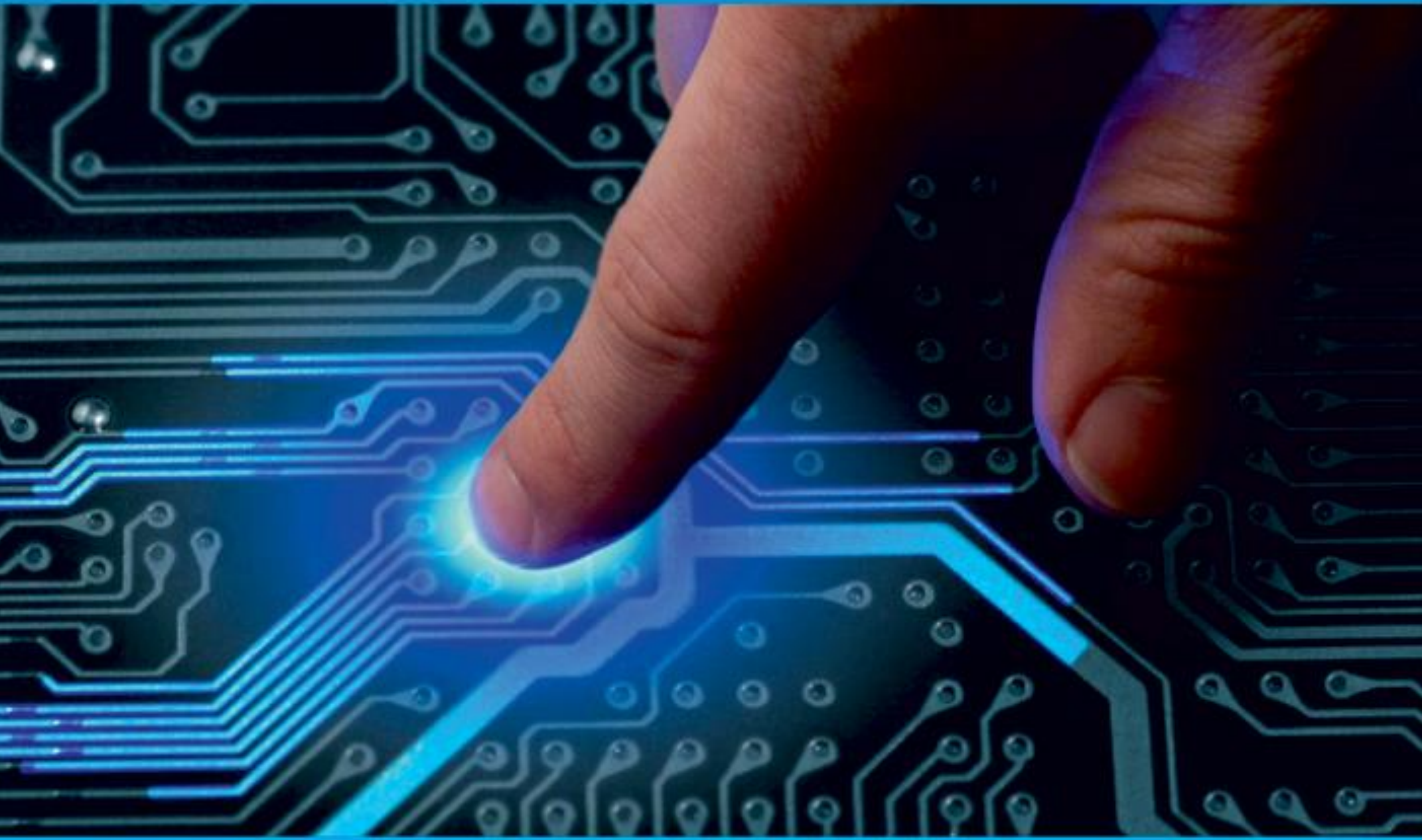
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205127 |

Video Classification a Deep Learning Approach

Mogali Madhu Babu[1], Seepana Ratna Kumari[2], Tappita Sai Appala Surendra Kumar[3], Kutcharlapati Sravya[4], Potnuru Venkat Dinesh[5]

Associate Professor, Department of CSE, Satya Institute of Technology and Management, Vizianagaram, India[1,2]

B. Tech Student, Department of CSE, Satya Institute of Technology and Management, Vizianagaram, India[3,4,5]

ABSTRACT: Nowadays, many platforms are available for uploading photos and videos, making it increasingly necessary to categorize this media content. Deep neural networks, process inputs and provide outputs to address real-time problems like classification. Deep neural networks are frequently used in reinforcement learning and supervised learning situations. Video classification is essential for supervised learning tasks such as hand gestures and human activity recognition.

