was no representation of information of the motion that took place within these frames, i.e, motion detection did not exist. Hence, various approaches have been proposed for representation of videos using CNNs incorporating the information in the time domain.

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

In the first approach, the first convolutional layer itself is extended in time. In the second approach, two separate 2-dimensional single frame networks are developed and merged later at the first fully connected layer

5.5 Face Recognition

Face detection using CNNs has been used since a long time. Local image sampling was combined with a convolutional neural network to extract larger features successively with layers arranged in a hierarchical manner along with a self-organizing map (SOM) neural network which primarily provided dimensionality reduction and invariance to minor changes in images to perform face recognition. However, it used a rather small dataset with limited individuals. Other approaches consisted of facial detection, rather than recognition which showed whether a given image has a face in it or not by dividing the image into small windows and detecting the presence of a face within each window. However, this approach detected only frontal views of faces and also only if the faces were upright. In, Garcia et al used CNNs to detect faces as an improvement to their previous effort in which used skin color for detection since, the previous method could not be used in many applications such as gray scaled images. Hence, CNNs proved to be robust even in the cases of various poses, lighting as well as facial expressions and also faster than previous approaches, one of which involved using SVM.

VI.COMPARISON TABLE

Algorithm	SVM Classifier	CNN(Our Approach)
1	SVM are linear classifiers.	CNNs are non-linear.
2	Not much restrictions on what kind of data it will work with.	CNNs work well with data having spatially recurring patterns. Images, speech.
3	No scope of representatives station learning.	CNN classifiers feature vectors are useful in many other problems. Very good representation
4	No way to increase model complexity	No way to increase model complexity
5	use SVM when u have less data	CNN when you have more.

There are at least couple differences:

1. CNNs are designed to work with image data, while SVM is a more generic classifier;
2. CNNs extract features while SVM simply maps its input to some high dimensional space where (hopefully) the differences between the classes can be revealed;
3. Similar to 2., CNNs are deep architectures while SVMs are shallow;
4. Learning objectives are different: SVMs look to maximize the margin, while CNNs are not (would love to know more)

Pros of NN:

- They are extremely flexible in the types of data they can support. NNs do a decent job at learning the important features from basically any data structure, without having to manually derive features.
- NN still benefit from feature engineering, e.g. you should have an area feature if you have a length and width. The model will perform better for the same computational effort.
- Most of supervised machine learning requires you to have your data structured in observations by features matrix, with the labels as a vector of length observations. This restriction is not necessary with NN. There is fantastic work with structured SVM, but it is unlikely it will ever be as flexible as NNs.

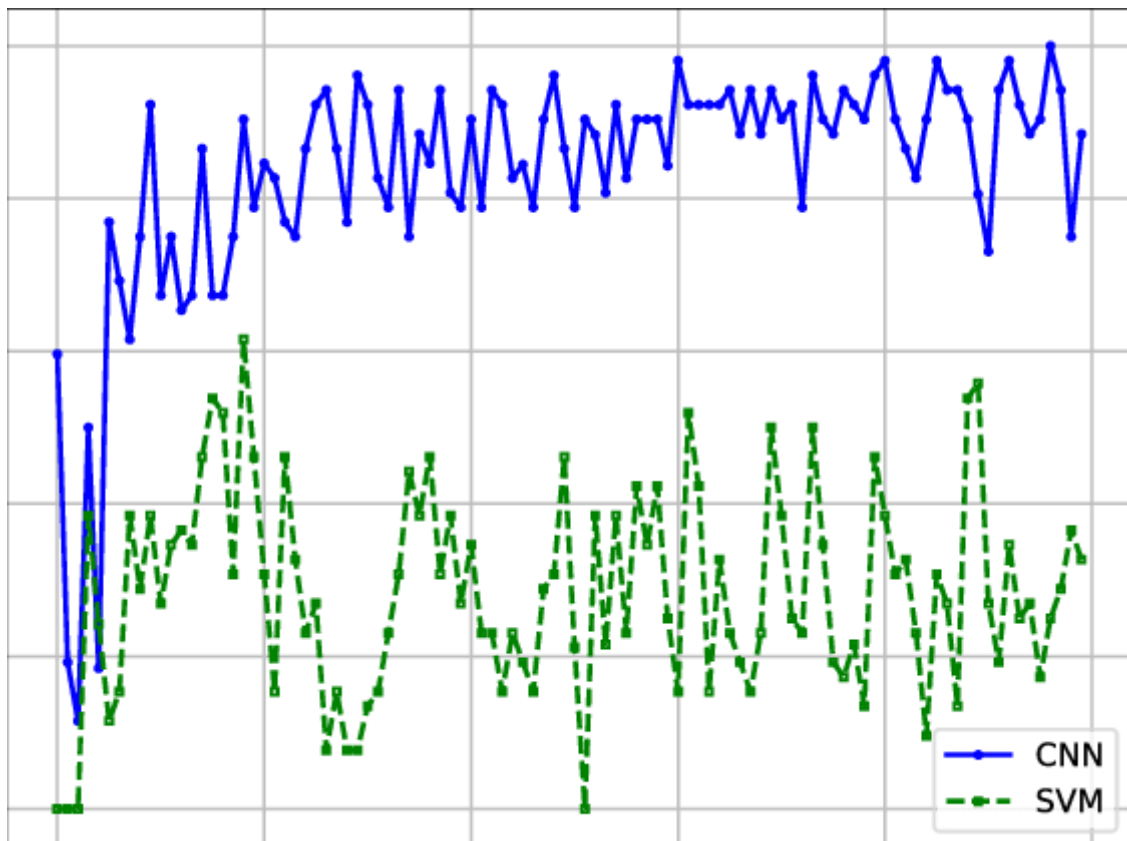
International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

