Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 132 (2018) 377–384

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

An Analysis Of Convolutional Neural Networks For Image Classification

Neha Sharma,Vibhor Jain, Anju Mishra

Amity University Uttar Pradesh, Noida, India

Abstract

This paper presents an empirical analysis of theperformance of popular convolutional neural networks (CNNs) for identifying objects in real time video feeds. The most popular convolution neural networks for object detection and object category classification from images are Alex Nets, GoogLeNet,and ResNet50. A variety of image data sets are available to test the performance of different types of CNN's. The commonly found benchmark datasets for evaluating the performance of a convolutional neural network are anImageNet dataset, and CIFAR10, CIFAR100,and MNIST image data sets. This study focuses on analyzing the performance of three popular networks: Alex Net, GoogLeNet,and ResNet50. We have taken three most popular data sets ImageNet, CIFAR10,and CIFAR100 for our study, since, testing the performance of a network on a single data set does not reveal its true capability and limitations. It must be noted that videos are not used as a training dataset, they are used as testing datasets. Our analysis shows that GoogLeNet and ResNet50 are able to recognize objects with better precision compared to Alex Net. Moreover, theperformance of trained CNN's vary substantially across different categories of objects and we,therefore, will discuss the possible reasons for this.

Keywords: Deep Learning; CNN; Object detection; Object classification; Neural network

1.Introduction

Nowadays internet is filled with anabundance of images and videos, which is encouraging thedevelopment of search applications and algorithms that can examine the semantic analysis [1] of image and videos for presenting the user with better search content and their summarization. There have been major breakthroughs in image labeling, object detection, scene classification [2] [3], areas reported by different researchers across the world. This leads to making it possible to formulate approaches concerning object detection and scene classification problems. Since artificial neural networks have shown a performance breakthrough in the area of object detection and scene classification, specially convolutional neural networks (CNN)[4] [5] [6], this work focuses on identifying the best network for this purpose. Feature extraction is a key step of such algorithms. Feature extraction from images involves extracting a minimal set of features containing ahigh amount of object or scene information from low-level image pixel values, therefore, capturing the difference among the object categories involved. Some of the traditional feature extraction techniques used on images are Scale-invariant feature transform (SIFT) [7], histogram of oriented gradients (HOG) [8], Local binary patterns (LBP) [10], Content-Based Image Retrieval (CBIR) [11], etc. Once features are extracted their classification is done based on objects present in an image. A few examples of classifiers are Support vector machine (SVM), Logistic Regression, Random Forest, decision trees etc.

Corresponding author: neha.sharma5852@gmail.com

10.1016/j.procs.2018.05.198

CNN has been presenting anoperative class of models for better understanding of contents present in an image, therefore resulting in better image recognition, segmentation, detection,and retrieval. CNN's are efficiently and effectively used in many pattern and image recognition applications, for example, gesture recognition [14], face recognition [12], object classification [13] and generating scene descriptions. Similarly, CNNs have achieved detection rates (CDRs) of 99.77% using the MNIST database of handwritten digits [23], 97.47% with the NORB dataset of 3D objects [24], and 97.6% on around 5600 images of more than 10 objects [25]. The successful integration of all the stated applications is due to advances and development in learning algorithms for deep network construction and moderately to the open source large labeled data set available for experimentation purpose, for example, ImageNet, CIFAR 10, 100, MNIST etc. [16] CNN has well known trained networks that uses these datasets available in open source networks and increases its efficacy of classification after getting trained over millions of images contained in the datasets of CIFAR-100 and Image-Nets. The datasets used are composed of millions of tiny images. Therefore, they can simplify well and accurate and hence successfully categorize the classes' out-of-sample examples. It is important to note that neural network classification and prediction accuracy and error rates are all most comparable to that of humans when such comparisons are made on a large data set such as Image-Net, CIFAR-10, 100 etc. This work aims at analyzing the capability of convolutional neural networks to categories the scene in videos on the basis of identified objects. A variety of image categories are included in CIFAR-100, CIFAR 10 and ImageNet datasets for training the CNN. The test datasets are videos of different categories and subjects. The contradiction branches out because of the feature extraction capabilities of different CNN. The primary contribution of our work is to present object detection methods using different types of trained neural networks where current up-to-date models show different performance rates for test images or videos when compared to trained images. After training these networks for different object classes presented as input in the form of images, and then testing for themore particular real-time video feed, we can better understand what is being learned and presented by these models. We therefore, can postulate that an image representation on the basis of objects detected in it would be significantly useful for high-level visual recognition tasks for scenes jumbled with numerous objects resulting in difficulty for the network to classify it. These networks also provide supplementary information about the extraction of low-level features. These networks are trained on datasets containing millions of tiny images [12]. We propose that the concept of object detection can be used as an attribute for scene representation. These networks used for our study are constructed using existing neural networks and each of these networks have different layers, therefore their performance varies considerably. Using complex real-world scenes the detection accuracy of the network can be checked. This paper is arranged as follows. We begin by presenting related prior works, following with the problem statement and our proposed methodology for comparing the networks chosen for the study, including descriptions of the models and data sets. We then present a comprehensive analysis of results obtained on different datasets. Finally, we conclude the paper and discuss about future work.

