Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

Object Recognition Using Deep Learning

Ogunti Erastus
Electrical Engineering Department,
Federal University of Technology
Akure, Nigeria

Ale Daniel
IPNX Nigeria Limited,
Lagos, Nigeria

Nwabueze Ifeoma Blessing
Department of Computer Engineering
Afe Babalola University,
Ado-Ekiti, Nigeria

***Abstract*— Object recognition is an active area of research and huge successes have been made in the last few years, for instance the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2012) which introduced a new frontier in object recognition. In this paper we develop a structured Convolutional Neural Network (CNN), model for supervised learning applied to object recognition. This paper attempts to investigate the performance of Adams Optimizer as opposed to the popular Stochastic Gradient Descent (SGD) approach that have been largely used in literature. The publicly available CIFAR(Canadian Institute For Advanced Research)10 datasets was used for the training of this network and an accuracy of 98.5% was achieved which proves that deep learning has surpassed human capability in recognition of tasks.**

***Keywords*— CNN, Adams, AI, Deep Learning.**

I. INTRODUCTION

Object recognition is one of the most fascinating abilities that humans easily possess since childhood. With a simple glance of an object, humans are able to tell its identity or category despite the appearance variation due to change in pose, illumination, texture, deformation, and occlusion. Furthermore, humans can easily generalize from observing a set of objects to recognizing objects that have never been seen before. For example, kids are able to generalize the concept of "chair" or "cup" after seeing just a few examples. Nevertheless, it is a daunting task to develop vision systems that match the cognitive capabilities of human beings, or systems that are able to tell the specific identity of an object being observed [1].

Generally, artificial intelligence (AI) is a way of making a computer, a computer-controlled robot or a software think intelligently, in a similar manner as an intelligent human. AI is accomplished by studying how human brains think, how human brains learn, decide, and work while trying to solve a problem, and then using the outcome of this study as a basis for developing intelligent software and systems. The study of human intelligence and its replicate has brought about several fields which include, but is not limited to neural networks, evolutionary computations, computer vision, robotics, expert systems, speech processing, natural language processing, planning, machine learning and fuzzy logic [2].

Machine learning (ML) is a subfield of AI, but is often also referred to as predictive analytics, or predictive modeling. Its goal and usage is to build new and/or leverage existing algorithms to learn from data, in order to build generalizable models that give accurate predictions, or to find patterns, particularly with new and unseen similar data. This helps to create systems which exhibit more intricate and complicated behavior. Even though the behavior of these system may be statistically consistent with the training data, they may find patterns we were unaware of, and thus may exhibit an unexpected behavior, which is somehow controllable via training data. This quality of machine learning is both its greatest strength and also its weakest, as it can introduce biases through over-fitting and under-fitting, or even learn undesired biases found in the training data. These dangers are more amplified as machine learning algorithms are relatively difficult to debug.

Machine Learning has been around for many decades. In 1948 Alan Turing describes how machines could be design to learn using "B-type unorganized machines" [3], conceptual precursor to modern day artificial neural networks. For many years, ML remained primarily an academic research area as ML was widely accepted due to its high computational requirements and inferior performance compared to other AI methods [4]. However, advances in ML algorithms, and increases in computing power, specifically high paralleled GPU (Graphics Processing Unit) computing has enabled dramatic advancements in how ML can be applied [5]. Gradually ML has outperformed other AI techniques in field such as speech recognition [6], natural language processing [7], computer vision [8], email spam filtering [9], image captioning [10], robotics and self-driving cars [11].

The major types of ML are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

A. Supervised Learning (SL)

In supervised learning (SL), the model is trained on labelled data, where each training example is an input-target pair. During training, the learning algorithm tries to learn the model parameters which effectively implement a function that maps the input of each training pair, to the associated target. These targets can also be thought of as a supervisory signal. Having input-target pairs makes it relatively straight forward to specify the objective function, thus right now SL is one of the most popular and successful branches of ML. However, the training pairs often need to be manually associated by people. This makes them cumbersome and very time consuming to prepare. Recent development has helped accelerate the preparation of large labelled datasets which is why we're starting to see more success in this field [12].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

B. Semi-supervised Learning (SSL)

Semi-supervised learning (SSL) is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. SSL falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, SSL can be of great practical value. SSL is also of theoretical interest in machine learning and as a model for human learning. [13].

C. Unsupervised Learning (UL)

In unsupervised learning (UL), training is performed on unlabeled data. Without an external supervisory signal, it can be more ambiguous as to how to specify the Objective function, and unsupervised learning is currently one of the big open problems in ML [14]. One of the common training objectives of UL is found in an Auto-Encoder, in which the target is the same as the input, and the learning algorithm tries to learn how to compress and decompress each training example with minimal loss. If successful, it finds regularities in the training data, and learns more compact and meaningful representations. Another common objective is clustering, in which the learning algorithm tries to organize the training data into groups based on similarities that it tries to learn. When new inputs are presented to a model trained with one of these unsupervised learning methods, the model can transform the input data to one of the more compact representations, or predict how it relates to other data based on the patterns it has already found.