Classification accuracy comparison of CNN vs SVM applied to network data



VII.CONCLUSION

Deep convolutional neural networks have been applied successfully to various applications. Besides providing breakthrough results with image recognition and classification tasks, they have also been applied to speech recognition, video analysis, natural language processing, etc., receiving positive results in terms of accuracy and cost they comprise of various components, with new components continuing to be added for the accomplishment of various tasks. Studies are constantly being done to find ways to improve their performance and widening their use to new tasks. Advances in GPU computing have helped CNNs to be used with a huge datasets and provide way faster computing which has eventually led to better generalization capacity by the networks. Depth has played a major factor in increasing efficiency of the network with various experiments continuously being done in order to achieve the most optimal performance with varying the number of hidden layers constantly. However, with advancements in the architectures of convolutional neural networks, more advances in technology for better software and hardware representations too are needed. Also, though methods such as deconvolution have been proposed to recognize the intuitive theory behind the working of CNNs, the solid theory behind how CNNs work and why they provide such results with visual tasks is still not formulated. The number of parameters to be reduced in order to obtain higher accuracy is constantly being experimented on, with different regularization techniques being introduced. The advent of fully convolutional neural networks, also has reduced the number of parameters used by the network, by completely eliminating the fully connected layers.

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 6, June 2018

REFERENCES

- 1.D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex", The Journal of physiology,1968,Vol.195,pp. 215-43
- 2.Yann Lecun,Leon Bottou,YoshuaBengio,Patrick Haffner,"Gradient based learning applied to document recognition", Proceedings of the IEEE,1998,Vol. 86, pp. 2278 - 2324
- 2.Alex Krizhevsky, IlyaSutskever, and Geoffrey Hinton,"Image Net Classification with Deep Convolutional Networks", Neural Information Processing Systems,2012, Vol 1,pp. 1097-1105
- 3.MatthewZeiler and Rob Fergus, "Visualizing and Understanding Convolutional Neural Network"s,European Conference on Computer Vision,2013,pp 818-833
- 4.Gu, Jiuxiang& Wang, Zhenhua&Kuen, Jason & Ma, Lianyang&Shahroudy, Amir &Shuai, Bing & Liu, Ting & Wang, Xingxing& Wang, Gang,"Recent Advances in Convolutional Neural Networks",Pattern Recognition,2015
- 5.M. D. Zeiler, R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks",International Conference on Learning Representations, 2013
- 6.Dumitru Erhan, Christian Szegedy, Alexander Toshev, DragomirAnguelov,"Scalable Object Detection using Deep Neural Networks", Conference on Computer Vision and Pattern Recognition,2013,pp. 2155-2162
- 7.ChristianSzegedy Alexander ToshevDumitruErhan,"Deep Neural Networks for Object Detection",Neural Information Processing Systems.,2013,Vol 2,pp.2553-2561
- 8.P. Jonathon Phillips ,"A Cross Benchmark Assessment of A Deep Convolutional Neural Network for Face Recognition",IEEE International Conference on Automatic Face & Gesture Recognition,2017
- 9.Hurieh Khalajzadeh, Mohammad Mansouri and Mohammad Teshnehlab,"Face Recognition using Convolutional Neural Network and Simple Logistic Classifier", Advances in Intelligent Systems and Computing,2017,Vol. 223,pp. 197-207
- 10.TaoWang,David J. Wu,AdamCoates,Andrew Y. Ng,"End-to-End Text Recognition with Convolutional Neural Networks",International Conference onPattern Recognition,2012,pp. 410-414
- 11.AlexisConneau,HolgerSchwenk,Yann Le Cun,"Very Deep Convolutional Networks for Text Classification,European Chapter of the Association for Computational Linguistics",2016,Vol 1,pp. 1107-1116
- 12.Kaiming He, Xiangyu Zhang, Shaoqing Ren, JianSunDeep,"Residual Learning for Image Recognition",Conference on Computer Vision and Pattern Recognition,2015,pp. 770-778
- 13.Ivan Laptev and Patrick Perez,"Retrieving actions in movies",International Conference on Computer Vision,2017,pp. 1-8
- 14.KarenSimonyan Andrew Zisserman,"Two-Stream Convolutional Networks for Action Recognition in Videos",Neural Information Processing Systems,2014,Vol 1,pp. 568-576
- 15.Andrej Karpathy et al,"Large-scale Video Classification with Convolutional Neural Networks",Conference on Computer Vision and Pattern Recognition,2014,pp. 1725-1732
- 16.Deepak S. Shete1, Dr. M.S. Chavan (2012) "Content Based Image Retrieval: Review", International Journal of Emerging Technology and Advanced Engineering ISSN, Volume 2, pp2250-2459
17. B. Darsana and G. Jagajothi (2014) "DICOM Image Retrieval Based on Neural Network Classification" International Journal of Computer Science and Telecommunications Volume 5, Issue 3 ISSN 2047-3338 pp.21-26.