2.Related Work

The Convolutional Neural Networks (CNN) are used in a number of tasks which have agreat performance in different applications. Recognition of handwritten digits [17] was one of the first application where CNN architecture was successfully implemented. Since the creation of CNN, there has been continuous improvement in networks with the innovation of new layers and involvement of different computer vision techniques [18]. Convolutional Neural Networks are mostly used in the ImageNet Challenge with various combinations of datasets of sketches [19]. Few of the researchers have shown a comparison between thehuman subject and a trained network's detection abilities on image datasets. The comparison results showed that human being corresponds to a 73.1% accuracy rate on the dataset whereas the outcomes of a trained network show a 64% accuracy rate [21]. Similarly, when Convolutional Neural Networks was applied to the same dataset it yielded an accuracy of 74.9%, hence outperforming the accuracy rate of humans [21]. The used methods mostly make use of the strokes' order to attain a much better accuracy rate. There are studies going on that aim at understanding Deep Neural Network's behavior in diverse situations [20]. These studies present how small changes made to an image can severely change the results of grouping. In the work also, presents images that are fully unrecognized by human's beings but are classified with high accuracy rates by the trained networks [20].

There has been a lot of development in the area of feature detectors and descriptors and many Algorithms and techniques have been developed for object and scene classification. We generally enticement the similarity between the object detectors, texture filters,and filter banks. There is anabundance of work in the literature of object detection and scene classification [3]. Researchers mostly use the current up-to-date descriptors of Felzenszwalb and context classifiers of Hoeim [4]. The idea of developing various object detectors for basic interpretation of images is similar to the work done in multi-media community in which they use alarge number of "semantic concepts" for image and video annotations and semantic indexing [22]. In the literature that relates to our work, each semantic concept is trained by using either the image or frames of videos. Therefore the approach is difficult to use and understand the image with many cluttered objects in the scene. The previous methods focused on single object detection and classification based on feature set defined by humans. These proposed methods explore the connection of objects in scene classification [3]. Many scene classification technique was performed on the object bank to compute its utility. Many types of research have been conducted emphasizing their focus on low-level feature extraction for object recognition and classification, namely Histogram of oriented gradient (HOG), GIST, filter bank, and abag of feature (BoF) implemented though word vocabulary [4].

3.Methodology of Evaluation

The main aim of our work is to understand the performance of the networks for static as well as live video feeds. The first step for the following is to perform transfer learning on the networks with image datasets. This is followed by checking the

prediction rate of the same object on static images and real-time video feeds. The different accuracy rates are observed and noted and presented in the tables given in further sections. Third important criteria for evaluating the performance was to check whether prediction accuracy varies across all CNNs chosen for the study. It must be noted that videos are not used as a training dataset, they are used as testing datasets. Hence we are looking for best image classifier where the object is the main attribute for classification of scene category. Different layers of the convolutional neural network used are:

