Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 10, No. 1 : 2019
ISSN : **1906-9685**

DEEP LEARNING ALGORITHM FOR HUMAN ACTIVITY RECOGNITION

L. RANGA SWAMY, ASSISTANT PROFESSOR, ranar404@gmail.com
N.A.V. PRASAD, ASSISTANT PROFESSOR, navpvlsi@gmail.com
C. RAMAMOHAN, ASSISTANT PROFESSOR, rammohan.mohan966@gmail.com

Department of ECE, Sri Venkateswara Institute of Technology, N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

ABSTRACT: As studies go on in areas like surveillance to identify offenders and lost objects in public spaces, as well as the elderly, the promise of machine learning—and deep learning in particular—becomes more apparent. Even while wearable sensor-based Human Action Recognition (HAR) methods exist, they have the potential to inflict needless psychological and physiological distress on individuals, particularly the young and the old. Automatization is a capability of deep learning.

Index Terms: ***Human Action Recognition (HAR), binary silhouettes, ALEXNET CNN.***

1. INTRODUCTION

Numerous fields may benefit from human behaviour recognition in the actual world. These include smart video surveillance, customer characteristics, and purchasing behaviour analysis. But with all the distractions, occlusions, and different viewpoints out there, it's not easy to accurately identify behaviour. Machines that belong to the class known as "deep learning models" automate the process of learning a feature hierarchy by building higher-level features from lower-level ones.

2. OVERVIEW OF HAR

Investigating the activities shown in video sequences or still pictures is the main goal of human activity identification. This is the driving force behind human activity recognition systems' efforts to accurately categorise incoming data. First, there are simple motions; second, atomic actions; third, interactions between humans or between objects; fourth, collective actions; fifth, behaviours; and sixth, events. The breakdown of the human activities is shown in Figure (1).

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

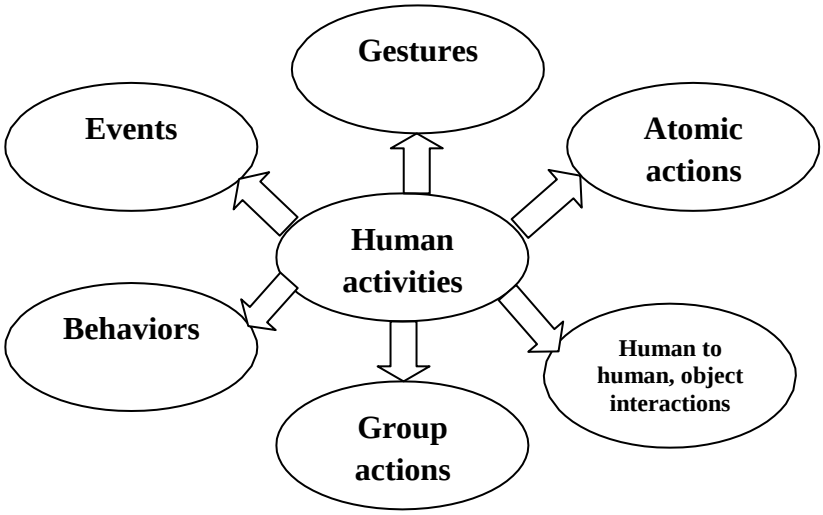


Figure: 1 Human Action Recognitions

3. PROPOSED SYSTEM

Deep Learning

Deep Learning is a sub-field of machine learning that deals with algorithms inspired by brain structure and function. In a word, deep learning accuracy achieves recognition accuracy at higher levels than ever before. This helps consumer electronics fulfill user standards and is important for safety-critical applications such as driverless cars. Recent advancements in deep learning have advanced to the point that deep learning outperforms humans in certain tasks, such as classifying objects in image.

How Deep Learning Works

Step1: The algorithm designer understands the problem and checks whether the deep learning is a good fit or not.

Step2: After understanding the problem he chooses relevant datasets and prepares them for analysis.

Step3: So, there are many deep learning algorithms are there, he chooses the best type of deep learning algorithm that suits to solve the problem.