D. Reinforcement Learning (RL)

In reinforcement learning (RL), the model is not trained with labels associating inputs with targets. It is neither supervised (with a direct supervisory signal) nor unsupervised (with no supervisory signal), but instead there is a delayed reward signal [10]. The terminology is slightly different in RL, where decisions or predictions are called actions, and the decision making (or action taking) entity is called an agent. This is because RL is based on a Markov Decision Process (MDP) [15].

The general objective of the algorithm is to learn what the optimal decisions are by maximizing its long term reward. This process also involves a balance between exploration (of new actions which haven't yet been made) and exploitation (of actions which are known to reward higher than others). RL can also be thought of as learning by trial and error.

The human brain contains over ten billion neurons. These neurons receive input signals from its dendrites and produce output signals along its axon. The axon branches out and connects via synapses to dendrites of other neurons. These connections are known as synapses, and the human brain contains about 60 trillion such connections. When the combination of input signals reaches some threshold condition among its input dendrites, the neuron is triggered, and its activation is communicated to successor neurons.

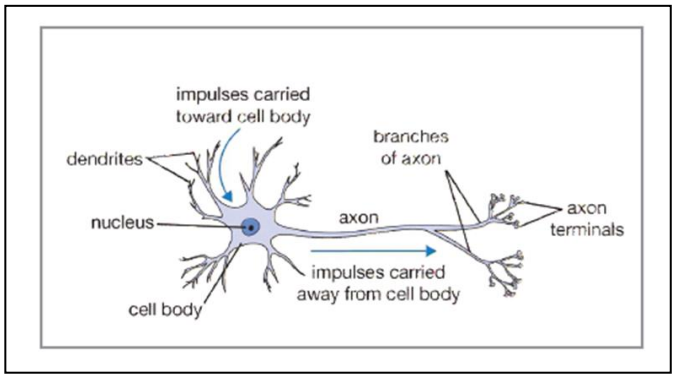

Fig. 1. Biological Neuron.

Artificial neural networks (ANN) are modeled on the human brain and consist of a number of artificial neurons. The neurons in artificial neural networks tend to have fewer connections than biological neurons, and neural networks are all significantly smaller in terms of number of neurons than the human brain. Each neuron (or node) in a neural network receives a number of inputs, each of these inputs have numeric weights that are tuned during the training process, so that a properly trained network will respond correctly when presented with an image or pattern to recognize. A function called the activation function is applied to these input values, which results in the activation level of the neuron, which is the output value of the neuron [16]. Below is a mathematical representation of artificial neural network.

$$y(k) = F\left(\sum_{i=0}^{m} w_i(k).x_i(k) + b\right) \quad (1)$$

where

$x_i(k)$ -input value in discrete time k where i goes from 0 to m,

$w_i(k)$ -weight value in discrete time k where i goes from 0 to m,

b - bias,

F - transfer function,

$y_i(k)$ - output value in discrete time k.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

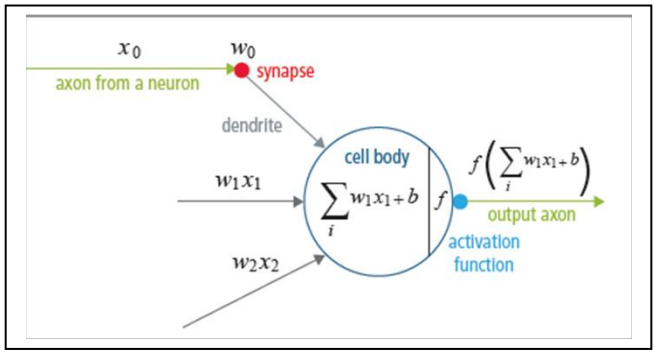

Fig. 2. Artificial Neuron.

CNN is a special case of the neural network, which consists of four sections; convolutional layer, non-linear layer, pooling layer, fully-connected layer. These layers will be further discussed in the system design

II. LITERATURE REVIEW

The field of AI research was founded at a conference at Dartmouth College in 1956. The attendees, including John McCarthy, Marvin Minsky, Allen Newell, Arthur Samuel and Herbert Simon, became the leaders of AI research [17] They and their students wrote programs that were astonishing to most people. Computers were winning at checkers, solving word problems in algebra, proving logical theorems and speaking English. By the middle of the 1960s, research in the U.S. was heavily funded by the Department of Defense [18] and laboratories had been established around the world. AI's founders were optimistic about the future, Herbert Simon predicted, "machines will be capable, within twenty years, of doing any work a man can do." Marvin Minsky agreed, writing, "Within a generation, the problem of creating 'artificial intelligence' will substantially be solved". They failed to recognize the difficulty of some of the remaining tasks. Progress slowed and in 1974, in response to the criticism of Sir James Lighthill [19] and ongoing pressure from the US Congress to fund more productive projects, both the U.S. and British governments cut off exploratory research in AI.