Video classification is the process of inferring a video's genre from its frames, such as whether it's a dance, yoga, or singing video. To train models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, this process depends on extracting features from the video frames. These models are used frame by frame to increase the video stream's classification accuracy. A high-quality video classifier describes the entire video and correctly labels individual frames.

KEY WORDS: Deep neural networks, Convolutional neural network, Recurrent neural networks, Video classification, LSTM.

I. INTRODUCTION

The recent years, a lot of information has been shared over the Internet using multimedia files, which include text, audio, images, and videos. Videos are one type of them that has a lot of information. Videos are more widely used in many applications, including hand gestures and human action recognition, acknowledgment, etc. For the videos to be used and managed properly, they must be categorized. A video is made up of an organized series of frames. Both temporal and spatial information are present in every frame. While temporal information is related to frames that deal with time, spatial information is extracted from frames. Videos cannot be processed using a fixed-size architecture due to their temporal properties, which are connected time elements. CNN and RNN architectures are used in video classification to extract features and forecast accurate class labels. A type of deep learning technique called convolutional neural networks (CNNs) is used to extract data from videos. Utilizing the sequences, train an LSTM network to anticipate the video labels. Recurrent neural networks (RNNs) and CNNs work together to create a potent architecture for video categorization issues because they can process spatial and temporal data efficiently and concurrently.

II. LITERATURE SURVEY

This paper uses a hybrid deep learning model to present a novel method for human action recognition. It uses a General Regression Neural Network (GRNN) to gather features from each frame and predict human actions. It combines Gaussian Mixture Models (GMM) and Kalman Filter (KF) methods to detect motion. This method's main benefit is that it can extract features based on time from each frame, enabling a thorough analysis. The method achieves high accuracies of 96.3%, 89.01%, and 89.30% on datasets like UCF sports, UCF101, and KTH, where it is evaluated.

- The study focuses on sports video classification using networks such as GoogLeNet and AlexNet. The main objective is to make it easier for athletes to locate pertinent training videos so they can perform better. On the UCF101 dataset, a customized convolutional neural network (CNN) is used to compare pre-trained neural networks. Pre-trained networks perform better, according to the results, with GoogLeNet achieving 91.67% accuracy and AlexNet achieving 92.67% accuracy in video classification.
- To increase the accuracy of human action recognition, a Convolutional Long Short-Term Memory (ConvLSTM) network based on attention mechanisms is presented. ConvLSTM is used by the framework to recognize actions

based on sequential information in videos, while GoogleNet is used for feature extraction. The model is used as an end-to-end video classification solution on the UCF-101, UCF-11, and HMDB-51 datasets.

- The study uses datasets classified into single viewpoint, multiple viewpoints, depth, and RGB datasets to address difficulties in human action recognition from video data. Group activities, object interactions, gestures, and actions are the different categories under which activities fall. To reliably detect videos, researchers have used 3D convolution operations for single and multiple viewpoint approaches.
- With a high accuracy of 94%, a hybrid VGG-GRU architecture is proposed to classify ten distinct football actions. Using a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), the model takes into account the difficulties presented by overlapping actions and uncontrolled video-capturing conditions in football matches.

III. VIDEO CLASSIFICATION: VIDEO CLASSIFICATION USING DEEP LEARNING

Problem Statement:

Develop a deep learning model for accurate classification of actions in videos, utilizing techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short Term Memory (LSTM) to handle spatial and temporal features.