Step4: Training an algorithm on large amount of labeled data.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

Step5: After training he tests the model against the unlabeled data

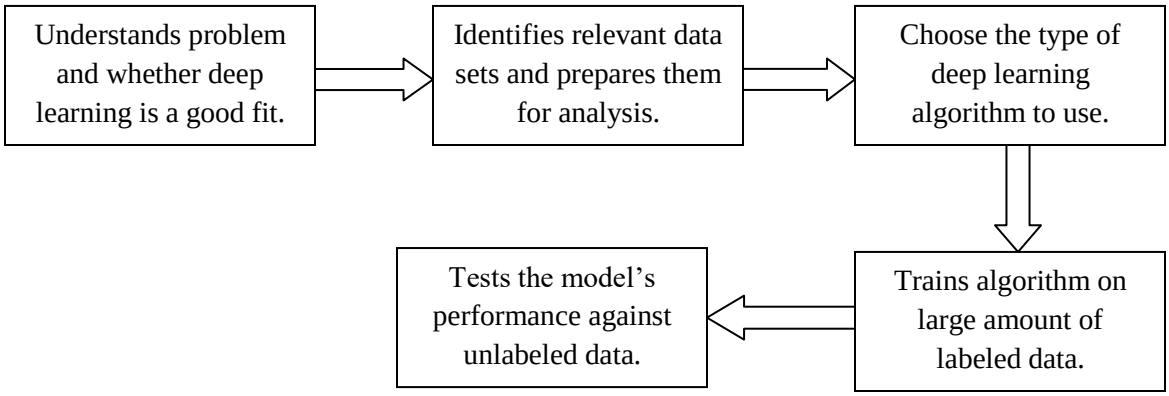


Figure 2: Deep Learning Process

4. DEEP NEURAL NETWORKS

Deep learning models are sometimes called deep neural networks since the majority of deep learning methods employ designs of neural networks. The number of hidden layers in a neural network is usually what the term "deep" means. Deep neural networks may contain up to 150 hidden layers, in contrast to traditional neural networks that only have 2-3. The training of deep learning models eliminates the requirement for human feature extraction by using neural network designs that learn features directly from data and big datasets of labelled data.

In the same way as neurons in the brain are little components, so are nodes in a network. There are some marked and linked nodes and some unmarked ones here. Layers are often used to organise nodes. In order to complete the job, the system has to handle data that is layered between the input and the output. Obtaining a satisfactory outcome requires processing more layers.

A regular neural network is very simple in comparison to a deep neural network. It is capable of a wide range of tasks including investigation, prediction, and creative thinking, including but not limited to: recognising voice instructions, audio and visual recognition, expert assessments, and more. The human brain is unique in that it can think about solutions to problems on a larger scale, make assumptions or draw inferences based on available information, and ultimately achieve the desired result. Even without many highlighted facts, it may solve the issue.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