In the early 1980s, AI research was revived by the commercial success of expert systems a form of AI program that simulated the knowledge and analytical skills of human experts. By 1985 the market for AI had reached over a billion dollars. At the same time, Japan's fifth generation computer project inspired the U.S and British governments to restore funding for academic research. However, beginning with the collapse of the Lisp Machine market in 1987, AI once again fell into disrepute, and a second, longer-lasting hiatus began. In the late 1990s and early 21st century, AI began to be used for logistics, data mining, medical diagnosis and other areas. The success was due to increasing computational power, greater emphasis on solving specific problems, new ties between AI and other fields and a commitment by researchers to mathematical methods and scientific standards. Deep Blue became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov on 11 May 1997.

According to Bloomberg's Jack Clark, 2015 was a landmark year for artificial intelligence, with the number of software projects that use AI within Google increasing from a "random usage" in 2012 to more than 2,700 projects. Clark also presents factual data indicating that error rates in image processing tasks have fallen significantly since 2011. He attributes this to an increase in affordable neural networks, due to a rise in cloud computing infrastructure and to an increase in research tools and datasets. Other cited examples include Microsoft's development of a Skype system that can automatically translate from one language to another and Facebook's system that can describe images to blind people [20].

III. SYSTEM DESIGN

This project centers on supervised learning, a learning algorithm in which the desired output for the network is also provided with the input while training the network. By providing the neural network with both an input and output pair it is possible to calculate an error based on its target output and actual output. It can then use that error to make corrections to the network by updating its weights.

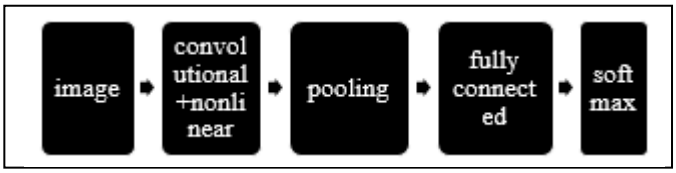

Fig. 3. Process Flow in a CNN.

The input layer defines the type and size of data (image) the CNN can process. In this project, CIFAR (Canadian Institute for Advanced Research) 10 dataset will be used to process the CNN, which are 32x32 RGB images. The middle layers are made up of repeated blocks of convolutional, nonlinear and pooling layers. The final layers (output layer) are typically composed of fully connected layer and a SoftMax layer.

A. *Image*

The CIFAR10 dataset is a collection of 32x32 pixel size images which belong to one of 10 distinct classes. Each class contains 5000 training samples and another 1000 samples for testing, making a total of 50000 training images and 10000 testing images. Even within the same class, images vary greatly from one another. The image may vary in size and color or texture of the image may vary greatly as well. [21], [22].

B. *Convolutional layer*

The Convolutional layer (conv) is the core building block of a Convolutional Network that does most of the computational heavy lifting. The CONV layer's parameters consist of a set of learnable filters. Every filter is small spatially (along width and height), but extends through the full depth of the input volume. The mathematical representation is given as:

$$y_{rc}^{l} = \sum_{i=1}^{F_r} \sum_{j=1}^{F_c} y_{(r+i-1)(c+j-1)}^{l-1} w_{ij}^{l} + b^{l} \quad (2)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

Where, y^l_{rc} is the output volume at $\{r, c\}$, F_r and F_c are the number of rows and columns in the 2D filter, w^l_{ij} is the value of the filter at position {i, j}, $y^{l-i}_{(r+i-1)(c+j-1)}$ is the value of input to this layer, at position $\{r+i-1, c+j-1\}$, and b^l is the bias term.

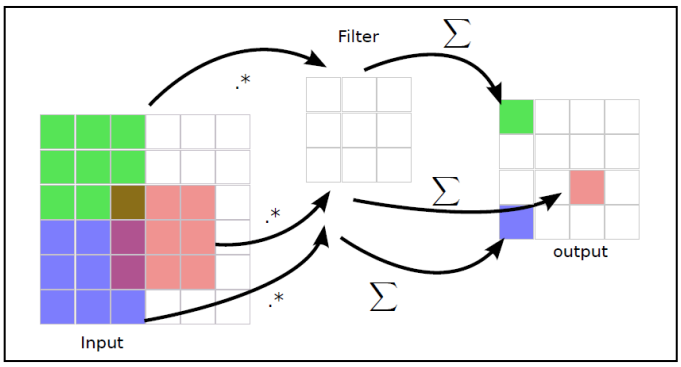

Fig. 4. Pictorial representation of Convolutional Layer.

C. Nonlinear Layer

After the convolutional layer, it is important to attach a nonlinear function (activation function). This nonlinear function increases the nonlinear (the output is not directly proportional to the input) properties of the model. There are two types of nonlinear function; rectified linear unit (ReLU) and continuous triggered (nonlinear) function. The continuous triggered function consists of hyperbolic tangent function, absolute hyperbolic tangent function, sigmoid function and tanh function. The ReLU applies the function f(x) = max (0, x) to all the values in the input volume (output from the convolutional layer). In basic terms, this layer just changes all the negative activations to 0. So, the input and output size of this layer are the same. For this model, ReLU was used, because with ReLU the network train many times faster compare to continuous triggered function [23].

