Sports Video Classification with Deep Convolution Neural Network: A test on UCF101 Dataset

M. Ramesh, K. Mahesh

***Abstract**: In the present era, Deep Learning has been applied on a variety of problems from image processing to speech recognition. Convolution Neural Network (CNN) has been extensively used as a powerful classification model for image recognition problems. Video classification presents unique challenges but the problem related to video data is similar to image classification or an object detection problem. The main purpose of video classification in sports is to help the viewers to find the video of their own interest for training and improve the performance. The proposed work is a preliminary attempt to evaluate the performance of deep convolution neural network architectures on the ordered sequence of frames of the sports video. Video classification and video content analysis is one of the ongoing research areas in the field of computer vision. The classification of each frames are recorded and the majority vote of the frames are used to classify the video. UCF101 Video action database has been used for the classification problem.*

***Keywords:** Convolutional Neural Network, Deep Learning, Video Classification, Sports Video, Video content Analysis.*

I. INTRODUCTION

A. Video and Video Classification:

Either directly or indirectly video has the important role in human’s day to day life. Video is sometimes called as sequence of images played in specified time interval. Images are sometimes referred to as frames. Video classification is one of the very challenging problems among computer vision researcher and also not so easy to solve because the shared actions that appear in the video are the major issue. The objective of this kind of research is to develop intelligent systems of video classification. The main purpose of video classification in sports is to assist the listeners to locate the video of their own concern for training and improve the performance of the players [7]. The following example demonstrates the working principles of video classification system. A high jump sport video consists of two different actions. One is running and one more is high jump. The same is shared with other videos, such as running, long jump or hurdling sports video [4].

A human can easily classify the correct event of the video only with one or two specific frames. That shows the distinctive action of the event. The same hypothesis can also work for video classification system by eliminating some of the frames and extracting few selected frames in the video and then apply classification process [4]. Selected frames among the extracted frames are also called as key frames.

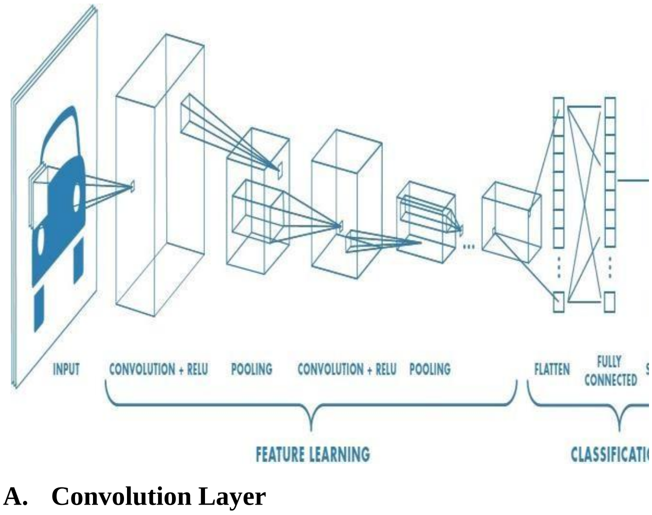
II. CNN ARCHITECTURE

ANN Artificial Neural Networks are used for various classification technique like text, audio, image and sometimes video. The followings are the various types of neural networks currently being used in ML (Machine Language). They are:

1. Feedforward Neural Network.
2. Radial Basis Function Neural Network.
3. Kohonen Self Organizing Neural Network.
4. Recurrent Neural Network (RNN).
5. Convolutional Neural Network (CNN).
6. Modular Neural Network (MNN).

FNN is mainly used for computer vision and speech recognition, RBF Neural Network is used predicting the sequence of words and text mining, similarly Image recognition, Object detection, Image classification and speech recognition are done through CNN. In this paper video is taken as input this can be converted in to frames, that is images. CNN takes input as an image, process and classify it under certain label (category).

The following diagram describes the absolute flow of CNN process. It takes input as an image and classifies based on some set of features. In general CNN is to train input image and test each input image, then pass it through series of convent layers with various filter size, followed by ReLu layer, pooling, fully connected layers and then apply softmax function to classify an image and video.