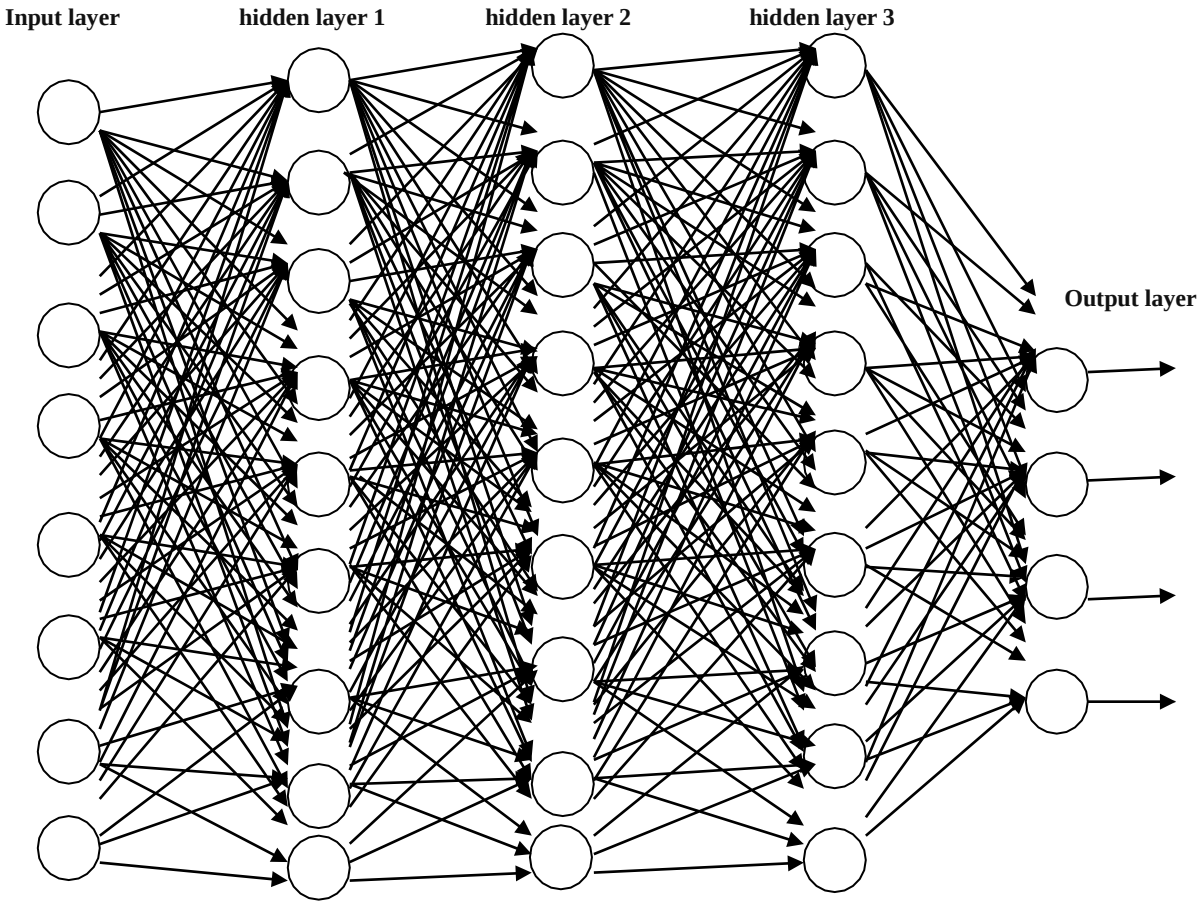


Figure 3: Deep Neural Network

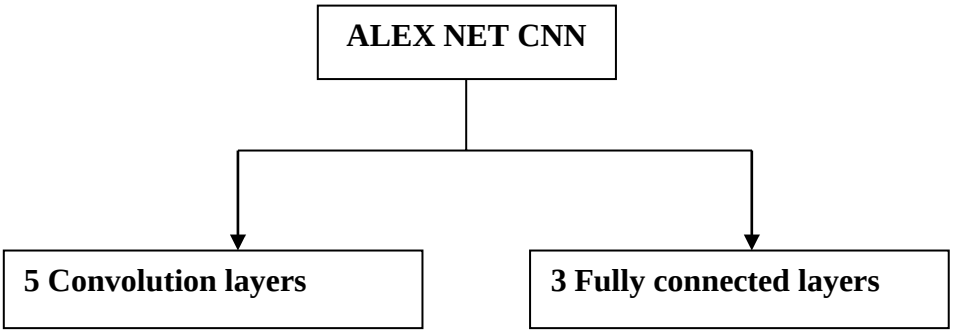
5. ALEX NET CNN ARCHITECTURE

ALEX NET has 8 layers. The first five are convolutionary layers, and the last three are entirely interconnected layers. In between, we also have a few layers called Pooling and Activation. The architecture consists of predefined filters, strings, padding for good object detection. Alex Net is commonly used for object detection tasks. The size of the input picture

227*227*3

Colored image

http://doi.org/10.36893/JNAO.2019.V10N01.013-026



ALEX NET CNN only performs the procedure when the input image has dimensions of 227*227*3. If not we need to reshape our input image. Here we send our input to the first convolution layer after we do the pooling, then the output of the pooling will be supplied to the second convolutionary layer and then again we do the pooling operation.