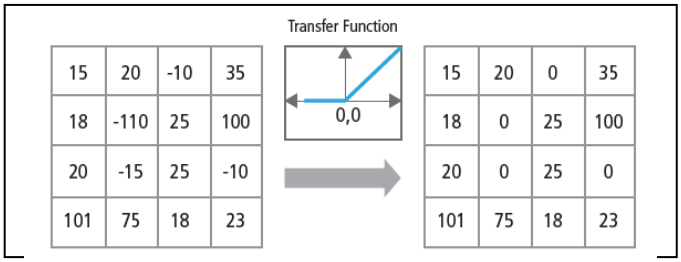

Fig. 5. Pictorial representation of RELU Function.

D. Pooling Layer

It is common to periodically insert a pooling layer in between successive nonlinear layer in the CNN, reasons being that it helps to reduce the amount of parameter and computations in the network and also to control overfitting. There are two types of pooling; max pooling and average pooling. For instance, if the input is of size 4x4. For 2x2 subsampling, a 4x4 image is divided into four non-overlapping matrices of size 2x2. In the case of max pooling, the maximum value of the four values in the 2x2 matrix is the output. In case of average pooling, the average of the four values is the output.

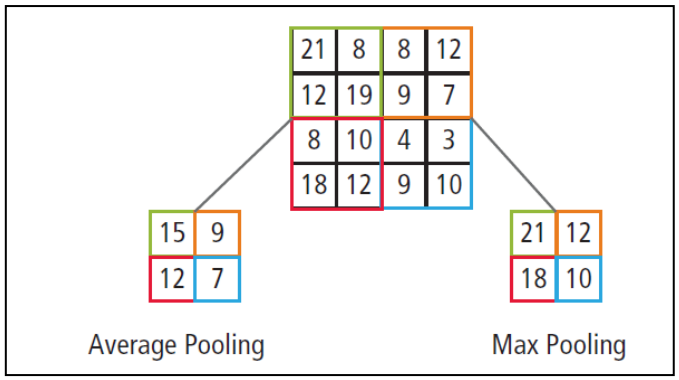

Fig. 6. Pictorial representation of Max and Average Pooling

For this model, max pooling was used, because overtime, researchers have shown that max pooling works better compare to average pooling [23].

E. Fully Connected Layer (FCL)

After several convolutional, ReLU and max pooling layers, the high-level reasoning in the network (CNN) is done via fully connected layers. The output from the convolutional, ReLU and max pooling layers represent high-level features of the input image [24]. The purpose of the FCL is to use these features for classifying the input image into various classes based on the training dataset. Apart from classification, adding a FCL is also a (usually) cheap way of learning non-linear combinations of these features. Most of the features from convolutional and pooling layers may be good for the classification task, but combinations of those features might be even better [23].

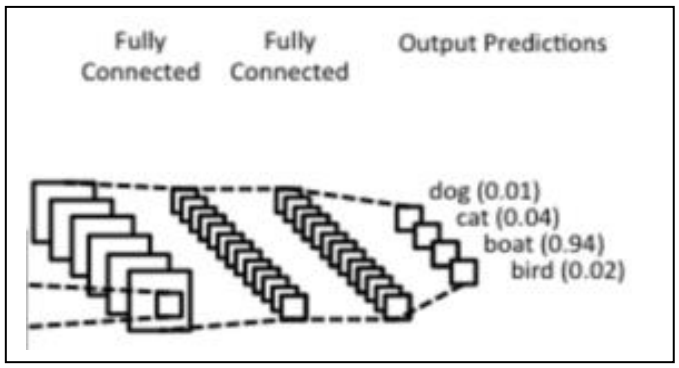

Fig. 7. Pictorial representation of Fully Connected Layer.

F. Softmax Classifier

The softmax classifier gives a slightly more intuitive output (normalized class probabilities) and also has a probabilistic interpretation. This is because it is much easier for us as humans to interpret probabilities rather than margin scores (such as in hinge loss). In the softmax classifier, the function mapping $f(xi;W) = Wxi$ stays unchanged, but we now interpret these scores as the unnormalized log probabilities for each class which amounts to swapping out the hinge loss function with cross-entropy loss.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

The loss function should minimize the negative log likelihood of the correct class:

$$L_i = -logP(Y = y_i|X = x_i) \quad (3)$$

The probability statement is interpreted as:

$$P(Y = y_i|X = x_i) = \frac{e^{s_{y_i}}}{\sum_j e^{s_j}} \quad (4)$$

Therefore, the Softnax Classifier is given as:

$$L_i = -log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right) \quad (5)$$

Computing the cross-entropy loss over an entire dataset is done by taking the average:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i \quad (6)$$

After the softmax classifier, the total error is being calculated and it is given as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - h_\theta(x_i))^2 \quad (7)$$

Where:

$h_\theta(x_i)$ = output from the softmax layer

θ = weight

m = number of images

Using backpropagation, the gradients of the error is calculated with respect to all weights (filter values) in the network and use Adams optimizer to update all weights (filter values) to minimize the output error.

Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on the vector are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: Stepsize

Require: $\beta_1, \beta_2 \in [0,1]$: Exponential decay rates for the moment estimates.

Require: $J(w)$: Stochastic objective function with parameters w

Require: w_0: Initial parameter vector

$m_0 \leftarrow 0$ (Initialize 1st moment vector)

$v_o \leftarrow 0$ (Initialize 2nd moment vector)

$t \leftarrow 0$ (Initialize timestep)

while w_t not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_w J_t(\theta_{t-1})$ (Get gradients w.r.t stochastic objective at timestep t)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)

$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)

$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)

end while

return θ_t (Resulting parameters).

IV. RESULT AND DISCUSSION

After training the network, an accuracy of 98.5% was achieved. The model was tested using the test images from the CIFAR10 dataset.

Before an image is passed through the model, the image is first smoothened as a pre-processing operation.

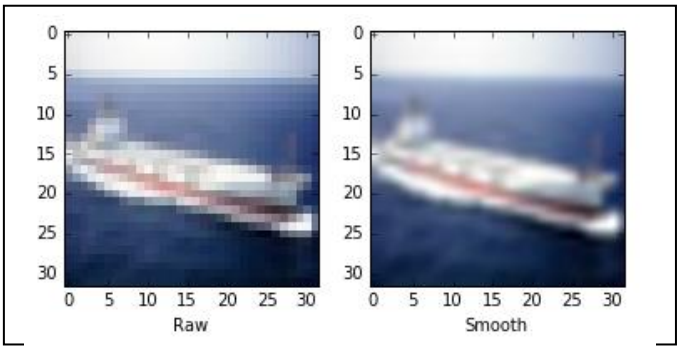

Fig. 8. Plot of a raw and smooth image.

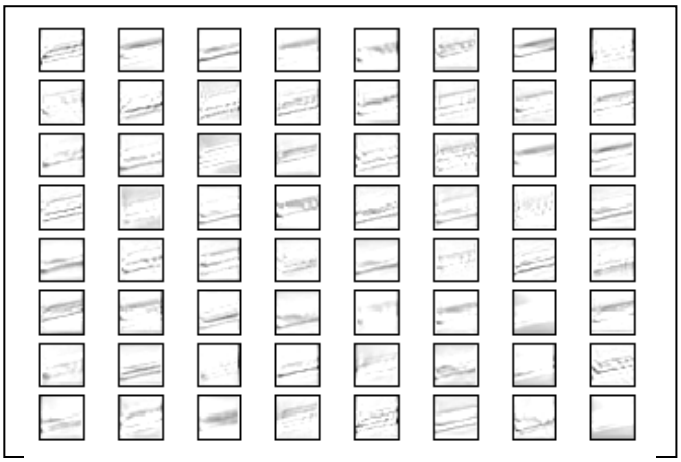

Fig. 9. Output from the first convolutional layer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

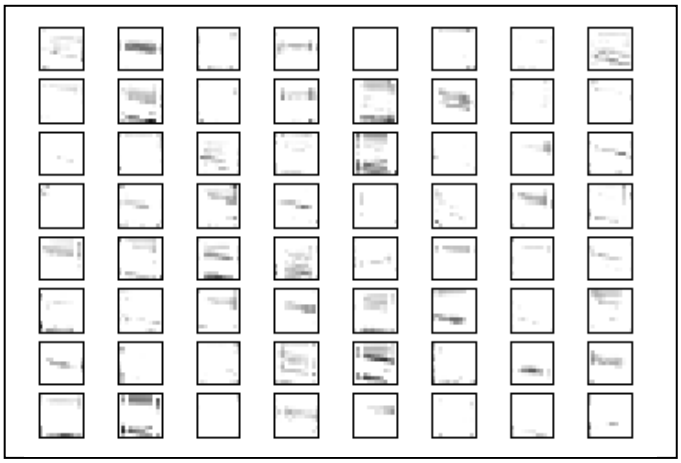

Fig. 10. Output from the second convolutional layer.

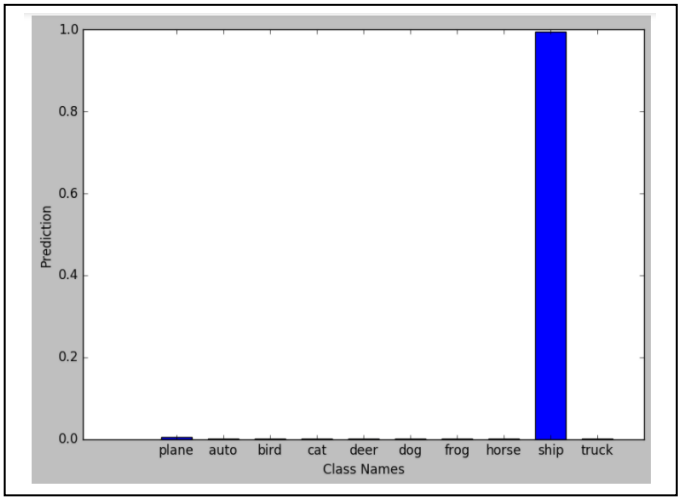

Fig. 11. Bar chart showing the output probability.

