iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43929

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Soccer Action Video Classification using Deep Learning

Omprakash Yadav[1], Rachael Dsouza[2], Rhea Dsouza[3] Janice Jose[4]
[1, 2, 3, 4]*Department of Computer Engineering, Xavier Institute of Engineering*

Abstract*: This paper proposes a deep learning approach for the classification of different soccer actions like Goal, Yellow Card and Soccer Juggling from an input soccer video. The approach used for the same included a Hybrid model which consisted of VGG16 CNN model and Bidirectional Long short-term memory (Bi-LSTM) a Recurrent Neural Network (RNN) model. Our approach involved manually annotating approximately 400 soccer clips from 3 action classes for training. Using the VGG16 model to extract the features from the frames of these clips and then training the bi-LSTM on the features obtained. Bi-LSTM being useful in predicting input sequence problems like videos.

Keywords: Soccer Videos, Convolution Neural Networks (CNNs), Recurrent Neural Network (RNN), Bidirectional Long short-term memory (Bi-LSTM)

I. INTRODUCTION

Classification of Videos into different genres or groups is a very effective way to retrieve information from a large set of data. The objective involves automatically tagging each segment of videos based on the group to which they belong to. Soccer being one of the most popular games having a massive fan base around the globe and a large set of video data available we chose to classify soccer action videos. Working with videos is quite trickier than images considering their dynamic nature. Convolution Neural Networks have been exhibited as a viable class of models for working with images and giving outstanding results on image classification, segmentation and detection. The factors behind these results were techniques for scaling up the network to countless parameters and large labelled dataset which can support the learning process. In these circumstances, CNNs have shown to find interpretable image features. In image classification, the image to be classified is passed through the trained model which predicts in what category the image belongs to. Video classification works in a similar way but with a few additional steps. Here the extracted frames from the videos pass through the different layers of the CNN model and the features are obtained. These features are fed to the RNN. RNNs are known to deal with time series data that are preferable for video classification.

II. BACKGROUND

CNNs have until recently been applied to relatively small-scale image recognition problems (on datasets such as MNIST, CIFAR10/100, NORB, and Caltech-101/256), thus improvement on GPU hardware have enabled CNNs to scale to networks of millions of parameters, which has in turn led to significant improvements in image classification, object detection, scene labelling [1]. Additionally, features learned by large networks trained on ImageNet have been shown to yield state-of-the-art performance across many standard image recognition datasets when classified with an SVM, even with no fine-tuning [2]. A convolution based neural network however is best for problems concerning Time series data and analysis. The deep-bidirectional LSTMs (Bi-LSTM) networks are a variation of normal LSTMs, in which the desired model is trained not only from inputs to outputs, but also from outputs to inputs [3]. In a Bi-LSTM model the given input data is utilized twice for training (i.e., first from left to right, and then from right to left). Bi-LSTM models outperforms the regular unidirectional LSTMs. The better performance of Bi-LSTM compared to the regular unidirectional LSTM is understandable for certain types of data such as text parsing and prediction of next words in the input sentence. Experimental analysis has proved that Bi-LSTM is better than regular LSTM [3].

III.PROPOSED APPROACH

A. Overview

Convolutional neural network and Bidirectional Long Short-Term Memory also referred to as CNN Bi-LSTM is an architecture designed for predicting sequence of spatial inputs like images or videos. LSTM architectures were primarily developed for visual time series description related problems. Using this architecture, it is possible to obtain textual description of a sequence of images or videos wherein CNN is used that is pre-trained on a challenging image classification task that is used as a feature extractor and given as an input to LSTM architecture.

Involves the use of Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to support sequence prediction by interpreting the features across time steps. We use Bidirectional LSTMs that helps in improving the model performance as compared to a traditional LSTM on sequence classification problems. In problems where all timesteps of the input sequence are available, Bidirectional LSTMs train two instead of one LSTM layer on the input sequence. The first on the input sequence as-is and also the second on a reversed copy of the input sequence. this will offer further context to the network and lead to quicker and even fuller learning on the problem.

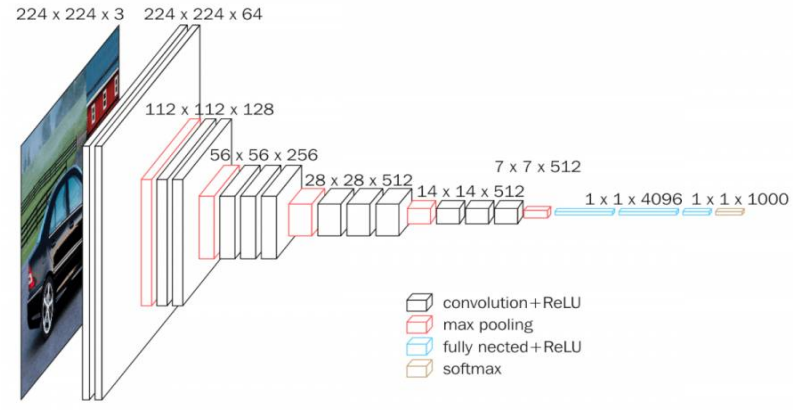


Fig. 1 VGG16 architecture [4]

