VIDEO TO TEXT SUMMERIZATION USING NEUTRAL NETWORKS

Project Reference No.: *45S_BE_2200*

College ***:*** *A.P.S. College of Engineering, Bengaluru*
Branch ***:*** *Department of Information Science and Engineering*
Guide(s) ***:*** *Ms. Pallavi H B*
Student(S) ***:*** *Mr. A Balaji*
Ms. Bhavana R
Ms. Varuna T S
Ms. Pavithra D

Keywords:

Convolution layer, pooling layer, flattering, Video captioning model

Introduction:

The prevalence of recording devices encourages more people to capture their daily life with video data content. But, the large amount of video data makes it more difficult to navigate, particularly long videos such as surveillance videos or CCTV footage. For larger videos, identifying the important parts/frames of the video content and enabling them with captions will give a richer and more concise condensation of the video. So, the video summarization has been proposed to extract a compact representation of the video data into textual form. The proposed system offers a brief semantic understanding of a long video just through a text summary.

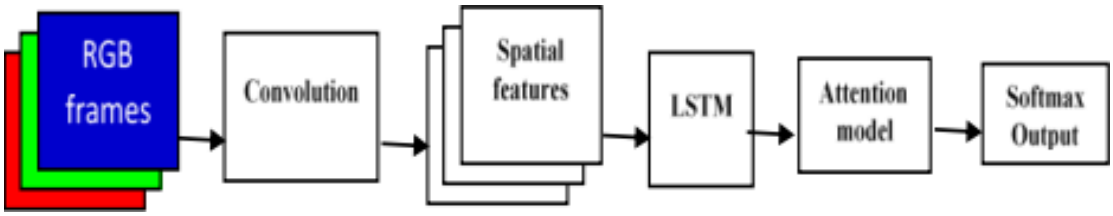
Traditionally, video summarization is done either by taking a holistic view of the entire video or by identifying the local differentiation among the adjacent frames. Some researchers utilize web media and metadata as prior knowledge to generate better summarization results. Visual attention is also used to select important frames. However, video summarization requires a semantical understanding of the video content and is hard to model with a heuristic design.

Objectives:

In this project, we propose a method for video scene classification with the particular intention of video summarization. This will help physically challenged students and senior citizens. It will help during the online classes, in which students can easily takedown the content present in the video.

Methodology:

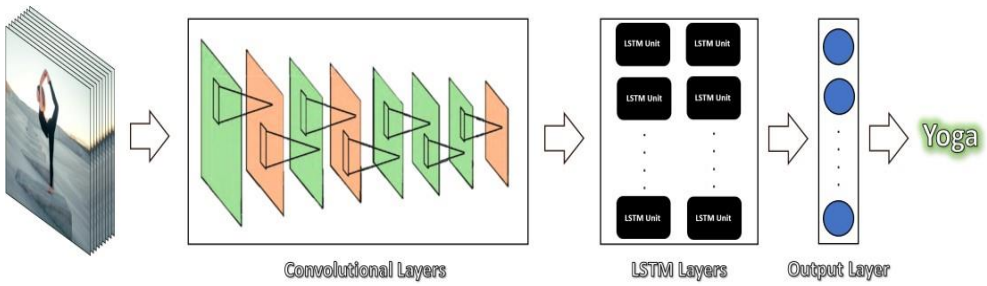
This paper proposes a novel architecture for human action recognition using LSTM based CNN from video frames. The proposed method contains CNN, LSTM and attention models. The convolution layer captures the spatial information. Consequently, the LSTM layer captures the temporal feature information. The attention model is combined with LSTM that captures the important feature information of video that avoid the unwanted noise from the frames. It improves recognizing performance of the LSTM based network. The output of LSTM is a vector that notifies temporal feature information of video frames. Fig 1. Shows the LSTM-CNN framework with convolutional features and attention model. CNN filtered the spatial information from each frame of video and LSTN-CNN explored the temporal information among the various frames in the video. Here, the attention model is combined with LSTM-CNN. The training phase of the network model uses the video labels for action recognition. The CNN captures the different spatial information such as curves, shapes, location, invariance, rotation invariance, etc.). Attention model is used to emphasize the moving objects than the entire image or static background that decreases the effect background effect. It led to increased performance of the designed network model.



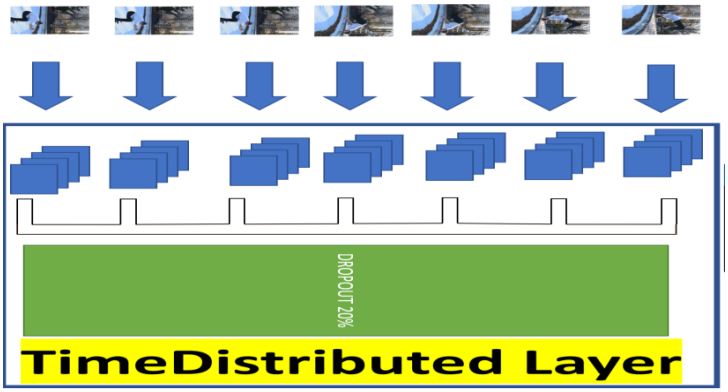
The output of CNN is fed as input to the LSTM layer. LSTM consists of many components such as input gate, forget gate, input modulation gate and output gate. The input feature vector represents xt, cell state as Ct, hidden state as ht and output state as Ot. The output is the tanh computation of hidden state.

Here, we will implement the LRCN Approach by combining Convolution and LSTM layers in a single model. Another similar approach can be to use a CNN model and LSTM model trained separately. The CNN model can be used to extract spatial features from the frames in the video, and for this purpose, a pre-trained model can be used, that can be fine-tuned for the problem. And the LSTM model can then use the features extracted by CNN, to predict the action being performed in the video.

But here, we will implement another approach known as the Long- term Recurrent Convolutional Network (LRCN), which combines CNN and LSTM layers in a single model. The Convolutional layers are used for spatial feature extraction from the frames, and the extracted spatial features are fed to LSTM layer(s) at each time- steps for temporal sequence modeling. This way the network learns spatiotemporal features directly in an end-to-end training, resulting in a robust model.

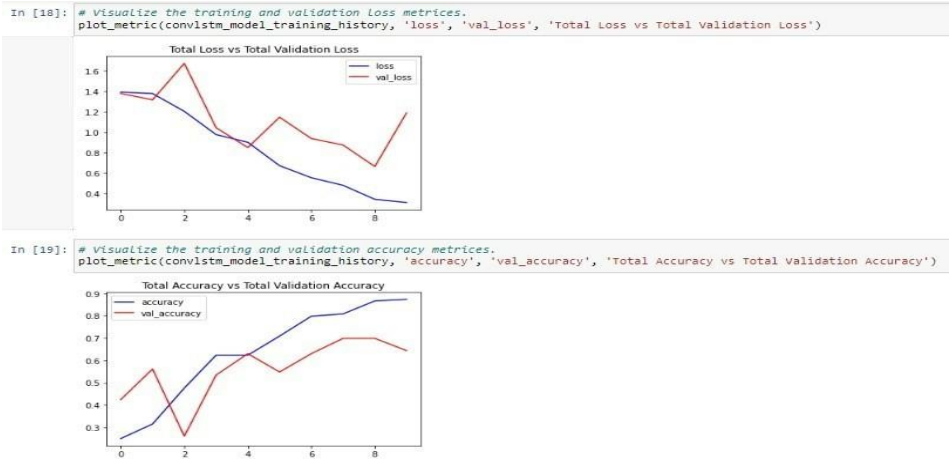


We will also use **Time Distributed** wrapper layer, which allows applying the same layer to every frame of the video independently. So it makes a layer (around which it is wrapped) capable of taking input of shape (no_of_frames, width, height, num_of_channels) if originally the layer's input shape was (width, height, num_of_channels) which is very beneficial as it allows to input the whole video into the model in a single shot.



Conclusion:

This project is proposed the integration of convolutional neural network and long short-term memory recurrent neural network for processing the video. The convolution processes the given input that produces the informative spatial features. It captures the highly valuable informative features in the frame of video. The actions are recognized from the informative features using softmax module. This model is used to recognize the human actions from video. The experimental results proved that proposed model performed better with accuracy.



Scope for future work:

In future we are going to add more number of images which will help blind people and we will try to process large video and give more information about that.