- **Input Layer:** The first layer of each CNN used is 'input layer' which takes images, resize them for passing onto further layers for feature extraction.
- **Convolution Layer:** The next few layers are 'Convolution layers' which act as filters for images, hence finding out features from images and also used for calculating the match feature points during testing.
- **Pooling Layer:** The extracted feature sets are then passed to 'pooling layer'. This layer takes large images and shrink them down while preserving the most important information in them. It keeps the maximum value from each window, it preserves the best fits of each feature within the window.
- **Rectified Linear Unit Layer:**The next 'Rectified Linear Unit' or ReLU layer swaps every negative number of the pooling layer with 0. This helps the CNN stay mathematically stable by keeping learned values from getting stuck near 0 or blowing up toward infinity.
- **Fully Connected Layer:** The final layer is the fully connected layers which takes the high-level filtered images and translate them into categories with labels.

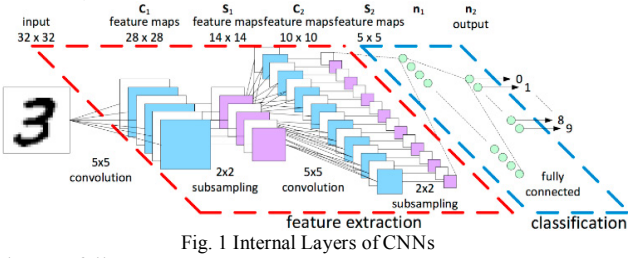

Fig. 1 Internal Layers of CNNs

The steps of proposed method are as follows:

1. **Creating training and testing dataset:** The super classes images used for training is resized [224,244] pixels for AlexNet and [227,227] pixels GoogLeNet and ResNet50, and the dataset is divided into two categories i.e. training and validation data sets.
2. **Modifying CNNs network:**Replace the last three layers of the network with fully connected layer, a softmax layer, and a classification output layer. Set the final fully connected layer to have the same size as the number of classes in the training data set. Increase the learning rate factors of the fully connected layer to train network faster.
3. **Train the network**: Set the training options, including learning rate, mini-batch size, and validation data according to GPU specification of the system. Train the network using the training data.
4. **Test the accuracy of the network:**Classify the validation images using the fine-tuned network, and calculate the classification accuracy. Similarly testing the fine tune network on real time video feeds for accurate results.

4.Models

There are various smart pre-trained CNN, these CNN have the capability of transfer learning. Therefore it just requires the training and testing datasets at its input layer. The architecture of the networks differs in terms of internal layers and techniques used. GoogLeNet has Inception Modules that perform different sizes of convolutions and concatenate the filters for the next layer [20]. On the other hand, AlexNet does not use filter concatenation, instead, it uses the output of the previous layer as the input. Both networks have been tested independently and use the implementation provided by Caffe, a Deep Learning framework [22]. ResNet is a short name for Residual Network. Many other visual recognition tasks have also greatly benefited from very deep models. So, over the years there is a trend to go deeper, to solve more complex tasks and to also increase /improve the classification/recognition accuracy. But, as we go deeper; the training of neural network becomes difficult and also the accuracy starts saturating and then degrades also [3]. Residual Learning tries to solve both these problems. In general, in a deep convolutional neural network, several layers are stacked and are trained to the task at hand. The network learns several low/mid/ high-level features at the end of its layers [15][2]. In residual learning, instead of trying to learn some features, the network tries to learn some residual. Residual can be simply understood as subtraction of feature learned from theinput of that layer. ResNet does this using shortcut connection (directly connecting theinput of nth layer to some $(n+x)^{the}$ layer [15]. It has proved that training this form of networks is easier than training simple deep convolutional neural networks and also the problem of degrading accuracy is resolved. The comparison is made among three existing neural networks i.e. the AlexNets, Google Nets and ResNet50 [21]. Followed by the transfer learning concepts for training these networks and generating new networks for further comparison. The new models have asame number of layers as that of original but the performance of these networks and existing networks varies considerably. On same images, the different accuracy rates were formulated in the tables presented in the following section.