The second pooling output will be given as input to the third convolutionary layer, where we perform three steps of covolutionary operation after three steps of the third pooling operation. The output of the third pooling layer is given to the first fully connected layer. The third fully connected layer will acts as a softmax function that is used to predict the final output.

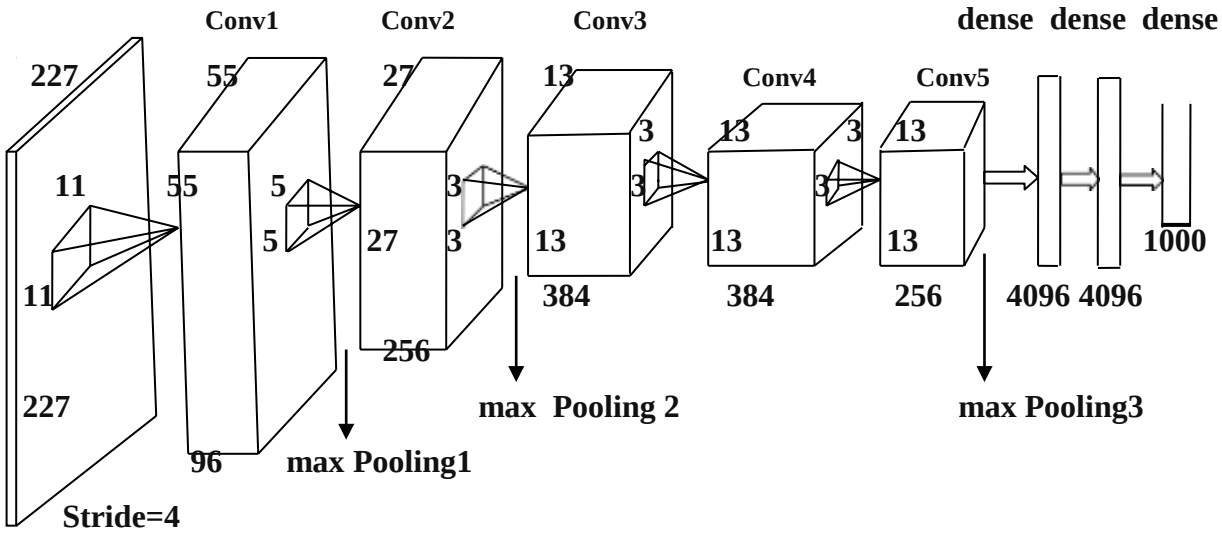


Figure 5: ALEX NET CNN Architecture

So, 227*227*3 is the input of the first convolution layer. 96 filters of size 11*11 with the 4-pixel phase will be added to the first convolution layer. We have a pooling layer after the first convolution layer where we use a window size of 3*3 with the 2 pixel phase. The output of the first convolution layer is given to the second convolution layer as the input.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

(a)