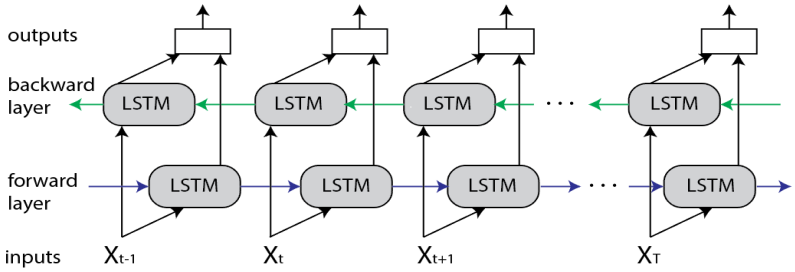


Fig. 2 Bi-LSTM architecture [5]

B. Algorithm

1) CNN

- Splitting the dataset and assigning 70% of the videos to the training set and 30% of the videos to the validation set.
- Extract frames from the videos.
- Train the CNN model using the dataset.
- Use the trained model to predict the frames of the video.

2) Bi-LSTM

- Input to the bi-Lstm would be the obtained features.
- Train the CNN bi-LSTM model with the dataset.
- Shuffle the dataset and use the trained model to predict the videos.
- Calculate the loss and accuracy of training and validation sets.
- Input random videos to check classification.

IV. EXPERIMENT AND PERFORMANCE EVALUATION

A. Dataset

We used our self-annotated dataset that contains soccer clips from SoccerNetv2, which we shall refer to as Soccer-3 from now on. Soccer-3 comprises 423 soccer clips. We annotated 3 action classes; goal action (135 clips), yellow card (141 clips) and soccer juggling (147 clips). All videos have a frame rate 25fps. Resolution of the clips in our dataset was 1080x1920 pixels which we then cropped and resized to 224x224 pixels. Furthermore, we split the dataset into training and testing sets.

B. Training

Initially we split our dataset by assigning 70% of the videos to the training set and 30% of the videos to the validation set. Our epochs were 20 and rmsprop being a good optimiser was used for adjusting the learning rate throughout the training.

C. Result

On Training the model, the features extracted were passed to the bi-LSTM model. We classified the videos and got an accuracy of 0.96. We use our model to input a random video other than the testing video to predict the label and classify the video into the given category. Using OpenCV we display the video clip along with the label it belongs to on the top left corner of the video.

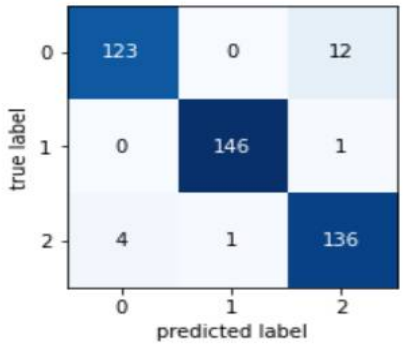


Fig.3 Confusion matrix

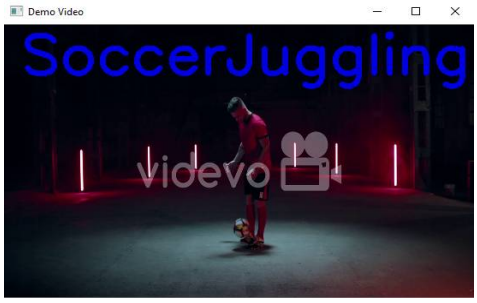


Fig. 3 Classified Video

Fig. 4 Classified Video

Fig. 5 Classified Video

V. CONCLUSIONS

We thus studied the performance of VGG16 Bi-LSTM hybrid model on video classification problem. There are many other existing methods for video classification that have given good results. This article has some limitations too like the training process takes longer, inability of handling multiple features at a time etc. Classifying longer actions also requires larger computation thus a limitation. Though extensive efforts have been made in video classification with deep learning, we believe we have just scratched the surface when it comes to using deep learning with large videos.

This can further be extended to include additional action classes like Free kicks, corner-Kicks, Red-cards, etc. or even more by using it for summarizing long soccer videos based on these action classes for highlight generation. This can be done by giving segments of videos as an input to the CNN LSTM model whose inclusion in a summarized video will be based on its validated relevance. Each segment is treated as a potential highlight that is independently evaluated to approve its inclusion in a summary video. And finally concatenating these segments to form the summarized videos.