5.Test Datasets

Image dataset of CIFAR- 100 which has numerous super-classes of general object images and a number of subclass categories of each superclass. CIFAR-100 has 100 classes of images with each class having 600 images each [15]. These

600 images are divided into 500 training images and 100 testing images for each class, therefore, making a total of 60,000 different images. These 100 classes are clubbed together into 20 superclasses. Every image in the dataset comes with a "fine" label (depicting the class to which it belongs) and a "coarse" label (superclass to the "fine" label detected). The selected categories for training and testing are abed, bicycle, bus, chair, couch, motorcycle, streetcar, table, train, and wardrobe [21][15]. For the proposed work, some wide categories of each super classes need to be used for training the networks, the superclasses used are Household furniture and vehicle. The chosen categories are shown in thetable below. The second dataset used was ImageNet datasets that has super-classes of images which is further divided into subclasses. ImageNet is an image dataset which is organized as per the WordNet hierarchy. The dataset is organized as meaningful concepts.

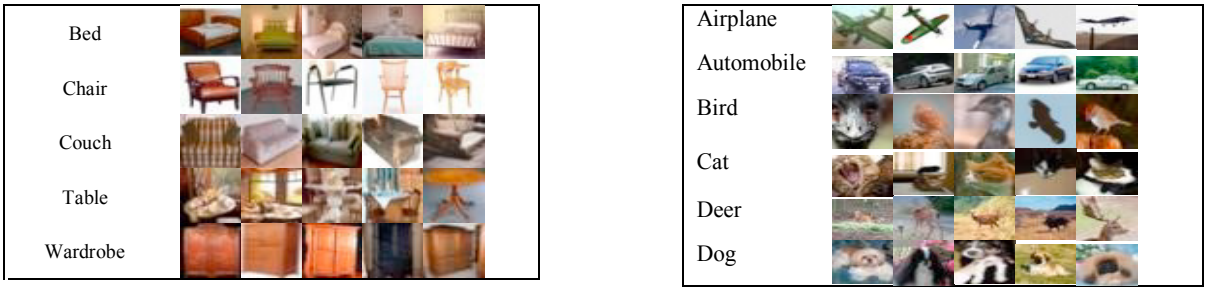

Fig 2: Few classes of CIFAR10 and CIFAR100 Datasets

Each concept in WordNet is described by many words called a "synonym set" or "sync set". The dataset contains more than 100,000 sync sets. All images are human-annotated. Furthermore, a grouping of ImageNet's less descriptive labels into more meaningful sets that matched that of the superclass was done for our study. For example, "table" was relabelled as "furniture", similarly many other images were grouped into their superclasses and created a more descriptive and meaningful label. The third dataset chosen for the study was aCIFAR-10 dataset of images. The CIFAR-10 dataset has 32x32 color images divided into 10 classes and 6000 images per class, which makes a total of 60000 images. The dataset consists of 50000 training images and 10000 test images. The dataset is divided into five training batches and one test batch, each of which has 10000 images. The test images are randomlyselected from each class.

Table 1. Performance of CNN's on CIFAR100 test dataset

CIFAR-100		AlexNet	GoogLeNet	ResNet50
Image Category	Bed	0.00%	70.80%	49.60%
	Bicycle	21.0	74.2%	55.00%
	Bus	84.00%	63.20%	36.80%
	Chair	90.00%	89.60%	57.60%
	Couch	11.00%	14.60%	76.40%
	Motorcycle	95.00%	74.60%	99.20%
	Streetcar	21.00%	0.84%	63.80%
	Table	00.00%	73.60%	33.40%
	Train	30.00%	95.60%	34.20%
	Wardrobe	89.00%	89.40%	92.20%

Table 2. Performance of CNN's on the CIFAR10 test dataset

CIFAR-10		AlexNet	GoogLeNet	ResNet50
Image Category	Airplane	41.80%	51.10%	90.80%
	Automobile	21.80%	62.10%	69.10%
	Bird	00.02%	56.70%	72.60%
	Cat	00.03%	78.80%	61.90%
	Deer	87.60%	49.50%	75.40%
	Dog	23.00%	57.50%	82.10%
	Frog	24.20%	90.20%	76.60%
	Horse	34.70%	78.20%	84.70%
	Ship	31.70%	95.50%	83.20%
	Truck	95.90%	97.10%	84.60%