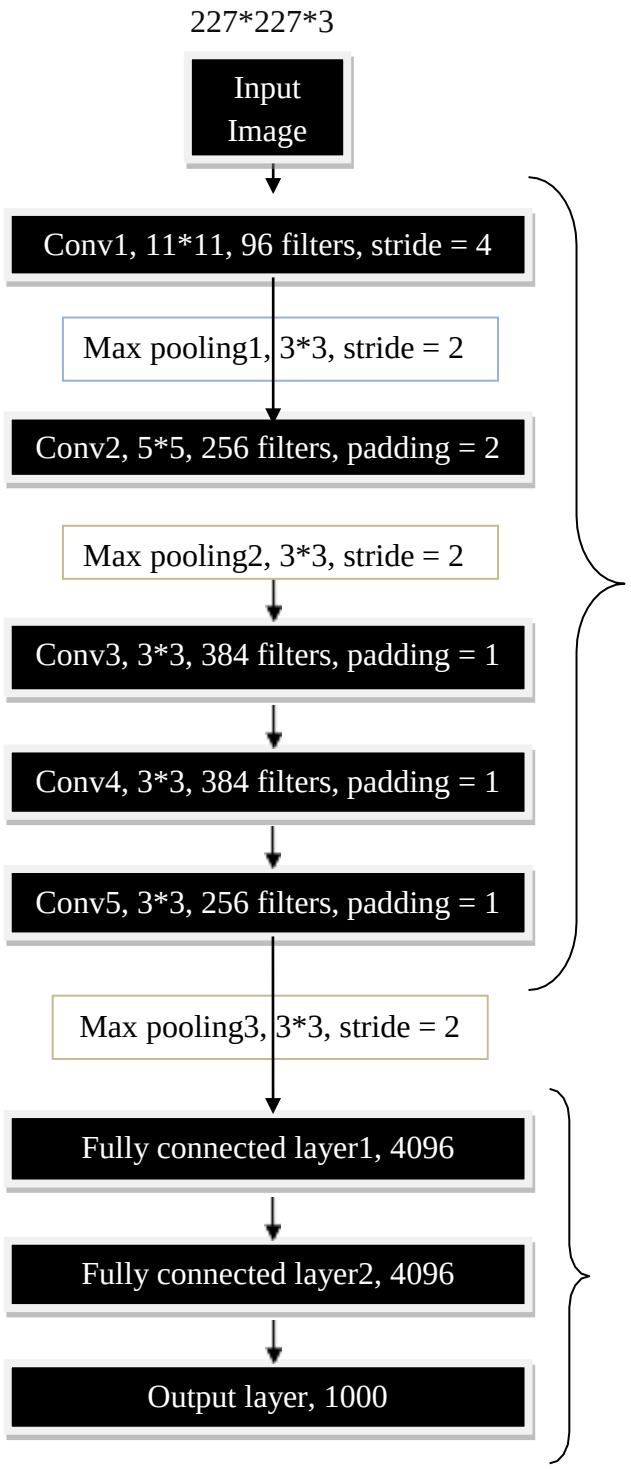


Figure 5)a): Alex net flow chart

With padding 2 and pooling layer 3*3 of step 2, we use 256 filters of size 5*5 in the second convolution layer. The output of the second convolution layer is given to the third

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

convolution layer as the input. Three, four, five layers of convolution are related to each other without any layer of pooling between them. Third convolution layer of 384 scale 3*3 filters with padding 1. Fourth, the same. The third and fourth have the same characteristics. The scale of 256 filters in the fifth convolution layer.

We have a 3*3 size and phase 2 maxpooling after these three layers. After that, we have three completely linked layers, the last one is used for the activation function of softmax that generates a distribution over 1000 class labels. 4096 neurons in the first fully connected layer, 4096 in the second fully connected layer, and 1000 neurons in the last fully connected layer. The neurons in the last fully connected layer rely on the dataset,

Layer		**Feature map**	**Size**	**Kernel size**	**Stride**	**Activation**
Input	Image	1	227*227*3	-	-	-
1	Convolution1	96	55*55*96	11*11	4	Relu
	Maxpooling1	96	27*27*96	3*3	2	Relu
2	Convolution2	256	27*27*256	5*5	1	Relu
	Maxpooling2	256	13*13*256	3*3	2	Relu
3	Convolution3	384	13*13*384	3*3	1	Relu
4	Convolution4	384	13*13*384	3*3	1	Relu
5	Convolution5	256	13*13*256	3*3	1	Relu
	Maxpooling3	256	6*6*256	3*3	2	Relu
6	FC	-	9216	-	-	Relu
7	FC	-	4096	-	-	Relu
8	FC	-	4096	-	-	Relu
9	FC	-	1000	-	-	Relu
						Soft-max

Table 1: Parameters used in Alexnet Cnn

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

Calculation of layers:

Without padding

$$\frac{n-f}{s}+1 * \frac{n-f}{s}+1$$

With padding

$$\frac{n+2p-f}{s}+1 * \frac{n+2p-f}{s}+1$$

Where, n = Image size

f = Filter size

s = Stride

p = padding

Layer 1: Convolution1

$$\frac{n-f}{s}+1 * \frac{n-f}{s}+1$$

$$\frac{227-11}{4}+1 * \frac{227-11}{4}+1$$

55*55*96

Max pooling1

$$\frac{n-f}{s}+1 * \frac{n-f}{s}+1$$

$$\frac{55-3}{2}+1 * \frac{55-3}{2}+1$$

27*27*96

Layer 2: Convolution2

$$\frac{n+2p-f}{s}+1 * \frac{n+2p-f}{s}+1$$

$$\frac{27+2(2)-5}{1}+1 * \frac{27+2(2)-5}{1}+1$$

27*27*256

Max pooling2

$$\frac{n-f}{s}+1 * \frac{n-f}{s}+1$$

$$\frac{27-3}{2}+1 * \frac{27-3}{2}+1$$

13*13*256

Layer 3&4: Convolution 3&4

$$\frac{n+2p-f}{s}+1 * \frac{n+2p-f}{s}+1$$

$$\frac{13+2(1)-3}{1}+1 * \frac{13+2(1)-3}{1}$$

Layer 5: Convolution5

$$\frac{n+2p-f}{s}+1 * \frac{n+2p-f}{s}+1$$

$$\frac{13+2(1)-3}{1}+1 * \frac{13+2(1)-3}{1}+1$$

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

13*13*384 **13*13*256**

Max pooling3 **Fully connected layer1**

$\frac{n-f}{s}+1 * \frac{n-f}{s}+1$ $6 * 6 * 256$

$\frac{13-3}{2}+1 * \frac{13-3}{2}+1$ **9216**

6*6*256

Convolutional layer:

To construct a feature map that summarizes the presence of detected features in the data, it applies a filter to an input. One image becomes a stack of filtered images in the convolutional layer, and the number of filtered images depends on the number of filters.

Input image ***** **Filter** **=** **Filtered image**

0	1	2
3	4	5
6	7	8

0	1
2	3

19	25
37	43

Pooling layer:

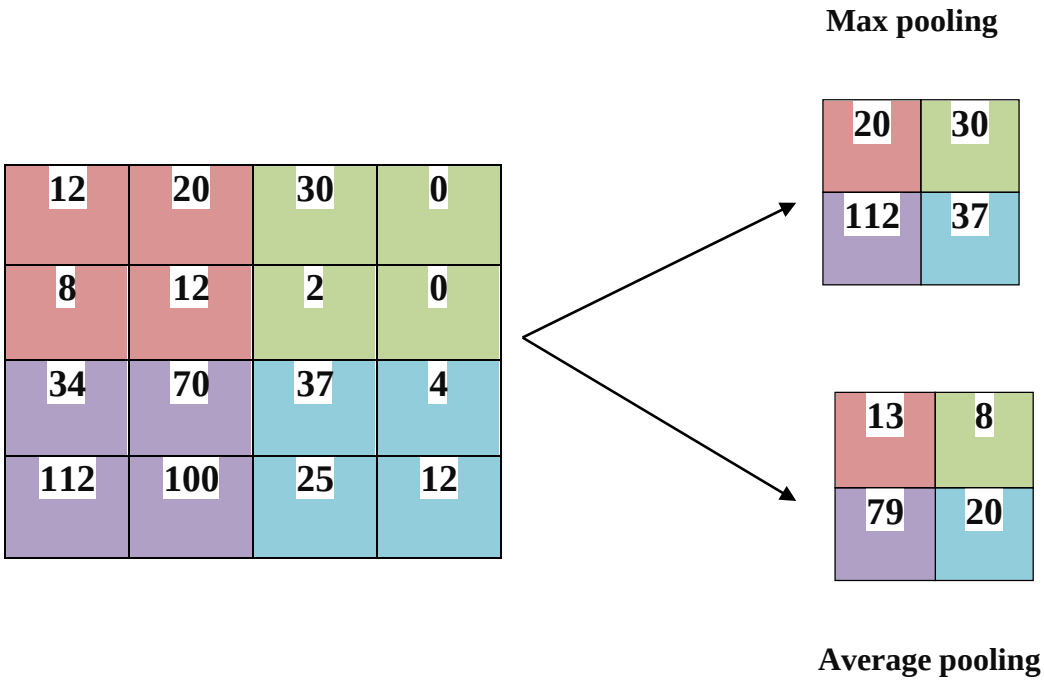
Pooling layer down samples the volume spatially, independently in each depth slice of the input volume. The most common down sampling operation is max, giving rise to max Pooling.