A. Convolution Layer

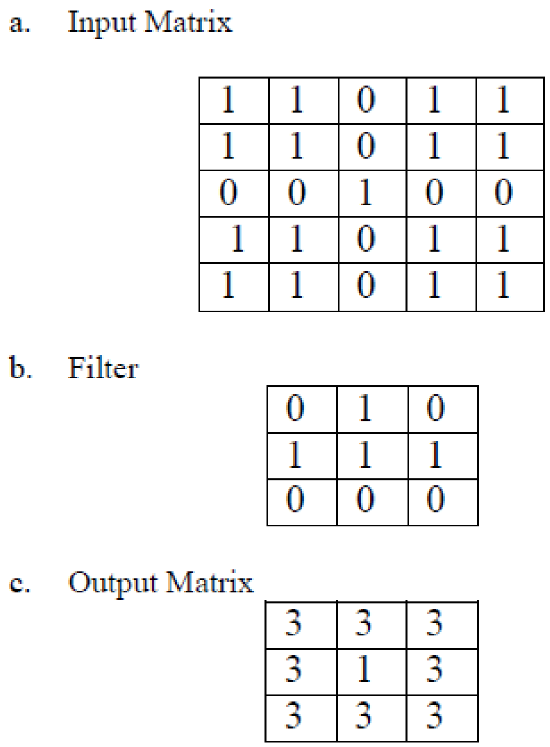
The initial component of a CNN is the convolution. The main function of this component is feature extractor of images [1]. The objective of convolution layer is to learn features of the inputs [2]. Convolution layer accepting two inputs image matrix (volume) of dimension, a filter and produce output matrix (volume).

Revised Manuscript Received on April 25, 2019.

***** Correspondence Author
M. Ramesh*, Department of Computer Science, Faculty of Science and Humanities, SRM IST, Kattankulathur, Chennai. India E-mail: ramcsit@gmail.com.
Dr. K. Mahesh, Professor, Department of Computer Applications, Alagappa University, Karaikudi, India. E-mail: mahesh.alagappa@gmail.com.

Input: An Image matrix (I_h X I_W X D) Filter (F_h x F_W x D)
Output: Output matrix (I_h-F_h+1) x (I_W – F_W +1) x 1

Where Ih is the input image height, Iw is the input image width, D is depth of the image, where Fh is the filter height, Fw is the filter width and D is depth of the filter. Consider a 5 x 5 whose image pixel values are either 0 or 1 and filter matrix 3 x 3 as shown in figure 2. Then the convolution of 5 x 5 image matrix multiplies with 3 x 3 filter matrix which is called feature map. This is calculated based on the strides. Stride is nothing but the number of pixels shifts over the input image matrix. Move the filter to one pixel at a time when the stride is one; similarly move the filter to two pixels at a time when the stride is two and so on. When the filter does not fit with the input image then implement padding.



a. Input Matrix

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

b. Filter

0	1	0
1	1	1
0	0	0

c. Output Matrix

3	3	3
3	1	3
3	3	3

Figure 2: Input Image matrix multiplies with filter and its output matrix.

B. ReLU Layer

ReLU stands for Rectified Linear Unit. The main aim of ReLu layer is to apply an activation function. For example max(0,x) function, for a non-linear operation. This type of operation does not affect or change the size of the volume [1].

C. Pooling Layer

The main purpose of pooling is down sampling in order to reduce the complexity for further layers [3]. It reduces the number of parameters. Parameters are reduced when the input images are too hefty. Down sampling is sometimes called as sub sampling (spatial pooling). This may reduces the dimensionality of each map and retains the significant information. Max pooling, average pooling, and sum pooling are three various types of spatial pooling [3].

III. PROBLEM STATEMENT

Video is an admirable tool for delivering content; it has one of the major roles in human daily life [7]. The main purpose of video classification in sports is to assist the viewers to locate the video of their own concern for training and improve the performance. Video classification and video content analysis is one of the ongoing research areas in the field of computer vision. Analyzing and classification are based on its contents on Video data in many applications [8]. Proposed work is a preliminary attempt to evaluate the performance of deep convolution neural network architectures on the ordered sequence of frames of the sports video. UCF101 Video action database has been used for the classification problem. The output of the proposed work is producing the class label of the video with lofty accuracy.

IV. METHODOLOGY

A. UCF101 Dataset

UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. This data set is extended from UCF50 data set. UCF50 has 50 actions of categories. In UCF101 data set has 13320 videos from 101 action categories. UCF101 gives the largest variety of sports in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc, it is the most challenging data set to date. The main goal of UCF101 is to encourage further research into action recognition by learning and exploring new realistic action categories.

B. Implementation

The following are the steps performed to implement deep convolution neural network on UCF101 sports dataset for video classification using MATLAB.

i. Load video dataset (UCF101Sports).
ii. Convert the video into frames.
iii. Find the number of frames in a video.
iv. Read the frames (images) one by one.
v. Resize the image with respect to [404,720].
vi. Store in separate folder.
vii. Provide resized image into convent.
viii. Apply filtering techniques with different strides.
ix. Apply padding when it is needed.
x. Do convolution on the image and apply ReLU activation to the matrix.
xi. Perform pooling to reduce dimensionality size.
xii. Add necessary CNN layers until satisfied with test image.
xiii. Flatten the output and feed into fully connected layer.
xiv. Output the class label using an activation function and classify video.