6.Results

The performance analysis of CNN's is done by testing each of the networks on CIFAR-100 and CIFAR-10 datasets. Table 1 depicts the accuracy of various image categories of CIFAR- 100 test dataset. For example, out of 100 test images of Bus, Alex Net predicts 84 images label correctly, whereas GoogLeNet detects bus in around 63 images and ResNet50 classifies 37 images labeled as abus. Table 1 and Table 2 show the prediction accuracy of CNN's when tested for various image categories of CIFAR- 100 and CIFAR-10 test datasets. For 100 images of Horse, AlexNet identifies horse in 35 images, GoogLenet finds ahorse in 78 images and ResNet50 classifies 85 images as horse labeled.Considering the probability values of all three CNN's calculated from confusion matrix after testing, a detailed preview of prediction done by three CNN's are as follow.

Table 3. Performance on Bicycle class of CIFAR-100 dataset

AlexNet's Output	Prediction Accuracy (%)	GoogLeNet's Output	Prediction Accuracy (%)	ResNet50 Output	Prediction Accuracy (%)
Motorcycle	45	Bicycle	74.2	Bicycle	55
Bus	28	Train	13	Motorcycle	35
Bicycle	21	Table	7.6	Streetcar	4.4
Chair	2	Motorcycle	4.4	Couch	2.6
Train	2	Chair	0.4	Bed	1
Streetcar	1	Wardrobe	0.2	Train	0.8
Wardrobe	1	Bus	0.2	Wardrobe	0.6
Couch	0	Streetcar	0	Table	0.6
Bed	0	Couch	0	Bus	0
Table	0	Bed	0	Chair	0

Table 4. Performance on Chair class of CIFAR-100 dataset

AlexNet's Output	Prediction Accuracy (%)	GoogLeNet's Output	Prediction Accuracy (%)	ResNet50 Output	Prediction Accuracy (%)
Chair	90	Chair	89.6	Chair	57.6
Wardrobe	5	Bed	7	Couch	21
Bus	3	Table	2.8	Bed	7.4
Motorcycle	1	Wardrobe	0.4	Wardrobe	5.8
Couch	1	Train	0.2	Train	5.4
Bed	0	Bicycle	0	Motorcycle	2
Bicycle	0	Bus	0	Streetcar	0.6
Streetcar	0	Couch	0	Bicycle	0.2
Table	0	Motorcycle	0	Bus	0
Train	0	Streetcar	0	Train	0

Table 3 depicts the prediction accuracy of all three networks for Bicycle class. We can see that AlexNet's top prediction for bicycle class is a motorcycle. GoogLe Net shows best performance and ResNet gives the average result. Similarly, Table 4 shows the output of CNN's for chair class.

Table 5. Performance on Deer class of CIFAR-10 dataset

AlexNet's Output	Prediction Accuracy (%)	GoogLeNet's Output	Prediction Accuracy (%)	ResNet50 Output	Prediction Accuracy (%)
Deer	87.6	Deer	49.5	Deer	75.4
Horse	3.7	Horse	24.4	Horse	10.7
Ship	3.4	Cat	13.3	Bird	3.5
Frog	2.2	Frog	6	Airplane	3.3
Truck	1.6	Bird	3	Dog	2.6
Airplane	1.2	Ship	2	Cat	2.5
Automobile	0.2	Airplane	1.1	Frog	1.6
Dog	0.1	Truck	0.3	Ship	0.3
Bird	0	Dog	0.4	Truck	0.1
Cat	0	Automobile	0	Automobile	0