VI. ACKNOWLEDGMENT

We would like to Thank and express our gratitude to prof. Omprakash Yadav for his guidance, valuable and constructive suggestions during the planning and development of this project.

REFERENCES

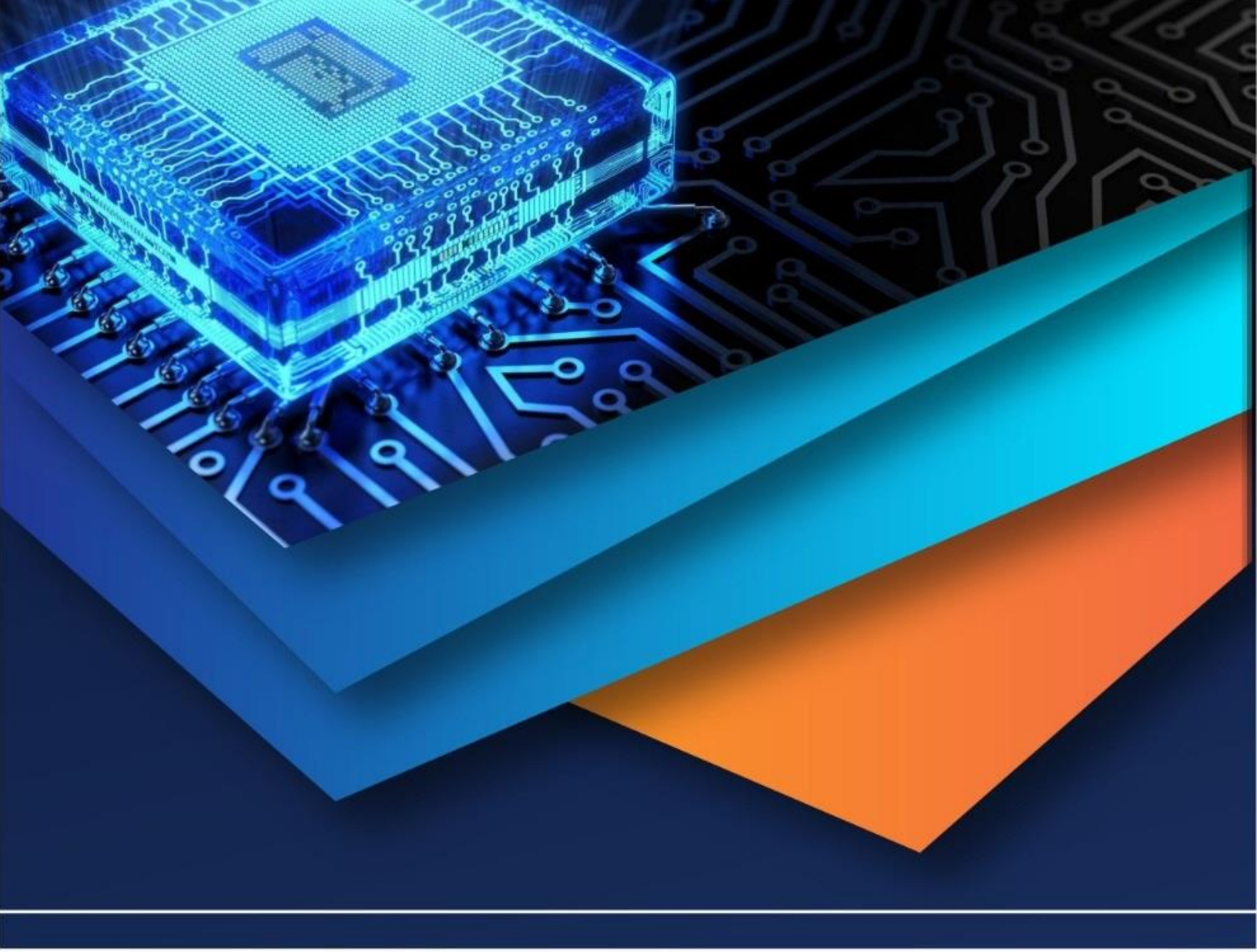
[1] Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," in *CVPR*, 2014.

[2] A. Sharif Razavian, H. Azizpour, J. S and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," 2014. [Online]. Available: arXiv:1403.6382v3.

[3] S. Siami-Namini, N. Tavakoli and A. Siami Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," in *IEEE International Conference on Big Data (Big Data)*, 2019.

[4] "VGG16- Convolutional Network for Classification and Detection", Neurohive, 20 November 2018. [Online]. Available: https://neurohive.io/en/popular-networks/vgg16/.

[5] E. Zvornicanin, "Bi-Lstm architecture", *Baeldung*, 5 February 2022. [Online]. Available: https://www.baeldung.com/cs/bidirectional-vs-unidirectional-lstm.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY