Image Classification Using Convolution Neural Network and Comparative Analysis

Vineet Kumar
Assoc. Professor, Department of Computer Science and Engineering
Noida Institute of Engineering and Technology
19 KP 2, Greater Noida, Uttar Pradesh India.

Amit Vikram Tripathi
Scholar, Department of Computer Science and Engineering
Noida Institute of Engineering and Technology
19 KP2, Greater Noida, Uttar Pradesh, India.

Dr. A. K. Sinha
Director, UST Software India Pvt Ltd
New Delhi, India

.

Deepali Gupta
Scholar, Department of Computer Science and Engineering
Noida Institute of Engineering and Technology
19 KP2, Greater Noida, Uttar Pradesh, India.

Abstract
Object classification solve many problems related to image processing and computer vision. In past researcher has proposed many diversified algorithms to classify the object present in an image. Convolution Neural Network (CNN) plays an important role in classifying the image and produce a high rate of accuracy. CNN uses the multi-layer neural network which is designed to recognize the visual patterns present in an image. LeNet, AlexNet, GoogLeNet, ResNet architecture of CNN were presented by researchers. This paper analyses the performance of existing architecture and their performances are compared with the proposed architecture of CNN. In this regard, CIFAR-10 and CIFAR-100 are datasets which have been used at the different batch size to check the limitation and capabilities of CNN. Proposed model presents good result on elective classes.

Keywords: convolutional neural network, image classification, gradient descent, AlexNet, GoogleNet, CIFAR-10, CIFAR-100.

Introduction
Object classification and detections are the most prominent research area in the field of computer vision and pattern recognition problems. With an abundance of images on the internet, the user is now focused on image searches. These search images need to be classified efficiently. Here comes a major role in image classification. Image classification can be defined as a task to categories the image into image classes. There are numerous algorithms which are being used in the image classification such as, bag-of-words, Histogram intersection kernel (HIK), support vector machines (SVM) [1], face landmark estimation (for face recognition), K-nearest neighbours (KNN), logistic regression etc. In the starting years of image classification, methods used for extracting patterns were SIFT (Scale-invariant feature transform) and HOG (histogram of oriented gradients). SIFT was patented in Canada by the University of British Columbia and published by David Lowe in 1999. In SIFT key-points of reference images were extracted and stored in a database then the features of a new image were compared to the database on the basis of the Euclidean distance of their feature vectors. Nowadays convolutional neural network is used for image classification and it is used as a base for other computer vision process such as image content analysis, object detection, and identification, image inpainting, face recognition etc. In 2017[2] Alex Krizhevsky suggests that CNN's are widely recognized and used to extract the essential feature for the classification. The main advantage of using CNN over SIFT and HOG is that the feature extraction and generalization of patterns are more similar to the process of the human visual system. For performing CNN we need to perform training and testing of the network that is to train data and to test data, so the data is divided into two sets train-set and test-set. For training the data we take train-set and based on the unique characteristics of images we categorize them into respective classes. Then the test-set is checked to find the accuracy of the model. To increase the accuracy of the model we fix all the parameters of the network and increase the number of convolutional layers. In this paper, we have proposed our own model by varying the numbers of convolutional layers and increasing the dense layer which further increases the number of neurons at the end of the network. We are presenting a comparative study of earlier proposed models, AlexNet

GoogLeNet, and our model by training and testing them on the standard dataset of CIFAR 10, CIFAR 100. We have recorded the variation in the image classification and presented into the result section of this paper. We have started the paper by Literature Survey which contains the information about all the prior works, following with our methodology for comparing the existing networks with our proposed network, including the description about models and datasets. The last section contains the conclusion in which the futuristic scope is discussed.

