PRAGMATIC ANALYSIS OF OBJECT CLASSIFICATION USING CNN

Chirag Sharma
Vellore Institute of Technology, Chennai
Tamil Nadu, India
0807chirag@gmail.com

Prof. Bharathi Raja S
Vellore Institute of Technology, Chennai
Tamil Nadu, India
bharathiraja.s@vit.ac.in

Abstract— :- Deep learning depends on learning level of portrayals, comparing to a request of highlights, factors, idea where tremendous volume of thoughts are described from lower level ones and the other way around. A subtype of a neural network called a convolutional neural system (CNN) is appropriate for picture related errands. The system is prepared to search for various features,like edges, corners and shading contrasts, over the picture and to consolidate these into more composite shapes. Convolutional neural network(CNN) is an extraordinary kind of forward counterfeit neural system dependent on visual cortex. We also show the funtionality of each and every layer of CNN. We had also discussed about the classifiers for image classification

Keywords:- Deep learning, neural network, classifiers, Convolutional neural network(CNN).

Introduction

Picture arrangement is a procedure of information extraction from a multiband raster picture. It is the procedure in which the picture can be named having a place with which classification and distinguishing it that whether it is a living thing or a non-living thing, evaluating the weight, measurements, surface region and so forth. There is essentially two kinds of picture characterization i.e. supervised and unsupervised.

"Supervised classification uses the spectral signature present in training set to classify the image".

When the group of pixel bind up together into cluster based on their properties or features then it is called Unsupervised classification. Analysts use image clustering algorithm like ISODATA to create cluster.

Deep learning works on Neural networks or we can say that it consist of neural networks. Neural networks are modifies under biological neural networks that allow computers to respond and work like a human. There are fundamentally three layers in neural system that is:- INPUT LAYER, HIDDEN LAYER and OUTPUT LAYER. Yann Lecun is the pioneer of Convolution Neural Network (CNN). Director of facebook artificial intelligence research group built the first Convolutional Neural Network (CNN) called LeNet in 1988 which is used for character recognition task like reading zip code, digits etc.

Convolution Neural System (CNN) is for the most part used to look at visual pictures by preparing information with matrix like topology Convolutional neural framework is wonderful sort of forward artificial neural framework in which the accessibility between neurons is roused by the visual cortex. Visual cortex is only a little part of our cerebrum(brain). It is a little segment which is touchy to particular field.

Generally there are five layers in CNN that we will study in this paper.

We will study the functioning of each and every layer of how it works in deep and its mathematical work, we will see some classifiers also that will help us in classifying the the given image.

Objective

In this project we will do the literature survey on object detection and classifictaion. In this semester our objevtive is to understand and learn the functioning of CNN and identification and classification of objects using CNN.

Object recognition is a computer innovation identified with computer vision and image handling that manages distinguishing occurrences of semantic objects of a specific classes, for example, creature, plant, human and so on Object detection is to identify the particular item from such huge numbers of other distinctive articles. Object detection is not only to say what the image is but it is also to recognize that where it is in the image.

Classification is the categorization of object based on previously define classes or type. Classification is identifying the object weather it is a car, pencil, marker, etc.

Motivation

We are living in a digital world where almost everything is based on digitalization and artificial intelligence, like ATM machine, robots, fully automatic washing machines, automatic cars etc. these day image recognition is also very popular and we are observing it in so many places where the machine after seeing the object easily identify the image that what that image is and to whom it belongs to, no matter it is a living thing or a non-living thing or any human or an individual.

Now a days everyone is using facebook, it is the simplest example where you can see this feature when you want to tag any of your friend in your group photo, it automatically identify the face and compare and match that face in your friend list and ask us that do you want to tag him/her.

Many research had being published in which they describe various method for object identification. In this paper we will discuss about what is Convolutional Neural Network (CNN), its functioning, different layers for image identification. Our main focus is to know the actual functioning that how the image is identified.

Working principle of Convolutional Neural Network (CNN)

In CNN every image is represented in the form of pixel values and it compare images piece by piece. It is generally use to analyze visual image by processing data with grid like topology. Convolutional Neural Network have different layers and each layer is important for getting the desired output. CNN have the following layers:

•Input layer

•Convolution layer

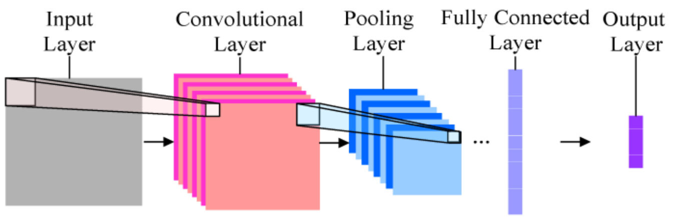
•ReLu layer

•Pooling layer

•Fully connected layer

•Output layer

If any of the layer from this fails to perform their task then the process will never be executed.



In the convolutional layer we first line up the component and the image and after that multiply the pixel values of the image with that of the comparing values of the filter and afterward adding then up and dividing them with the absolute number of pixels. ReLu layer represents Rectified layer unit.

The work of this layer is to remove all the negative value from the filter image and then change it with zero. It activates the node if the input is above the certain quality and it has the linear relationship with the dependent variable when the input is above threshold.

The work of pooling layer is to shrink the image which it got from from the Relu layer.

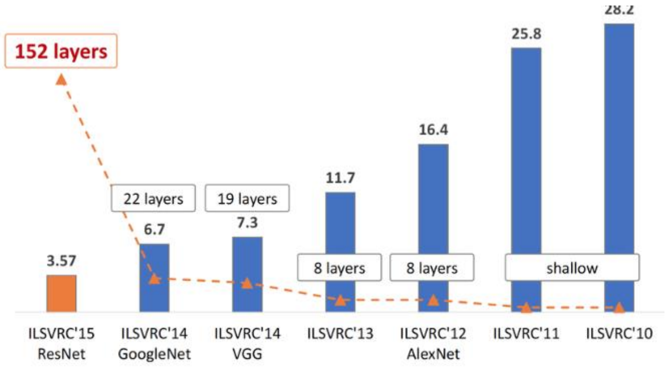
The actual classification is done in fully connected layer. We take the shrinked image and up that in a single list. And then we compare that image with our previously stored list and judge the image.

The last is the output layer which give the output of the classified image.

Related ImageNet

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a completion in which researchers come with their research and evaluate their algorithm on the given topic or data set and tries to win by showing their algorithm with higher accuracy on several visual recognition task. It was started in 2010 and held once in every year. Name of some winner teams are: AlexNet (2012), ZFNet (2013), VGG Net, GoogleNet, ResNet(2015).

The research is going on continuously and new ideas are coded which helps in increasing the image recognition accuracy. In every new and better research every year the number of layers are increasing to overcome the existing problem in the previous research and the error rate is falling every year.



Shallow Net:- shallow net come into existence in 2010 and 2011. A model was created to identify the pictures and images based by following some step. The layers are very less in shallow net and the error rate is very high of somewhat around 28.2% in 2010 then they modify the same model and the error rate is then decreases to 25.8% but still that was not a small error rate. The image identification quality is very low.

***AlexNet*:-** AlexNet is the name of a convolutional neural network(CNN), designed by Alex Krizhevsk, and publish with Ilya Sutskever and Geoffrey Hinton . AlexNet has had a huge affect on the field of machine learning, specificaly in the application of deep learning to machine. As of 2018 it has been cited over 25,000 times AlexNet competed in the ImageNet Large Scale Visual Recognition Challenge(ILSVRC) in 2012. The framework achieved a fundamental 5 blame of 15.3%, in overabundance of 10.8 rate centers lower than that of the contender. The first papers essential outcome was that the profundity of the model was fundamental for its superior, which was computationaly expensiive, yet made possible because of the usage of GPUs amid trainingg.AlexNet contained eight layers; the first five were convolutional layers, and the last 3 were completely associated layers.. It utilized the non-saturating ReLU initiation work, which indicated enhanced preparing execution over tanh and sigmoid. AlexNet was initially composed with CUDA to keep running with GPU bolster.

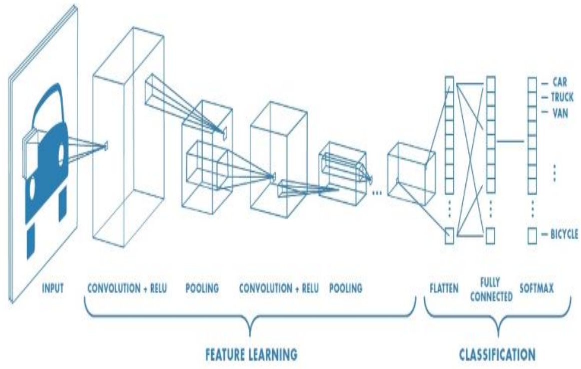
AlexNet was substantially bigger than past CNNs utilized for PC vision errands (e.g. Yann LeCuns LeNet paper in 1998). It has 60 million parameters and 650,000 neurons and took five to six days to prepare on two GTX 580 3GB GPUs. Today there are significantly more mind boggling CNNs that can keep running on quicker GPUs effectively even on huge datasets. Be that as it may, in 2012, this was colossal! Numerous Convolutional Kernels (a.k.a channels) extricate fascinating in a picture. In a solitary convolutional layer, there are generally various parts of a related size. For instance, the principal Conv Layer of AlexNet contains 96 portions of size 11x11x3. Note the width and tallness of the bit are typically the equivalent and the profundity is the equivalent as the quantity of channels. The first two Convolutional layers are surveyed by the Overlaping Max Pooling layers that we define next. The 3rd, 4th and 5th convolutional layers are joined directly. The fifth convolutional layer is folowed by an Overlaping Max Poolling layer, the output of which goes into a series of two fuly connected layers. The 2nd fully connected layer feads into a softmax classifier with 1000 class labels

***GoogleNet*:-** The winner of the ILSVRC 2014 competition was GoogleNet(a.k.a. Origin V1) from Google. It accomplished a best 5 error rate of 6.67%! This was near human level performance which the coordinators of the test were currently compelled to judge. Notably, this was entirely difficult to do and required some human preparing with the end goal to beat GoogLeNets precision. Following a couple of long periods of preparing, the human master (Andrej Karpathy) could accomplish a best 5 error rate of 5.1 %(single model) and 3.6%(ensemble). The system utilized a CNN motivated by LeNet yet executed a novel component which is named an origin module. It utilized bunch standardization, picture distortions and RMSprop. This module depends on a few little convolutions with the end goal to definitely lessen the quantity of parameters. Their design comprised of a 22 layer profound CNN however decreased the quantity of parameters from 60 million (AlexNet) to 4 million. Google isnt just about content any longer. The hunt mammoth is making incredible walks in comprehension and ordering pictures. Googles GoogLeNet venture was one of the triumphant groups in the 2014 ImageNet expansive scale visual acknowledgment challenge (ILSVRC), a yearly rivalry to gauge upgrades in machine visual innovation The GoogLeNet expands on the possibility that a large portion of the enactments in a profound system are either unnecessary(value of zero) or excess in light of connections between them. Along these lines the most productive engineering of a profound system will have a scanty association between the initiations, which infers that every one of the 512 yield channels wont have an association with all the 512 information channels. There are systems to prune out such associations which would result in an inadequate weight/association. Be that as it may, parts for scanty framework augmentation are not advanced in BLAS or CuBlas(CUDA for GPU) bundles which render them to be considerably slower than their thick partners. So GoogLeNet conceived a module considered origin module that approximates a scanty CNN with an ordinary thick development (appeared in the figure). Since just few neurons are viable as made reference to before, the width/number of the convolutional channels of a specific part estimate is kept little.