Table 6. Performance on Ship class of CIFAR-10 dataset

AlexNet's Output	Prediction Accuracy (%)	GoogLeNet's Output	Prediction Accuracy (%)	ResNet50 Output	Prediction Accuracy (%)
Truck	50.6	Ship	95.5	Ship	83.2
Ship	31.7	Truck	2.2	Airplane	14.4
Airplane	12.3	Cat	1.2	Truck	0.5
Deer	3.1	Airplane	0.6	Cat	0.5
Automobile	1.5	Automobile	0.3	Horse	0.4
Horse	0.8	Bird	0.2	Dog	0.3
Bird	0	Deer	0	Bird	0.3
Cat	0	Dog	0	Deer	0.2
Dog	0	Frog	0	Automobile	0.1
Frog	0	Horse	0	Frog	0.1

Table 5 compares the output of three networks for Deer class. In other words, both the networks provide consistently correct classifications. By observing all the tables, the classifications accuracy obtained for all images across all categories, are different. AlexNet essentially see a Motorcycle in top prediction, while GoogLeNet and ResNet50 see a bicycle in top prediction for bicycle class. For other less frequent classes, there is still a large overlap across different categories.Similarly, Table 6 presents results for the ship class. The predicted label along with its score shows how accurately the object is detected by a particular network. While analyzing each table independently, one can observe that for most of the categories of Cifar-100 dataset, GoogleNet does the correct labeling and classification while ResNet50 identifies an average number of classes of CIFAR-100 dataset.But for CIFAR – 10 ResNet50 shows best classification results and GoogLeNet remains average. Nonetheless, both networks are quite consistent, having high counts for a small subset of classes. The reason for this behavior seems to be the fact that most classifiers are trained for object categories that contain simple, thin traces in their composition, such as safety pins and bowstrings. It is therefore understandable that the networks may mistake with appearance and properties of objects.

Table 7. Performance of CNNs on live video feeds

Object Category	AlexNet Prediction Accuracy(%)	GoogleNet Prediction Accuracy(%)	ResNet50 Prediction Accuracy(%)	Object Category	AlexNet Prediction Accuracy(%)	GoogleNet Prediction Accuracy(%)	ResNet50 Prediction Accuracy(%)
Bed	12	85	25	Airplane	14	84	96
Bicycle	11	80	55	Automobile	12	59	56
Bus	14	74	25	Bird	11	45	53
Chair	12	47	30	Cat	11	62	49
Couch	12	25	90	Deer	12	45	33
Motorcycle	14	50	35	Dog	12	57	58
Streetcar	11	45	25	Frog	13	60	25
Table	11	63	50	Horse	12	87	65
Train	15	72	45	Ship	15	91	25
Wardrobe	14	84	32	Truck	22	95	52

The real-time analysis of the performance of convolutional neural networks shows that Alex Net has overall 13% accuracy of detecting correct objects in the scene. Similarly, GoogleNet and ResNet50 classification is 68.95% and 52.55% correct. It can be observed that performance of CNN's on images vary substantially compared to live testing results. In live testing, CNNs get confused between few objects, for example, ResNet50 often has a problem in classifying dog and deer. It detects them as a horse in most of the scenes. The accuracy results prove that GoogleNet performance is better and detection accuracy is highest compared to all other nets.

7. Evaluation

Both of the CNN produce a probability distribution in the possible input classes. Two different methods were used to calculate the results. The first method only considers the 10 most probable classes and the second register the position of the correct class in the full probability range. In the first method, we classify the results of the network according to their probability and consider only the ten most probable classes. We count how many times each class appears for each image in

each target category. This method allows you to evaluate if a good and useful probability is assigned to the correct result, but also to observe qualitatively the consistency of the results for each category i.e., it is expected that for each category, the top 10 probabilities do not vary significantly. In the second method, we construct descriptive statistics about the position of the correct class in the probability range. This is achieved by ranking the results obtained by the classifier. The higher the rank, the better the classification is. Ideally, the correct class will be in first place. Calculate the mean and the standard deviation for each category. A low average corresponds to a higher position in the rankings, while a low standard deviation is a proof of the consistency of production for the different instances of the same category. It also allows you to capture the best and worst instances of each category that we use to analyze the possible reasons for the observed results. Finally, we can infer from the obtained results that the average performance of these three networks on CIFAR100 dataset is found to be as: for AlexNet average performance is 44.10 %, for GoogLeNet it is 64.40% and for ResNet50 an average performance of 59.82% is reported by our experimental study [20]. Similarly, the average performance of CNN's for theCIFAR10 dataset is as follows: for AlexNet- 36.12 %, for GoogLeNet- 71.67%, and for ResNet50- 78.10% is found.