Literature Survey

The image classification problems are affected due to the presence of noise, occlusion and poor quality. It is very difficult to categorize the object present in the image and it became very difficult if multiple objects are present in the image. Most classification techniques are based on the feature extracted from the image. Image classifications are categorized as supervised and unsupervised. In supervised classification, the Learning process is designed to form a mapping from one set of data (feature) to another set of data (information classes) in the presence of teacher while unsupervised classification is free from human intervention. Many techniques like Artificial Neural Networks (ANN), Minimum Distance from Mean (MDM), Support Vector Machines (SVM), Maximum Likelihood (ML), are in supervised classification for assigning pixels to informational classes. The support vector machine (SVM) [4] is applicable to both regression & pattern recognition and it is a new universal machine learning. Due to the presence of a teacher in supervised classification errors are detected during the training and is corrected at that time also but this process is highly time-consuming with high expenses. In the case of unsupervised classification, is free from human intervention because no prior information is needed. In the absence of reliable training data, it is possible to understand the structure of the data using statistical methods such as clustering algorithms. Popular clustering algorithms are k-means and ISODATA. These techniques are faster, error-free and there is no need for detailed prior knowledge. Maximally-separable clusters bring the major drawback in this technique[3]. Further soft classification is based on the pixel these are maximum likelihood(ML) classifier, subpixel classifier, K-NN classifier, fuzzy-set classifiers. Nowadays for each and everything done like speech recognition, health monitoring, object detection, marketing, behavioural analysis and many more all of these uses machine learning, artificial neural network, and deep learning. So now ANN is playing a very important role in development. Although using ANN has many advantages like fault tolerance, once it is trained it can work with incomplete information and many more but the main disadvantage ANN is facing is that it takes a large amount of time in training the sets which contains a large number of features and millions of patterns. So it is very important to extract important features from every layer so that they can be processed further by the next layer which makes classification easy. Vectorization is the process of transforming the original data structure into a vector form. Vectorization played an important role in scaling up the neural network model. This introduced deep learning or hierarchical learning which is a part of machine learning working on various layers where the output of every layer works as input to the other layer and like this it helps in both supervised and unsupervised learning. Several types of deep learning network have been proposed which helps in extracting useful information from digital images such as CNN(Convolutional Neural Net) [4], SAE (Stacked Auto Encoder)[6] and RBMs (Restricted Boltzmann Machines)[5] networks. Although Convolutional neural network is widely used for speech recognition, image recognition, and image inpainting, object detection [7]. CNN uses various layers which are the input layer, output layer and hidden layers such as convolution layer, pooling, fully connected layer, and normalization layers. By the mid-1980s the first realization of Convolutional Neural Network (CNN) [8] was seen, the receptive field was the basis of CNN. This hierarchical structured artificial neural network was first neuron based connectivity. The algorithm was proposed to deal with a problem with shifts in distortions and positions in shape of patterns in images, so the algorithm proposed was a multi-layered network made up of neurons like nodes. Then researchers moved forward and started working on the ways in which Artificial Neural Network can be used in daily life problems[9]. Further backpropagation algorithm was proposed which was used for determining the error gradient [10]. David Rumelhart, Geoffrey Hinton, and Ronald Williams published this algorithm in 1985. This algorithm turned out to be very effective and Artificial Neural Network has expanded since then. Earlier in 1990 backpropagation algorithm was used the first time for training CNN model[11] when the handwritten digital identification is studied by them and MNSIT [12] where handwritten digital data was studied with relative time constraints. The CNN was simplified by Simard et al.[13] in 2003 by fusing convolution and pooling operations which improved the network and document analysis. In 2006, Chellapilla et al.[14] the same architecture of convolution layers was converted into a matrix-matrix product which resulted in faster recognition. However, this computing result is much smaller than the modern deep CNN's due to low computing power available at that time. A Krizhevsky, G Hinton –2009[15] trained a multilayer generative model of natural images and used the dataset CIFAR-10.

Methodology of Evaluation

The principal objective of this work is to study the performance of the different network as well as the proposed network for images. There are several pre-trained CNN's. Each of these networks has a different number of layers through which they are building up and there exist different approach in designing them. GoogLeNet has Inception modules that are used for the different size of convolution and concatenate the filters for further processing. In AlexNet instead of using concatenation of filters, it uses the output of antecedent layers as input to the succeeding layer. Caffe [16] and Tensorflow are used for the implementation of these networks independently. In this paper, we proposed a CNN architecture with modification in the existing model by replacing the end most two layers with dense layers and a SoftMax layer. Firstly, we have taken the size of prediction neurons as 1024 which is followed by 512 and then 256 and at

the last equal to the classes taken from the test dataset, consecutively to get output from the final layer. It gives a significant change in the learning rate of architecture.