***ResNet*:-** Finally, at the ILSVRC 2015, the purported Residual Neural Network (ResNet) by Kaiming He et al presented anovel engineering with &quot;skip associations&quot; and highlights overwhelming bunch standardization. Such skip associations are otherwise called gated units or gated intermittent units and have a solid similitude to later fruitful components connected in RNNs. On account of this strategy they could prepare a NN with 152 layers while as yet having lower multifaceted nature than VGGNet. It accomplishes a best 5 blunder rate of 3.57% which beats human-level execution on this dataset. Profound remaining systems surprised the profound learning world when Microsoft Research discharged Deep Residual Learning for Image Recognition. These systems prompted first place winning sections in every one of the five primary tracks of the ImageNet and COCO 2015 rivalries, which secured picture order, question recognition, and semantic division. The strength of ResNets has since been demonstrated by different visual acknowledgment assignments and by non-visual errands including discourse and dialect. Before ResNet, there had been a few different ways to bargain the vanishing angle issue, for example, includes an assistant misfortune in a center layer as additional supervision , however none appeared to truly handle the issue for the last time. The center thought of ResNet is presenting a purported &quot;personality alternate way association&quot; that avoids at least one layers ResNet acquires and greater ubiquity in the examination network, its design is getting considered intensely. In this segment, I will initially present a few new structures dependent on ResNet , at that point present a paper that gives a translation of regarding ResNet as a gathering of numerous littler systems.

CNN image classification

CNN image classification takes an input image, process it and classife y it under certain categories for example dog, cat, tiger, lion, ect. Computer see an input image in an array of pixels and it depends on the image resolution. According on the resolution, it will see **h** x **w** x **d** where h= height, w = width and d = dimension
example – an image of 6 x 6 x 3 array of matrix of RGB(3 refer to RGB values) and an image of 4x4x1 array of matrix of gray scale image.



Convolutional layer:

- Convolutional is the first layer to extract features from an input image.
- Convolutional preserves the connection between the pixels by learning input highlights utilizing little squares of input information.
- It is a mathematical operation that takes two inputs such as image matrix and a filter kernel.

- An image matrix (volume) of dimension **(h x w x d)**
- A filter **(f_h x f_w x d)**
- Outputs a volume dimension **($h - f_h + 1$) x ($w - f_w + 1$) x 1**

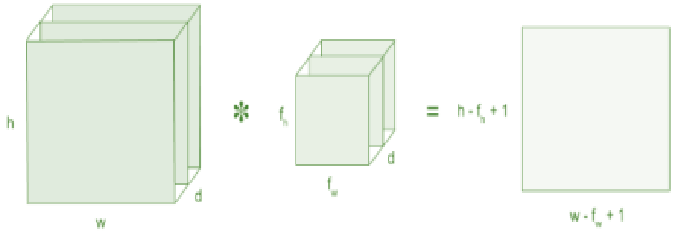


figure: image matrix multiply filter matrix

- consider a 5 x 5 whose image pixel values are 0,1 and filter matrix 3 x 3 as shown below.

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

*

1	0	1
0	1	0
1	0	1

5 x 5 – Image Matrix

3 x 3 – Filter Matrix

Then 5 x 5 image matrix multiply with 3 x 3 filter matrix which is called "feature Map" as output shown in below.