Max pooling with 2*2 filter and stride4

There are two types of pooling. They are:

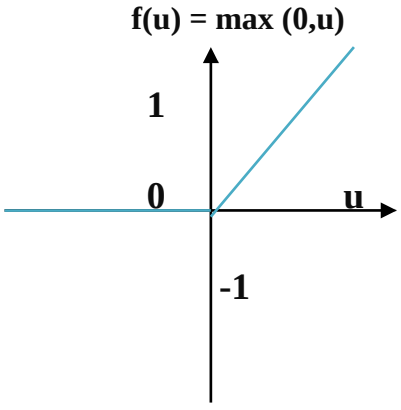
1. Max pooling: From above example in a 2*2 window we choose max value.The process called max pooling.

2. Average pooling: From above example we take the average of 2*2 window. The process called average pooling.



Rectified Linear unit（ReLU):

- Activation function of ReLU produces 0 when u < 0, and is linear with slope 1 when u > 0. Rectified linear function, f(u) = max(0,u)



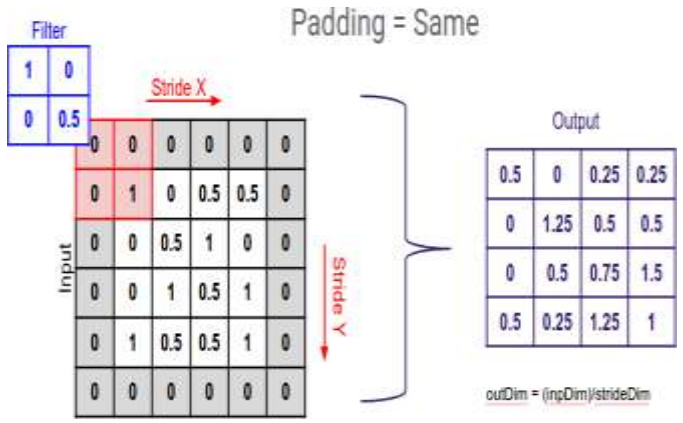
Fully connected layer:

- This is the layer where image classification actually happens and we convert our filtered images into a 1-Dimensional array.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

Padding and Stride:

- ❖ Adding zero rows and columns to the image is known as padding.
- ❖ Number of columns and rows are shifting towards right and downside is known as stride.



Soft-max function:

The soft-max function is applied after the output layer of ALEX NET CNN in order to obtain the probability of the possible actions.