Fig. 11, shows the output of the model using a bar chart from the bar chart, one can say the predicted image by the network is a ship. Which happens to correspond with the true image (Fig. 8).

Below are some of the test carried out and their result:

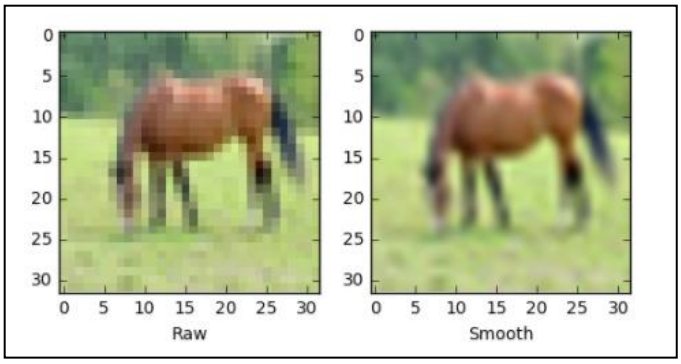

Fig. 12. Plot of a raw and smooth image.

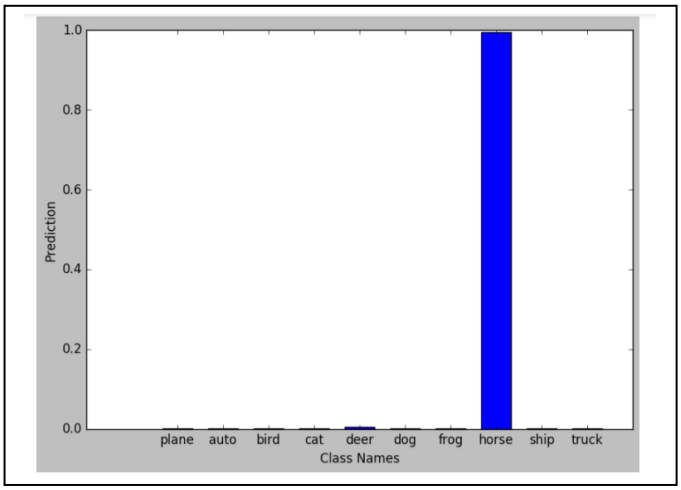

Fig. 13. Bar chart showing the Output Probability.

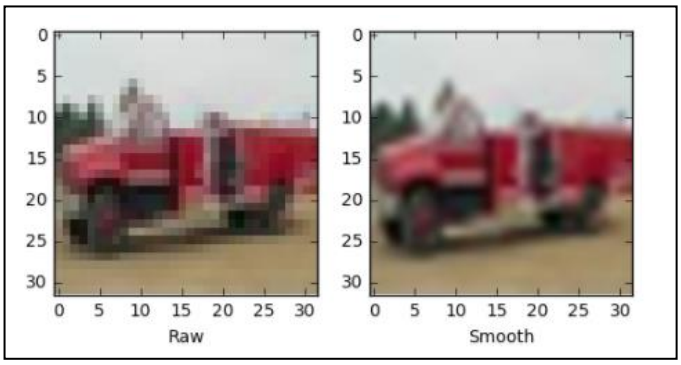

Fig. 14. Plot of a raw and smooth image.

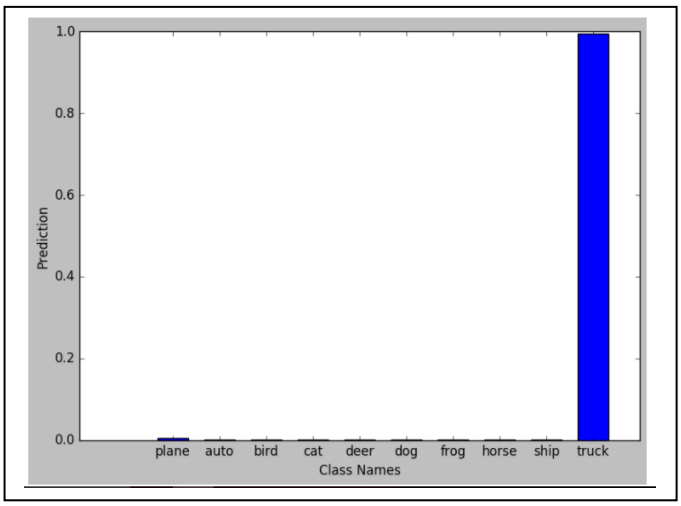

Fig. 15. Bar chart showing the Output Probability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