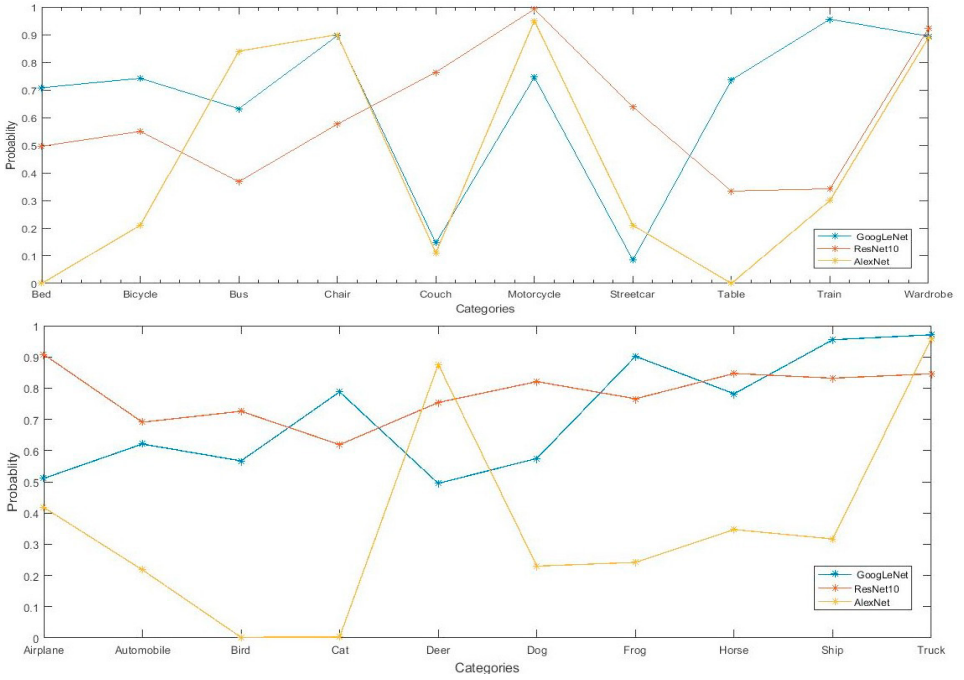

Fig 3: (a) Probability vs Categories graph for CIFAR- 100 dataset (b) Probability vs Categories graph for CIFAR- 10 dataset

8. Conclusion

The work analyzed the prediction accuracy of three different convolutional neural networks (CNN) on most popular training and test datasets namely CIFAR10 and CIFAR100. We focused our study on 10 classes of each dataset only. Our main purpose was to find out the accuracy of the different networks on same datasets and evaluating the consistency of prediction by each of these CNN. We have presented a thorough prediction analysis for comparing the networks' performance for different classes of objects. It is important to note that complex frames often create confusion for the network to detect and recognize the scene. It was also noted that though in real-world beds and couches as well as chair are different and easily recognized objects but the trained networks showed confusion and therefore differ in accuracy rates. The results suggested that trained networks with transfer learning performed better than existing ones and showed higher rates of accuracy. Few objects like "chair", "train" and "wardrobe" were perfectly recognized by 147 layered networks whereas objects like "cars" were perfectly recognized by 177 layered networks. From our experiments, we could easily conclude that the performance of 27 layered networks was not much appreciated. Hence, more the number of layers, more will be the training and therefore, higher the rate of accuracy in prediction will be achieved. It can further be summed up that neural networks are new and best emerging techniques for making amachine intelligent for solving many real-life object categorization problems. Many types of research and works are being done on it. It has wide applications and it is easy and flexible to integrate into various platforms. The hardware requirements may not allow the network to be trained on normal desktop work but just with nominal requirements one can train the network and generate the desired model.