$$\sigma(z)_j = e^{z_j} / \sum_{k=1}^{N} e^{-z_k}$$

Where, j = each action

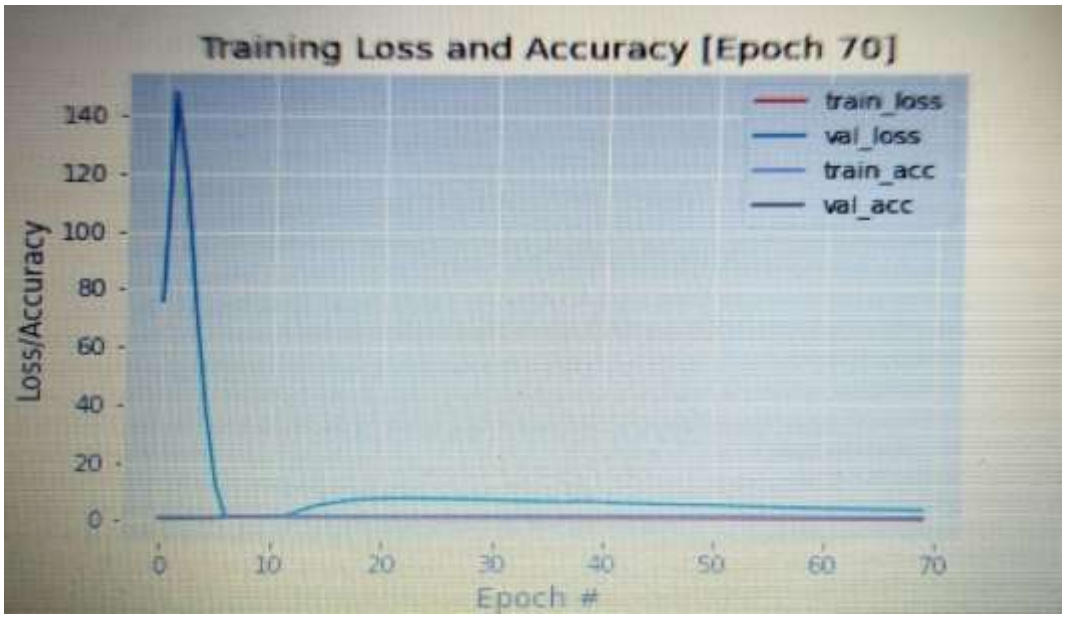
Z = network output

N = Total number of actions

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

6. SIMULATION RESULTS AND ANALYSIS

Accuracy scores ranging from 0.7 to 0.95 are often achieved by using picture variations in the datasets used by Convolutional Neural Networks. However, in order to attain greater accuracy levels (0.3 - 0.55), we have chosen a dataset with more picture similarities. Some preprocessing adjustments, such as picture scaling and cropping, were performed to emphasise the Region of Interest (ROI), and adjustments were made to the algorithm to get a greater accuracy of 0.63.



Due to false positives identified in the algorithm false higher accuracy is achieved till 1 to 12 epochs then as the number of epochs increases false positives are minimized and true positives are identified.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026

7. CONCLUSION

We suggest creating a consumer electronics device that can automatically monitor and identify the everyday activities of older persons living alone. The system should have cheap computing cost and provide high accuracy outputs. Additionally, the system's quick processing time makes it very promising for real-time applications, and it can be used regardless of ambient circumstances or domain architecture. This method has successfully addressed the issues of view-variance (single camera) and intra-class variation. Both the CAD-60 daily activity datasets and the experimental findings demonstrate that the suggested approach outperforms existing state-of-the-art systems. The goal of this project was to create a low-cost, highly accurate human action recognition system for use in consumer electronics.

8. REFERENCES

[1] Cho Nilar Phyo, T. T. Zin and P. Tin, "Deep Learning for Recognizing Human Activities using Motions of Skeletal joints" DOI 10.1109/TCE.2019,IEEE transactions on consumer electronics.

[2]C. N. Phyo, T. T. Zin and P. Tin, "Skeleton motion history based human action recognition using deep learning", in Proc. of 2017 IEEE 6th Global Conf. on Consumer Electronics, Nagoya, Japan, 24-27 Oct. 2017, pp. 784-785.

[3]J. Wang et al., "An enhanced fall detection system for elderly person monitoring usingconsumer home networks", IEEE Trans. Consumer Electronics, vol. 60, no. 1,pp.23- 29,Apr.2014, 10.1109/TCE.2014.6780921.

[4] A. Jalal et al., "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home", IEEE Trans. Consumer Electronics, vol. 58, no. 3, pp. 863-871, Sept. 2012, 10.1109/TCE.2012.6311329.

[5] T. T. Zin, P. Tin and H. Hama, "Visual monitoring system for elderly people daily living activity analysis", in Proc. of the Int. MultiConf. of Engineers and Computer Scientists 2017, Hong Kong, 15-17 Mar. 2017, pp. 140-142.

[6] L. Zaineb et al., "A Markovian-based approach for daily living activities recognition", in Proc.

of the 5th Int. Conf. on Sensor Networks, Rome, Italy, 17-19 Feb. 2016, pp. 214-219.

[7] L. H. Wang et al., "An outdoor intelligent healthcare monitoring device for the elderly", IEEE Trans. Consumer Electronics, vol. 62, no. 2, pp. 128-135, Jul. 2016, 10.1109/TCE.2016.7514671.

[8] J. Wang et al., "An enhanced fall detection system for elderly person monitoring using consumer home networks", IEEE Trans. Consumer Electronics, vol. 60, no. 1, pp. 23-29, Apr. 2014, 10.1109/TCE.2014.6780921.

http://doi.org/10.36893/JNAO.2019.V10N01.013-026