C. Experimental Setup

CNN Network Design Parameters: Hardware Resource : Single CPU
No. of Classes : 5
{0-Golf,1-diving,2-football,3-Running..} Image Size : 404 X 720
Filter Size : 5

No. of Filters : 20 Stride : 2

Padding : 1

No. of Convolutional2d Layer (Nested): 2

Output Size in Fully Connected Layer: 5 (No. of Classes) Training Options:

Optimizer : Stochastic Gradient Descent with Momentum (SGDM)

Epochs : 15

MiniBatch Size : 64 Initial Learning Rate: 0.01

D. Results

(i) Experimental Results with MiniBatch Size= 15 Training on a Single CPU

Epoch	Iteration	Time Elapsed (Seconds)	MiniBatch Loss	Mini Batch Accuracy	Base Learning Rate
1	1	75.49	3.5458	30.30%	1.00e-04
15	15	1089.15	0.0005	100.00%	1.00e-04

(i) Experimental Results with MiniBatch Size=64 Training on a Single CPU

Epoch	Iteration	Time Elapsed (Seconds)	MiniBatch Loss	Mini Batch Accuracy	Base Learning Rate
1	1	9.38	0.9782	25.00%	1.00e-04
7	50	450.50	-0.0000	100.00%	1.00e-04
13	100	898.47	-0.0000	100.00%	1.00e-04
15	120	1077.69	-0.0000	100.00%	1.00e-04

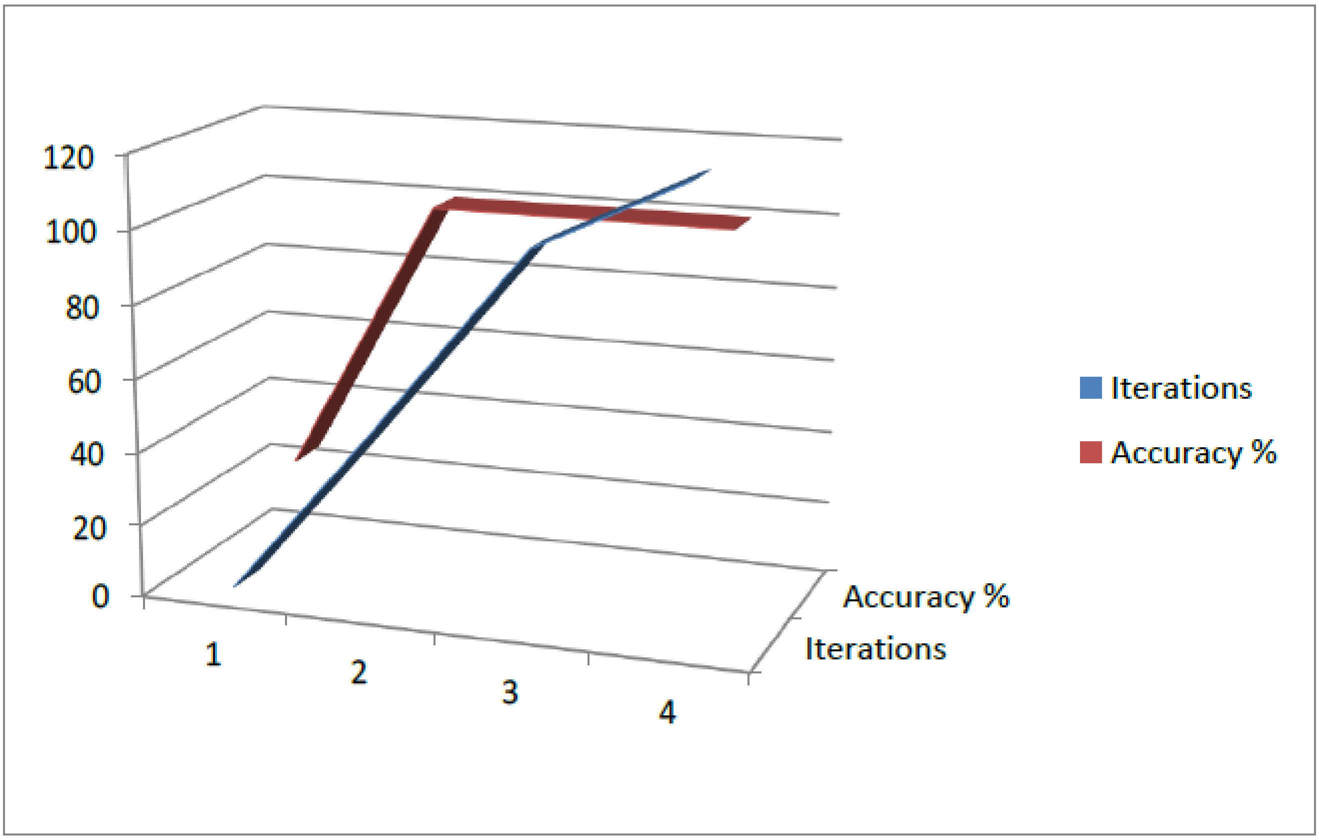


Figure 3: Epochs Vs Accuracy Rate

E. Discussion

The results show that the Convnet trains itself and it is capable of achieving high accuracy rates. The depth of the network is very important because if any one of the layer is removed, it leads to significant performance degradation in the network. So, depth is really significant for achieving the results.