Proposed Model

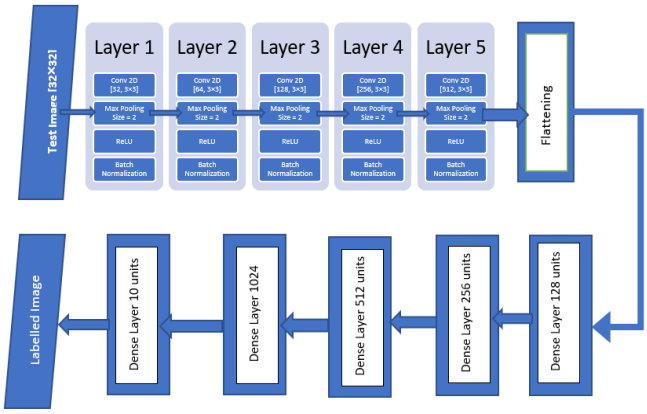

Figure 1. Block diagram of Proposed Neural Network Architecture

Training the models is a highly time-consuming process. It increases the cost and use the high amount of energy and hence not feasible for the researchers. Sometimes the system configuration does not allow to train the deep neural networks to predict less accurate results. Transfer learning is an efficient way of predicting the results on our own datasets. All the datasets are categorized into training and testing datasets and these superclass images that are used for training purpose are resized to [244,244] pixels for AlexNet and [227,227] pixels in GoogLeNet and in our proposed model these images are resized to [32,32] pixels.

The proposed model has 5 layers initially which have following components with them. All of them are described in the details below:

Conv2D

This layer creates a matrix which is convolved on the input image and hence produces the feature matrix for the input image.

Max Pooling

Max pooling is used on the output of the conv2D matrix and a stride of size 2 is scanned over it which drops the extra values from the matrix and hence reduces the size of it.

ReLU

It is an activation function which is used at each layer. The ReLU work as a node which is kept on the output end. It helps us to determine the value of output between 0 to infinity.

Batch Normalization

The process through which we normalize the input layer neurons by adjusting and scaling the activation nodes. Batch normalization works on the output of the previous layer and normalizes it by subtracting the batch mean and further dividing by the batch standard deviation.

Despite these neural layers, we have few more layers of neurons which are described below:

Dense Layer

It is the collection of output nodes. The output is calculated on this layer as, output = activation (dot (input, kernel) + bias), where activation, input, kernel are the matrices derived on the previous neural layers. Bias is a vector created by the layer which helps keep the balance between weights.

Softmax Layer

It is used to obtain the desired result for the multi-class classification. The probability sum for the Softmax will not exceed more than 1. Softmax layer is used because the classes in test dataset were mutually exclusive.

Considering the first layer of proposed model in which the [32,32] pixel size of image is passed and after convolving over this input matrix, a new matrix has been generated which is further pooled using a stride of size 2. Then the resultant matrix is normalized using the batch normalization which help to reduce the overfitting of model.

Test Dataset

For testing the models, 2 standard datasets CIFAR-10 and CIFAR-100 are used [17]. CIFAR-10 has a total of 60000, 32x32 colour images which belong to 10 different classes. They all are physically existing entities. While doing the experiment this data is imported and is divided into two subsets Training dataset and test dataset. The training set contains 50000 images while the test set contain 10000 images. Similarly, in CIFAR-100 dataset there are 100 classes of images and each class contains 600 images. Now, these classes have two sets of test data and train data which contains 500 and 100 images respectively. Again 10 classes are selected from CIFAR-100 dataset. It is clear that we have used the same data for training, testing, and validation. We have listed all those classes into the results.