MODELS THAT CAN BE USED FOR THE PROJECT

Convolutional Neural Network:

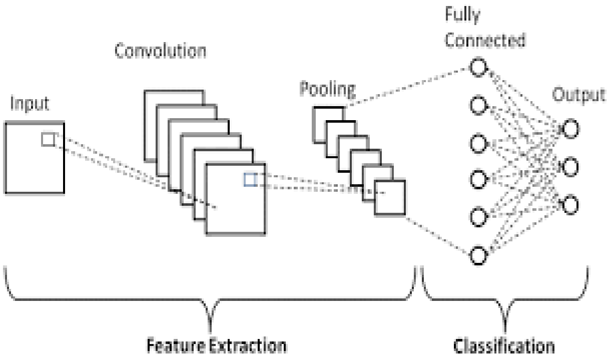
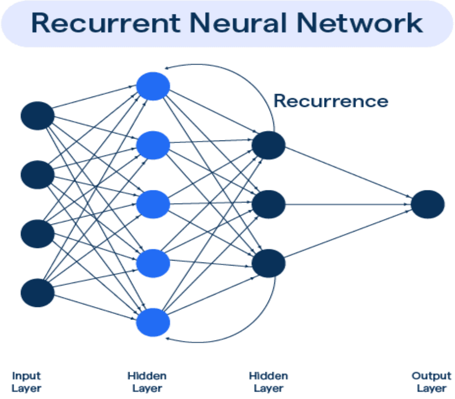


Fig 3.1. Extracting the Features

A frame will be fed into the CNN-LSTM model's first convolution layer.

Its output will then be subjected to the maximum pooling layer, which will enable it to pool into a smaller dimension. After that, it is fed into an LSTM layer. The output layer is the last component, and it will classify videos into different groups and evaluate how accurately the model performs. This model is based on the idea that the convolution layer will extract local features. The LSTM layer is used to determine the relationship between the frames in the sequence. These linkages will help with the videos' classification.



3.2 Classifying the output

- **Recurrent Neural Network:**
  One kind of artificial neural network that makes use of sequential or time series data is the recurrent neural network (RNN). These deep learning algorithms are integrated into well-known programs like Siri, Voice Search, and Google Translate. They are frequently utilized for ordinal or temporal problems like language translation, natural language processing (NLP), speech recognition, and picture captioning. Recurrent neural networks (RNNs), like feedforward and convolutional neural networks (CNNs), learn from training data. Their ability to use information from previous inputs to affect the current input and output sets them apart. Recurrent neural networks rely on the previous elements in the sequence to determine their output, in contrast to traditional deep neural networks, which assume that inputs and outputs are independent of one another.
- **Long Short-Term Memory:**
  One kind of RNN, or recurrent neural network, that can preserve long-term dependencies in sequential data is the LSTM or long short-term memory. Text, speech, and time series are examples of sequential data that LSTMs can process and analyze. They avoid the vanishing gradient issue that besets conventional RNNs by controlling the information flow using gates and a memory cell, which enables them to maintain or discard information as needed. LSTMs are extensively employed in many different applications, including time series forecasting, speech recognition, and natural language processing.

IV. METHODOLOGY

We begin with importing necessary libraries. Next, we take frames from the input and normalize them and then we create a data set based on the labels and the features then we give that dataset to a convolution neural network which is used for extracting the feature by applying kernels and max pooling to extract the features from the dataset and then we save the trained model. Then we gave test data to our saved model so that it could predict the output.

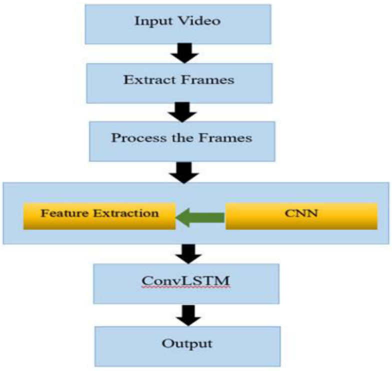


Fig 4.1. Methodology

These include:

- We use categorical to handle categorical data.
- CNN for extracting the features and labels from the video.
- RNN and LSTM for the model predicting the output labels
- Train-test split for model validation.
- Evaluate method from CNN is used to evaluate the model performance

V. RESULT

After Evaluation, the accuracy of our model is 89%.