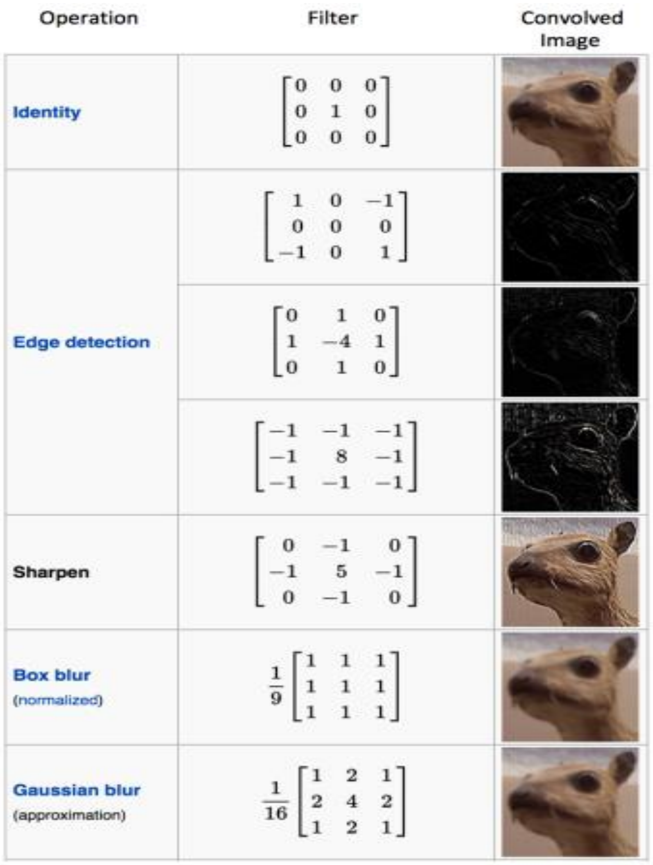
$1_{\times 1}$	$1_{\times 0}$	$1_{\times 1}$	0	0
$0_{\times 0}$	$1_{\times 1}$	$1_{\times 0}$	1	0
$0_{\times 1}$	$0_{\times 0}$	$1_{\times 1}$	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature

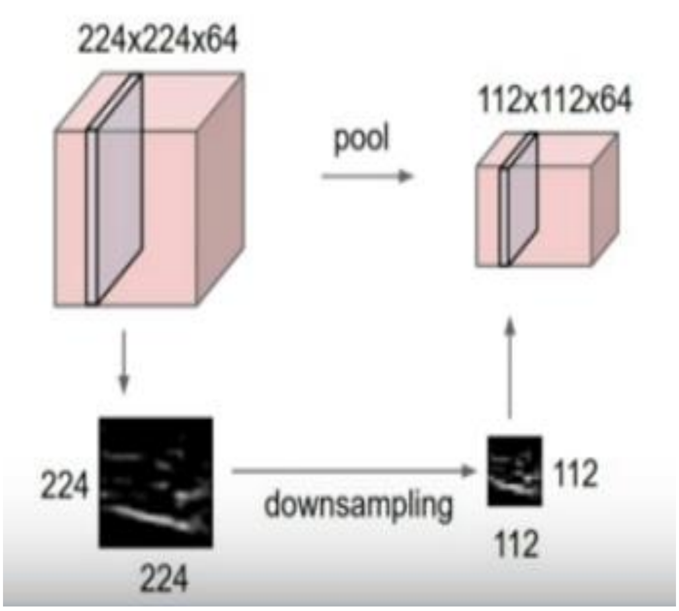
Convolutional of a picture with various channel can perform activity, for example, edge detection, obscure and sharpen by applying filters. The underneath model shows different convolutional picture in the wake of applying different sorts of filters.