Results

We have analysed AlexNet and GoogleNet neural networks by testing them on CIFAR-100, CIFAR-10. We have trained and tested the dataset separately on each model. Table 1 contains the performance of CIFAR 100 dataset on different networks and Table 2 contains the performance of CIFAR 10. The percentage notation denotes that one the set of 100 images certain number of images are predicted correctly. It was noticeable that the transfer learning performed better than the existing networks. Hardly some entities like “chair”, “train” and “wardrobe” were perfectly recognized by the network. It is indicating that the network is trained on a great parameter and after all the model has learned to distinguish between the objects on which it has been trained. We are not claiming it as a perfect model, there's a lot of improvement needs to be done. Few objects of size 32*32 were predicted wrong by the proposed model but we think that's fine because achieving 100% accuracy is not possible due to many limitations including hardware configuration of the machine.

Results of various CNN with the proposed model is given below.

TABLE 1. RESULTS FOR CIFAR10 ON DIFFERENT NETWORKS

CIFAR 10	AlexNet	GoogLeNet	Proposed Model
Airplane	41.80%	51.10%	75.80%
Automobile	21.80%	62.10%	55.90%
Bird	0.02%	56.70%	82.90%
Cat	0.03%	78.80%	51.90%
Deer	87.60%	49.50%	29.40%
Dog	23.00%	57.50%	72.10%
Frog	24.20%	90.20%	66.60%
Horse	34.70%	78.20%	74.70%
Ship	31.70%	95.50%	78.20%
Truck	95.90%	97.10%	64.60%

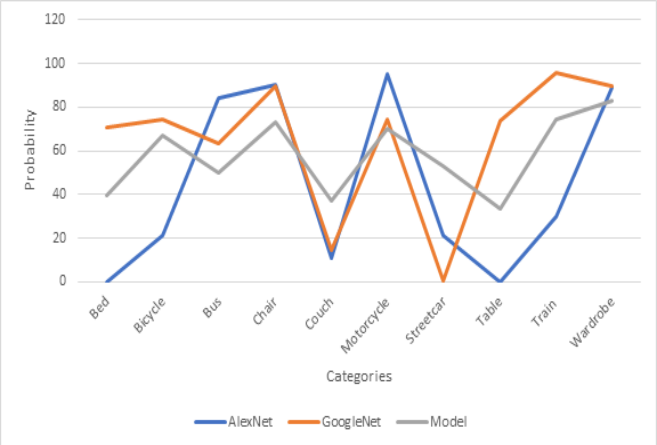

Figure 2. Probability v/s category graph for CIFAR10 datasets

TABLE 2. RESULTS FOR CIFAR100 ON DIFFERENT NETWORKS

CIFAR 100	AlexNet	GoogLeNet	Proposed Model
Bed	0.00%	70.80%	39.80%
Bicycle	21.00%	74.20%	67.10%
Bus	84.00%	63.20%	49.70%
Chair	90.00%	89.60%	73.20%
Couch	11.00%	14.60%	36.90%
Motorcycle	95.00%	74.60%	69.80%
Street Car	21.00%	0.84%	53.20%
Table	0.00%	73.60%	33.50%
Train	30.00%	95.60%	74.20%
Wardrobe	95.90%	89.40%	82.80%