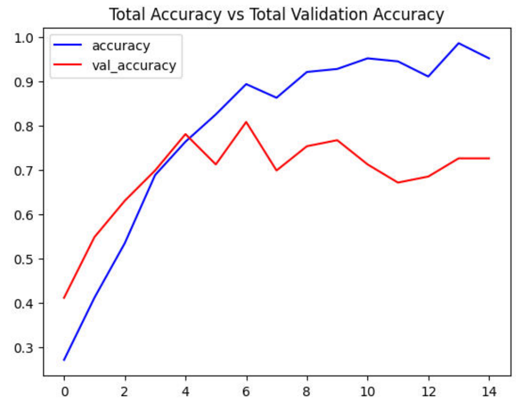


Fig 5.1 Training vs Validation Accuracy

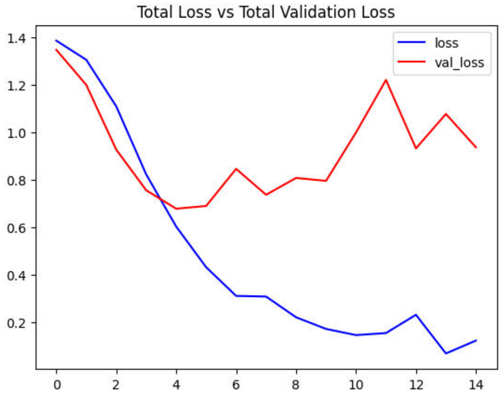
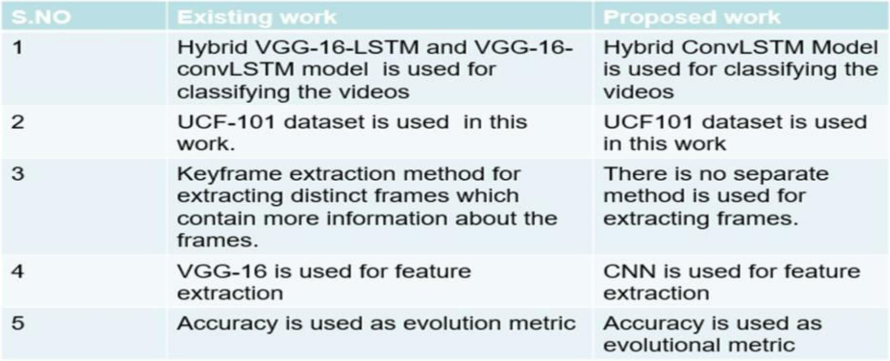


Fig 5.2 Total loss vs Validation loss



S.NO	Existing work	Proposed work
1	Hybrid VGG-16-LSTM and VGG-16-convLSTM model is used for classifying the videos	Hybrid ConvLSTM Model is used for classifying the videos
2	UCF-101 dataset is used in this work.	UCF101 dataset is used in this work
3	Keyframe extraction method for extracting distinct frames which contain more information about the frames.	There is no separate method is used for extracting frames.
4	VGG-16 is used for feature extraction	CNN is used for feature extraction
5	Accuracy is used as evolution metric	Accuracy is used as evolutional metric

Fig 5.3 Result

The Accuracy score of our model is equal to **0.8997957559744329**

VI. CONCLUSION

The advantage of using ConvLSTM for video classification is that it's an effective method for deciphering and interpreting videos. A particular kind of neural network called ConvLSTM can recognize spatiotemporal patterns in videos and forecast the actions that are performed within. In addition to taking into account the spatial information in each frame, the model can identify both short- and long-term temporal dependencies in the video frames. This method has been used for several video classification applications, such as sentiment analysis, action recognition, and surveillance systems.

This work involves some challenges.

- Selecting the dataset
- Complexity of computation
- Optimizing hyperparameters

VII. FUTURE WORK

Enhancing the architecture of the model by incorporating an attention mechanism. Transfer learning, which can modify pre-trained models for tasks involving video classification. To increase the model's accuracy, multimodal learning can be used to combine data from various sources, such as audio and visual data.

REFERENCES

1. Muhammad, K., Ullah, A., Imran, A. S., Sajjad, M., Kiran, M. S., Sannino, G., & de Albuquerque, V. H. C. (2021). Human action recognition using attention-based LSTM network with dilated CNN features. Future Generation Computer Systems, 125, 820-830.
2. Wang, Jinzhuo, Wenmin Wang, and Wen Gao. "Multiscale deep alternative neural network for large-scale video classification." IEEE Transactions on Multimedia 20, no. 10 (2018): 2578-2592.
3. Tran, D., Wang, H., Torresani, L., & Feiszli, M. (2019). Video classification with channel-separated convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5552-5561).
4. Pandeya, Y. R., & Lee, J. (2021). Deep learning-based late fusion of multimodal information for emotion classification of the music video. Multimedia Tools andApplications, 80(2), 2887-2905.
5. Ge, H., Yan, Z., Yu, W., & Sun, L. (2019). An attention mechanism-based convolutional LSTM network for video action recognition. Multimedia Tools and Applications, 78(14), 20533-20556.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details