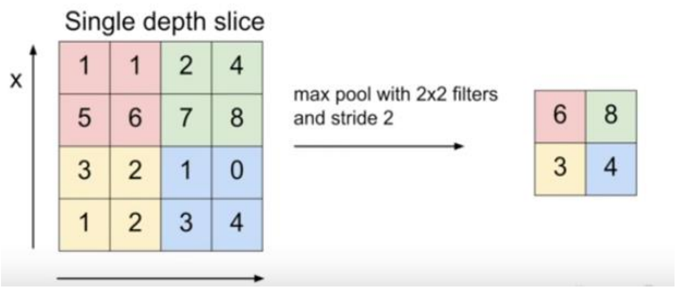
Operation	Filter	Convolved Image
Identity	$\begin{bmatrix}0&0&0\\0&1&0\\0&0&0\end{bmatrix}$	
Edge detection	$\begin{bmatrix}1&0&-1\\0&0&0\\-1&0&1\end{bmatrix}$	
	$\begin{bmatrix}0&1&0\\1&-4&1\\0&1&0\end{bmatrix}$	
	$\begin{bmatrix}-1&-1&-1\\-1&8&-1\\-1&-1&-1\end{bmatrix}$	
Sharpen	$\begin{bmatrix}0&-1&0\\-1&5&-1\\0&-1&0\end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix}1&2&1\\2&4&2\\1&2&1\end{bmatrix}$	

Pooling layer :

It makes the introduction littler and progressively sensible for preparing and works over every activation map inpendently.



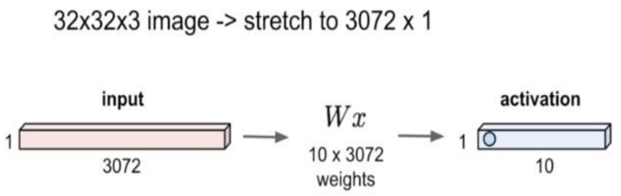
Max Pooling :



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Fully Connected Layer :



(1 x 3072) x (3072 x 10) = (1 x 10)

CLASSIFIER

Parametric approach : linear classifier

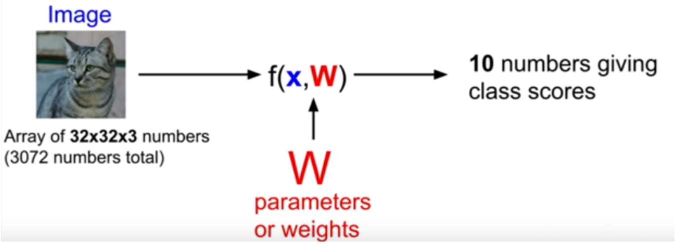
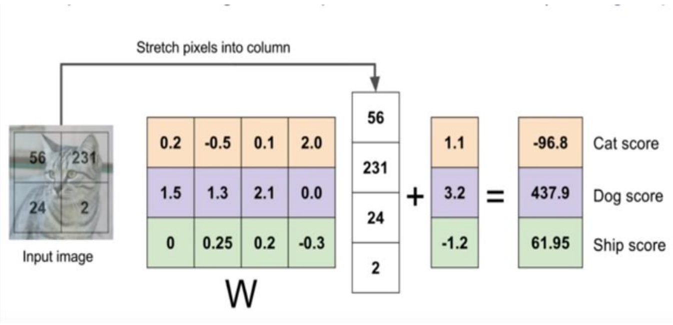


Image
Array of 32x32x3 numbers (3072 numbers total)
f(x,W)
10 numbers giving class scores
W
parameters or weights

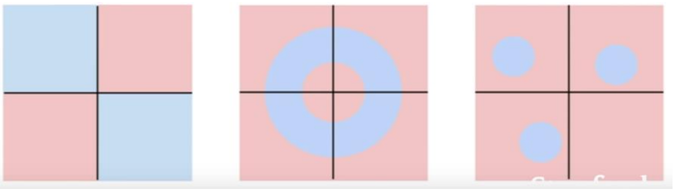
In k-nearest neighbor we do not have parameters. We have to keep around on the whole training data and use that at the test time. But now in parametric approach we goinmg to summarised the knowledge of the training data and stick on the knowledge into the parameters W.
Now in the test we do not need the actual training data, we only need the parameter W.
f(x, W) = Wx + b

Here b is the biased data, if the image is not balanced and it is different to decide that it is a cat or a dog then this biased is used. Biased element corresponding to the cat is higher than the other one.



Columnar matrix of 3X4
3 rows is the category i.e cat dog and ship
4 columns is the pixels .
Here dog value is the highest.