V. CONCLUSION AND FUTURE DIRECTION

In this paper, a simple convolutional neural network on Video Classification has been made. This network gives less computational cost. This is a very shallow network that gives good accuracy rate. Experiments by varying learning rate and Epochs have great influence on the networks performance. The temporal features of Video Sequences are important rather than the features in static images and this can be extended in the future work. In future our system will be added with different stride, padding, and more numbers of convolution layers to different optimizer and solve the classification problem in order to increase accuracy and performance.

REFERENCES

1. Nadia Jmour, Sehla Zayen, Afef Abdelkrim, Convolutional Neural Network for image classification, 2018, IEEE.
2. Tianmei Guo, Jiwen Dong, Henjian Li, Yunxing Gao, Simple Convolutional Neural Network on Image Classification, in 2017 IEEE 2nd International Conference on Big Data Analysis.
3. Saad ALBA WI, Saad AL-ZAWI, Understanding of a Convolutional Neural Network, in ICET 2017, Antalya, Turkey.
4. Rexa Fuad Rachmadi, Keiichi Uchimura, and Gou Koutaki, Video Classification using compacted dataset based on selected keyframe, in 2016 IEEE.
5. Darin Brezeale and Diane J. Cook , Fellow, IEEE, Automatic Video Classification: A Survey of the Literature, IEEE Transactions on System, Man, and Cybernetics, Applications and Reviews, vol. 38, no. 3, May 2008.
6. M.Ramesh, K.Mahesh, Multidimentional View of Automatic Video Classification: An Elucidation, in IJCSE, vol. 6, Special Isse-4, may 2018.
7. M.Ramesh, K.Mahesh, A Preliminary Investigation on a Novel Approach for Efficient and Effective Video Classification Model, in ICCS18, Loyola College, Chennai, Dec 2018.
8. Dong-Chul park, Classification of Video Data Using Centroid Neural Network with Bhattacharyya Kernel, in 2009, International Conference on Electronic Computer Technology.
9. S.Maheshwari, P. Arockia Jansi Rani, “Human Action Recognition System Based on Silhouette”, International Journal of Computer and Information Engineering, VOl. 9, No:110, 2015.
10. M. Ramesh1*, K. Mahesh2, ,” Significance of various Video Classification Techniques and Methods: A Retrospective”,International Journal of Pure and Applied Mathematics, Volume 118 No. 8 2018, 523-526.
11. Nirav Bhatt, - “A survey on video classification techniques‖, International Journal of Emerging Technologies and Innovative Research”, March 2015, Volume 2, Issue 3.

AUTHORS PROFILE

Mr. M. Ramesh pursed Bachelor of Science in Computer Science, Master of Science in Computer Science and Information Technology and M.Phil in Computer Science from Madurai Kamaraj University, Tamil Nadu, India. He is currently pursuing Ph.D (part time) in Algappa University and currently working as Assistant Professor in Depatment of Computer Science, Faculty of Science and Humanities, SRM IST, Chennai since 2011. He has more than 11 years of teaching experience.

Dr. K Mahesh pursed Master of Computer Applications, M.Phil in Computer Science and Ph.D in Computer Science. He is currently working as Professor in Department of Computer Applications in Alagappa University, Tamil Nadu, India. He has published 45 International journals, 9 International Conference papers, 3 National Journals and 23 National Conferences. He has completed funded research project titled “Collaborative Directed Basic Research in Smart and Secure Environment” from july 2007 to august 2012. He is a member of International Association of Engineers (IAENG). He is Reviewer, ICTACT Journal on Image and Video Processing (IJIVP) Publisher: ICT Academy of Tamil Nadu and also Reviewer, International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Publisher: ACE Publishers. He has Presented Key Note Address in the National Seminar on Current Trends in Computing Technologies organized by the PG Department of Computer Science, Government Arts College for women, Ramanathapuram, Feb 28, 2015 and he aslo presented Presented Key Note Address in the Intercollegiate Meet (TECHNO’15) organized by the PG Department of Computer Science, Idhaya College For Women, Sarugani, Sep 9th ,2015. He has 26 years of teaching experience and 9 years of Research Experience.