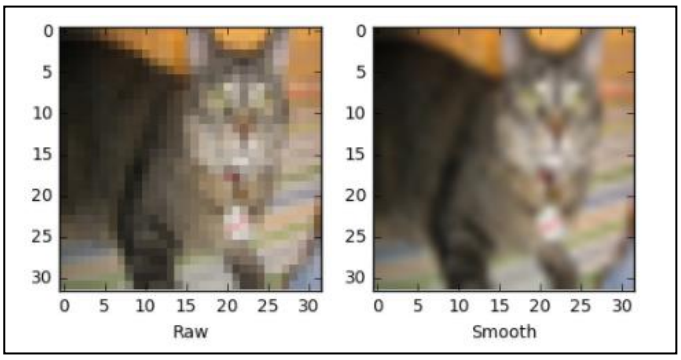

Fig. 16. Plot of a raw and smooth image.

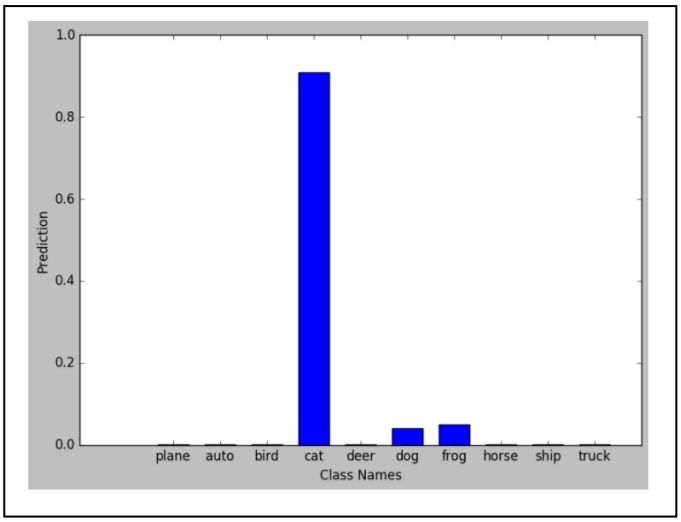

Fig. 17. Bar chart showing the Output Probability.

Looking at the test carried out and their result, one cannot easily give a prior knowledge (recognition) of the image, due to the fact that the image is not so clear. From the bar chart, the network was able to produce a correct output, despite the fact that the image is not so clear, with this, one can say that the network is accurate.

Although, some errors were made by the model due to the fact that an accuracy of 98.5% was achieved, that is out of the 10,000 test images, only 9850 images were correctly recognized whereas the remaining 150 were not. That does not mean that the network is not accurate because it is impossible to achieve an 100% accuracy, but an accuracy higher than 98.5% can be achieved by increasing the depth of the network and also increasing the number of epoch (forward and backward propagation for all the training images). The number of images affects the accuracy of the network that is, the larger the dataset, the more accurate the network. Below are some of the errors made by the model.

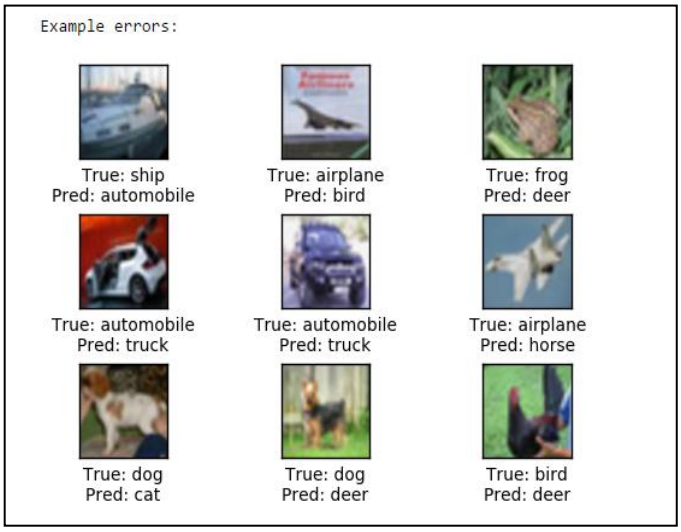

Fig. 18. Examples errors made by the model during testing.

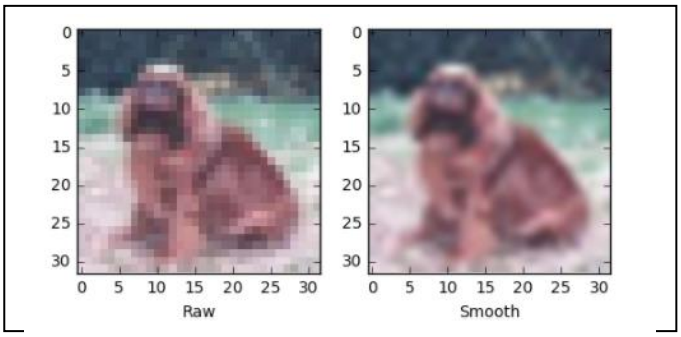

Fig. 19. Plot of a raw and smooth image.