Three cases where linear classifier fails is :-



Multi-class SVM loss: **(linear classifier)**

It "needs" the correct class for each image to a have a score higher than the incorrect classes.The SVM "needs" a certain outcome in the sense that the outcome would yield a lower loss. Given an example **(x_i , y_i)** where **x_i** is the image and **y_i** is the (integer) label and using the short hand for the scores vector:
S=f(x_i , w)
The sum loss has the form :

$$L_i = \sum_{j \neq y_i} \begin{cases} 0 & \text{if } s_{y_i} \geq s_j + 1 \\ s_j - s_{y_i} + 1 & \text{otherwise} \end{cases}$$
$$= \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

SVM loss function is calculated by :-

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

Here Δ is some fixed margin.

In the the minimum loss is 0 and the maximum loss is infinite. Let us take one example to find out the loss.
Suppose that we have three classes that receive the scores s=[13,−7,11] and that the first class is the true class . Also assume that Δ is 10.

L_i = max(0, -7-13+10) + max(0,11-13+10)
= -10+8
= 0

Softmax Classifier (Multinomial Logistic Regression)

In multiclass SVM loss we did'nt have the interpretation of the scores. We do some classification and get some number. We don't care about the interpretation of those number or scores. We said that the score of correct class is greater than the incorrect classes and beyond that we don't say that what those score means. But in multinomial logistic regression we we actually will indulge in those score with some additional meaning and in particular we are going to use that score to complete a probability distribution over a class so we this softmax function where we take all our score, we expontiate them so that now they become positive then we normalise them by the sum of the exponents, so after we send the score to the softmax function to end up with probability distribution.

Scores = unnormalized log prob. of the classes

$$\boxed{P(Y = k|X = x_i) = \boxed{\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}}} \text{ where } \boxed{s = f(x, W)}$$

Softmax function

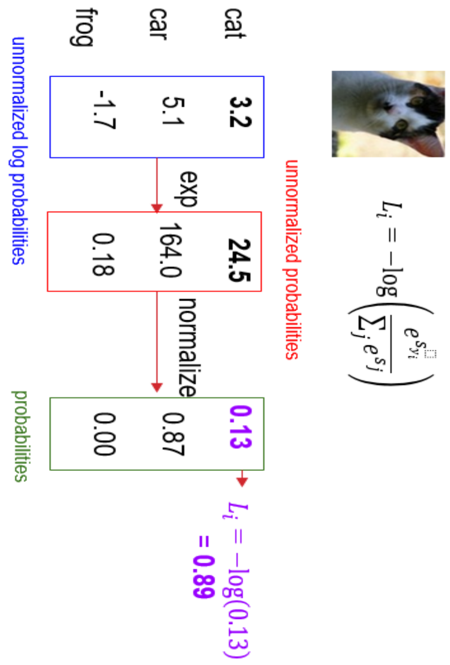
The interpretaion is that, if we know that the interpretation is a cat then the target probability distribution would plug all the probability mass on cat. So we have probability of cass =1 and zero for all the other classes, so what we want to do is to encourage our computive probability distribution coming out of the softmax function to match this target probabilty distribution that have all the mass of correct class, the way you want to do this is in the equation.

$$\boxed{L_i = -\log P(Y = y_i|X = x_i)}$$

In the end the probability of true class is high or close to 1.