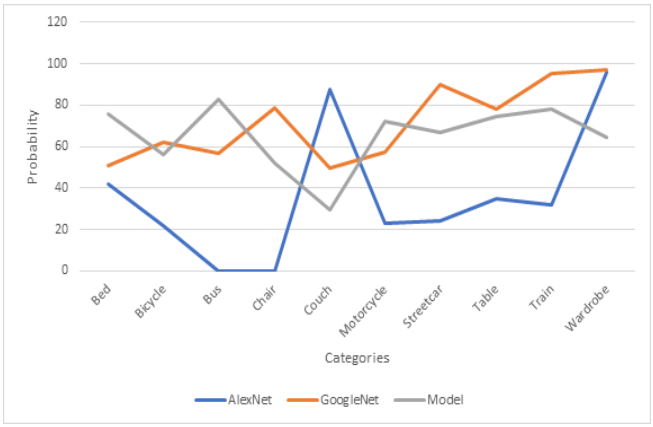

Figure 3. Probability v/s category graph for CIFAR100 datasets

Analysis and comparison of the existing network and our proposed model is the core of this paper. We have limited our study to only 10 classes of each dataset of CIFAR-10 and CIFAR-100. The principal objective of this paper evaluating the stability of each network. One of the major notable points is that multiplex picture create uncertainty for the network and hence result varies. The result shows that on increasing the dense layer some of the objects like "wardrobe", "airplane", "bird" is classified accurately by our proposed model. Existing network AlexNet which contains 8 layers in which first 5 are the convolutional layer and last 3 are the fully connected layer. AlexNet uses 60 million parameters. Similarly, GoogLeNet has 22 convolutional layers and it has 4 million parameters. In our proposed model we have attained a value of near about 1 million parameters which is quite a good number. One more thing which we have noticed that increasing the number of layers increases training rate hence it can be useful in reducing the training cost of datasets. Further, we can conclude that neural networks are the emerging technology and they are useful in dealing with daily life problem. It is easy to integrate these neural networks with various platforms. This model can be further modified to achieve the best results in the form of supervised learning. The weights generated during the training of this model can we used to deploy it on the web servers and can be used in creating the web-app for the classification. Image classification can be used on social media as well as on the e-commerce platforms where the user can shoot a picture from their phone and use it to search the similar item available on the platform. Although there's an existing system which functions, in the same manner, we hope our research work will help them to improve their standards high.

References

[1] Jianxin Wu, "Efficient Hik SVM Learning For Image Classification", IEEE Transactions On Image Processing, Vol. 21, No. 10, October 2012.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks", In Advances in Neural Information Processing Systems 25, pages 1106–1114, 2012

[3] Lizhen Lu, Liping Di, Senior and Yanmei Ye, "A decision-tree classifier for extracting transparent plastic-mulched landcover from Landsat-5 TM images", IEEE Journal Of Selected Topics In Applied Earth

Observations And Remote Sensing, Vol. 7, No. 11, November 2014.

[4] Zhu, G. and Blumberg, D. G. 2002. "Classification using ASTER data and SVM algorithms: the case study of Beer Sheva", Israel *Remote Sensing of Environment*, 80: 233–240.

[5] P. Vincent, H. Larochelle, Y. Bengio, "Extracting and composing robust features with denoising autoencoders", ICML. ACM, pp.1096-1103, 2008.

[6] Larochelle, Y.Bengio, "Classification using discriminative restricted Boltzmann machines". Proceedings of the 25th international conference on Machine learning. ACM, pp.536-543, 2008.

[7] S. Ren, K. He, R. Girshick, "Faster r-CNN: towards real-time object detection with region proposal networks". Advances in neural information processing systems, pp,91-99, 2015.

[8] K. Fukushima, S. Miyake. "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position". Pattern Recognition, 1982, 15 (6): 455-469.

[9] D. W. Ruck, S. K. Rogers, M. Kabrisky. "Feature selection using a multilayer perceptron" Journal of Neural Network Computing, 1990.

[10] D. E. Rumelhart, G. E. Hinton, R. J. Williams. "Learning representations by back-propagating errors". Nature, 1986, 323: 533-538.

[11] Y. LeCun, et al. "Handwork digit recognition with a back-propagation network", Advances in Neural Information Processing Systems. Colorado, USA: [s. N], 1990: 396-404.

[12] Y. LeCun, C. Cortes, "MNIST handwritten digit database [EB / OL]". Http: // yann. Lecun. Com / exdb / mnist, 2010.

[13] Simard, D., Steinkraus, P. Y. & Platt, J. C. "Best practices for convolutional neural network"s. In *Proc. Document Analysis and Recognition* 958–963, 2003.

[14] Kumar Chellapilla, Sidd Puri, Patrice Simard, "High performance convolutional neural networks for document processing". In International Workshop on Frontiers in Handwriting Recognition, 2006.

[15] A. Krizhevsky, "Learning multiple layers of features from tiny images". Tech report, 2009.

[16] Karpathy, A., Toderici, G., Shetty, Leung, T., Sukthankar, R., & Fei-Fei, L. "Large-scale video classification with convolutional neural networks." in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[17] X. Wang, M. Yang, S. Zhu, and Y. Lin. "Regionlets for generic object detection." In ICCV, 2013.