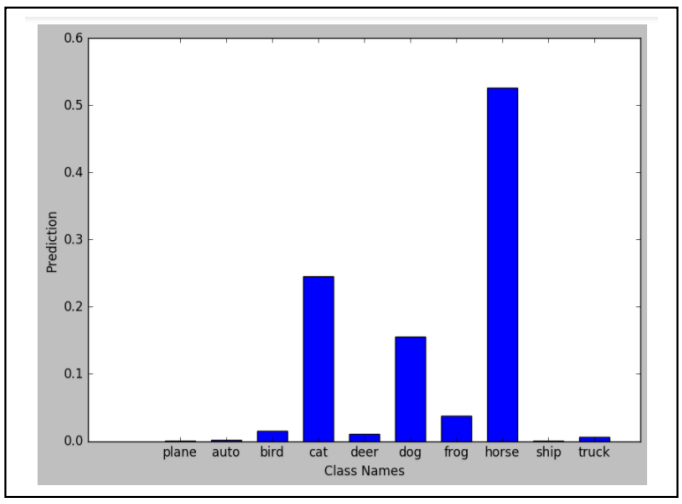

Fig. 20. Bar chart showing the Output Probability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 7 Issue 06, June-2018

V. CONCLUSION

CNNs give the best performance in pattern/image recognition problems and even outperform humans in certain cases.

In this paper a structured Convolutional Neural Network (CNN) was developed, model for supervised learning applied to object recognition. In this model, the CIFAR10 datasets was used as the input images, which must undergo some pre-processing before it is passed through the convnet layer by layer. And this model has proven that with Adams optimizer, the network tends to converge rapidly with gradient descent compare to the popular stochastic gradient descent approach that have been largely used in literature.

After training the network, an accuracy of 98.5% was achieved. The model performed well by giving accurate prediction of some the test images. Though some errors where encountered but that doesn't mean that the network is not accurate, the network is accurate but not 100% accurate.

The task of object recognition has relevant applications in a wide range of domains, and although being successful in many areas of computer vision, including image retrieval and video surveillance. And also, deep learning advanced so fast that is used in many aspects of modern society such as deep learning analytics, named-entity recognition, image recognition and speech recognition.

REFERENCES

[1] M. Yang, "Object Recognition," *Encycl. Database Syst.*, pp. 1936–1939, 2009.
[2] "Artificial Intelligence Overview." .
[3] A. Memo, "Review of machine _ deep learning in an artistic context – Machine Intelligence Report – Medium." 2016.
[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[5] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible , High Performance Convolutional Neural Networks for Image Classification," 2011.
[6] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," pp. 1–27, 2012.
[7] R. Collobert, J. Weston, and M. Karlen, "Natural Language Processing (almost) from Scratch," vol. 1, pp. 1–34, 2000.
[8] C. Couprie, L. Najman, and Y. Lecun, "for Scene Labeling," pp. 1–15, 2012.
[9] W. M. Guzella, Thiago S., Caminhas, "A review of machine learning approaches to Spam filtering _ Walmir Caminhas - Academia." 2009.
[10] A. Karpathy, "Deep Visual-Semantic Alignments for Generating Image Descriptions," 2015.
[11] S. Thrun *et al.*, "Stanley : The Robot that Won the DARPA Grand Challenge," vol. 23, no. April, pp. 661–692, 2006.
[12] Y. Lecun, "The Unreasonable Effectiveness of Deep Learning," 2014.
[13] "Semi-supervised learning - Wikipedia." 2017.
[14] Q. V Le *et al.*, "Building high-level features using large scale unsupervised learning Dataset constructions," pp. 1–10, 2013.
[15] R. Bellman, "THE THEORY OF DYNAMIC PROGRAMMING." 2008.
[16] B. Coppin, *Artificial Intelligence Illuminated*. 2008.
[17] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, vol. 9, no. 2. 1995.
[18] P. Mirowski, "Mc{C}orduck's \emph{{M}achines {W}ho {T}hink} after Twenty-Five Years: Revisiting the Origins of {AI}," *{AI} Mag.*, vol. 24, no. 4, pp. 135–138, 2003.
[19] J. Lighthill, "Lighthill Report: Artificial Intelligence: a paper symposium," *Sci. Res. Counc. London*, 1973.
[20] "Artificial intelligence_wikipedia." 2016.
[21] S. Majumdar, "Deep Columnar Convolutional Neural Network," vol. 145, no. 12, pp. 25–32, 2016.
[22] A. Krizhevsky, V. Nair, and G. Hinton, "CIFAR-10 and CIFAR-100 datasets," *Dataset Accessed From Https://Www.Cs.Toronto.Edu/~Kriz/Cifar.Html*. 2009.
[23] B. S. Hijazi, R. Kumar, C. Rowen, and I. P. Group, "Using Convolutional Neural Networks for Image Recognition," pp. 1–12, 2015.
[24] Ujjwalkarn, "An Intuitive Explanation of Convolutional Neural Networks – thedata science blog." 2016.