In summary:

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$


$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$
unnormalized probabilities
cat
car
frog
3.2
5.1
-1.7
exp
24.5
164.0
0.18
normalize
0.13
0.87
0.00
$L_i = -\log(0.13)$
= 0.89
unnormalized log probabilities
probabilities

Summary

AUTHOR	DESCRIPTION	ADVANTAGES
1. AlexKrizhevsky, IlyaSutskever, GeoffreyE.Hinton	In this paper they prepared a vast, profound convolutional neural system to group the 1.2 million high-goals pictures in the ImageNet LSVRC-2010 challenge into the 1000 distinct classes. The neural system, which has 60 million parameters and 650,000 neurons, comprises of five convolutional layers, some of which are trailed by max-pooling layers, and three completely associated layers with a final 1000-way softmax. To make preparing quicker, they utilized non-immersing neurons and an extremely efficient GPU execution of the convolution task.	Confirm the importance of depth in visual representation.
2. Christian Szegedy	In this paper they propose a deep convolutional neural network architecture, which was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). The main hallmark of that architecture is the improved utilization of the computing resources inside the network. To optimize quality, the architectural decisions were based on the intuition of multi-scale processing	this method is a significant quality gain at a modest increaseofcomputationalrequirementscomparedtoshallowerandlesswidenetworks.
3. Karen Simonyan∗ & Andrew Zisserman	In this paper they investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Their main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters. They also show that their representations generalize well to other datasets, where they achieve state-of-the-art results. They have made their two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.	They show that their models discover well to a large range of tasks and datasets, matching or outperforming more compound recognition.
4. Christian Szegedy	In this paper they propose a deep convolutional neural network architecture, which was responsible for setting the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14).	This method accure a significant quality at a humble increase of computational requirements compared to shallower and less wide networks.

5. Md Zahangir Alom1, Tarek M. Taha1, Chris Yakopcic1, Stefan Westberg1, Paheding Sidike2, Mst Shamima Nasrin1, Brian C Van Essen3, Abdul A S. Awwal3, and Vijayan K. Asari1	This report continues a hasty survey on the advances that have happen in the field of deep learning (DL), starting with the Deep Neural Network, This survey then covers the Convolutional Neural Network. They also included recent developments such as advanced variant deep learning techniques based on these deep learning approaches. This work considers most of the papers distributed after 2012 from when the history of deep learning started.	They have provide an in-depth analysis of deep learning and its applications.
6. Suraj Srinivas, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas S S Kruthiventi and R. Venkatesh Babu	In this paper they clearly inspect one form of deep networks largely used in computer vision – Convolutional Neural Networks. They started with AlexNet as their base convolutional neural network.	They empower tricky hand tuned algorithms being replaced by single monolithics algo. trained in an end to end manner.

Conclusion

In this paper we examined different techniques for the grouping of pictures, for example, neural systems, convolutional neural network, we understood and learn the functioning of convolutional neural network (CNN) and neural networks and identification and classification of objects using CNN. We covered the standard CNN architecture, CNN classification for image classification and the functioning of classifiers.

References

1.https://ieeexplore.ieee.org/document/5206848.

2.VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION, Karen Simonyan∗ & Andrew Zisserman+ Visual Geometry Group, (2015).

3.Dynamic Routing Between Capsules, SaraSabour NicholasFrosst GeoffreyE.Hinton, (2017)

4.Doctoral research scientist on deep Learning, computer vision for remote sensing and hyper spectral imaging (e-mail: pehedings@slu.edu).

5.The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, Md Zahangir Alom1, Tarek M. Taha1, Chris Yakopcic1, Stefan Westberg1, Paheding Sidike2, Mst Shamima Nasrin1, Brian C Van Essen3, Abdul A S. Awwal3, and Vijayan K. Asari1, (2015)

6.Convolutional neural networks for document image classification, Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann, (2014).

7."TimeNet: Pre-trained deep recurrent neural network for time series classification." Malhotra, Pankaj, et al. (2017).

8.Image Net Classification with Deep Convolutional Neural Networks, Alex Krizhevsky, Ilya Sutskever, Geoffrey E.Hinton. (2012).

9.Going deeper with convolutions, Christian Szegedy, Yang qingJi, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, (2015)

10.https://in.mathworks.com/help/deeplearning/ug/introduction-to-convolutional-neural-networks.html

11.https://becominghuman.ai/building-an-image-classifier-using-deep-learning-in-python-totally-from-a-beginners-perspective-be8dbaf22dd8