References

[1] Kou, F., Du, J., He, Y., & Ye, L. (2016) "Social Network Search Based on Semantic Analysis and Learning." *CAAI Transactions on Intelligence Technology*.

[2] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V.,& Garcia-Rodriguez, J. (2017) "A Review on Deep Learning Techniques Applied to Semantic Segmentation."

[3] Li, L. J., Su, H., Lim, Y.,& Li, F. F. (2010, September) "Objects as Attributes for Scene Classification."*ECCV Workshops***(57-69).**

[4] Srinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., &Babu, R. V. (2016) "A taxonomy of deep convolutional neural nets for computer vision."

[5] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., &Torralba, A. (2014) "Object detectors emerge in deep scene cnns."

[6] Wang, Y., & Wu, Y. "Scene Classification with Deep Convolutional Neural Networks."

[7] Lowe, D. G. (2004) "Distinctive image features from scale-invariant keypoints." *International journal of computer vision* **60(2).**

[8] Dalal, N., &Triggs, B. (2005, June) "Histograms of oriented gradients for human detection."*In Computer Vision and Pattern Recognition, 2005. CVPR 2005.*

[9] Yang, J., Jiang, Y. G., Hauptmann, A. G.,& Ngo, C. W. (2007, September)"Evaluating bag-of-visual-words representations in scene classification." in Proceedings of *the international workshop on Workshop on multimedia information retrieval*.

[10] Cheung, Y. M., & Deng, J. (2014, October) "Ultra local binary pattern for image texture analysis." in Security*Pattern Analysis, and Cybernetics (SPAC), 2014 International Conference*.

[11] Khan, S. M. H., Hussain, A., &Alshaikhli, I. F. T. (2012, November) "Comparative study on content-based image retrieval (CBIR)." in Advanced Computer Science Applications and Technologies (ACSAT), 2012 *International Conference*.

[12] Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997) "Face recognition: A convolutional neural-network approach."*IEEE transactions on neural networks, 8(1)***:**98-113.

[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ...&Rabinovich, A. (2015)"Going deeper with convolutions." in Proceedings of the *IEEE conference on computer vision and pattern recognition* (pp. 1-9).

[14] Bobić, V., Tadić, P., &Kvaščev, G. (2016, November) "Hand gesture recognition using neural network based techniques." in Neural Networks and Applications (NEUREL), *2016 13th Symposium on (pp. 1-4). IEEE.*

[15] Krizhevsky, A., & Hinton, G. (2009)"Learning multiple layers of features from tiny images."

[16] LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., ... &Vapnik, V. (1995) "Learning algorithms for classification: A comparison on handwritten digit recognition."*Neural networks: the statistical mechanics perspective*(pp 261-276).

[17] LeCun, Y., Bottou, L., Bengio, Y., &Haffner, P. (1998) "Gradient-based learning applied to document recognition." proceedings of the *IEEE* **86(11)**: 2278-2324.

[18] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., &Salakhutdinov, R. (2014) "Dropout: a simple way to prevent neural networks from overfitting."*Journal of machine learning research* **15(1)**: 1929-1958.

[19] Eitz, M., Hays, J., & Alexa, M. (2012) "How do humans sketch objects?"*ACM Trans. Graph., 31(4)*.

[20] Ballester, P., & deAraújo, R. M. (2016, February)"On the Performance of GoogLeNet and AlexNet Applied to Sketches." in *AAAI*.

[21] Yang, Y., &Hospedales, T. M. (2015)"Deep neural networks for sketch recognition".

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

[23] Ciresan, Dan, Ueli Meier, and JurgenSchmidhuber, (2012) ""Multi-column deep neural networks for image classification."*2012 IEEE Conference on Computer Vision and Pattern Recognition.*

[24] Ciresan, Dan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and JurgenSchmidhuber. (2011) "Flexible, High Performance Convolutional Neural Networks for Image Classification." *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Two:* 1237–1242.

[25] Lawrence, Steve, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. (1997) "Face Recognition: A Convolutional Neural Network Approach." *IEEE Transactions on Neural Networks, Volume 8; Issue 1.*