ACTION RECOGNITION USING RNN

SAI BHARAT KONAKALLA
PRANAY REDDY DAVA

University at Buffalo The State University of New York

Abstract

Research in human action recognition has accelerated significantly since the introduction of powerful machine learning tools such as Convolutional Neural Networks (CNNs). However, effective and efficient methods for incorporation of temporal information into CNNs are still being actively explored in the recent literature. Recurrent neural network (RNN) and long short-term memory (LSTM) have achieved great success in processing sequential multimedia data and yielded the state-of-the-art results in speech recognition, digital signal processing, video processing, and text data analysis. In this report, we propose a novel action recognition method by processing the video data using convolutional neural network (CNN) and LSTM network.

Data Set

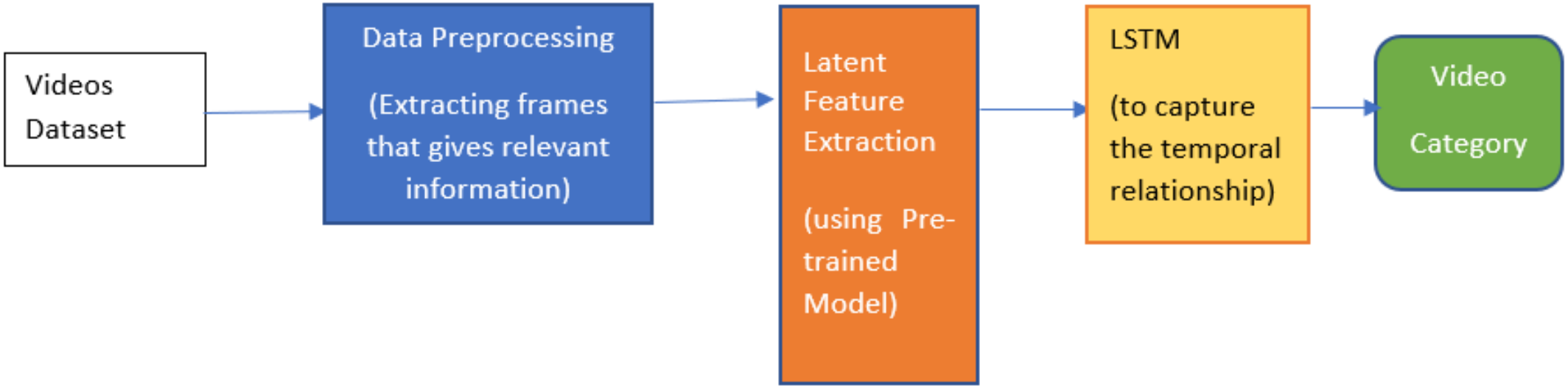
- We have used UCF-101 dataset which contains 101 actions and 13,320 videos in total. With 13320 videos from 101 action categories, UCF101 gives one of the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. As most of the available action recognition data sets are not realistic and are staged by actors, UCF101 aims to encourage further research into action recognition by learning and exploring new realistic action categories. The action categories can be divided into five types: 1)Human-Object Interaction 2) Body-Motion Only 3) Human-Human Interaction 4) Playing Musical Instruments 5) Sports.

Data Flow

This whole process can be categorized into three steps:

- Segmentation.
- Feature Extraction.
- Model training.

Data Flow

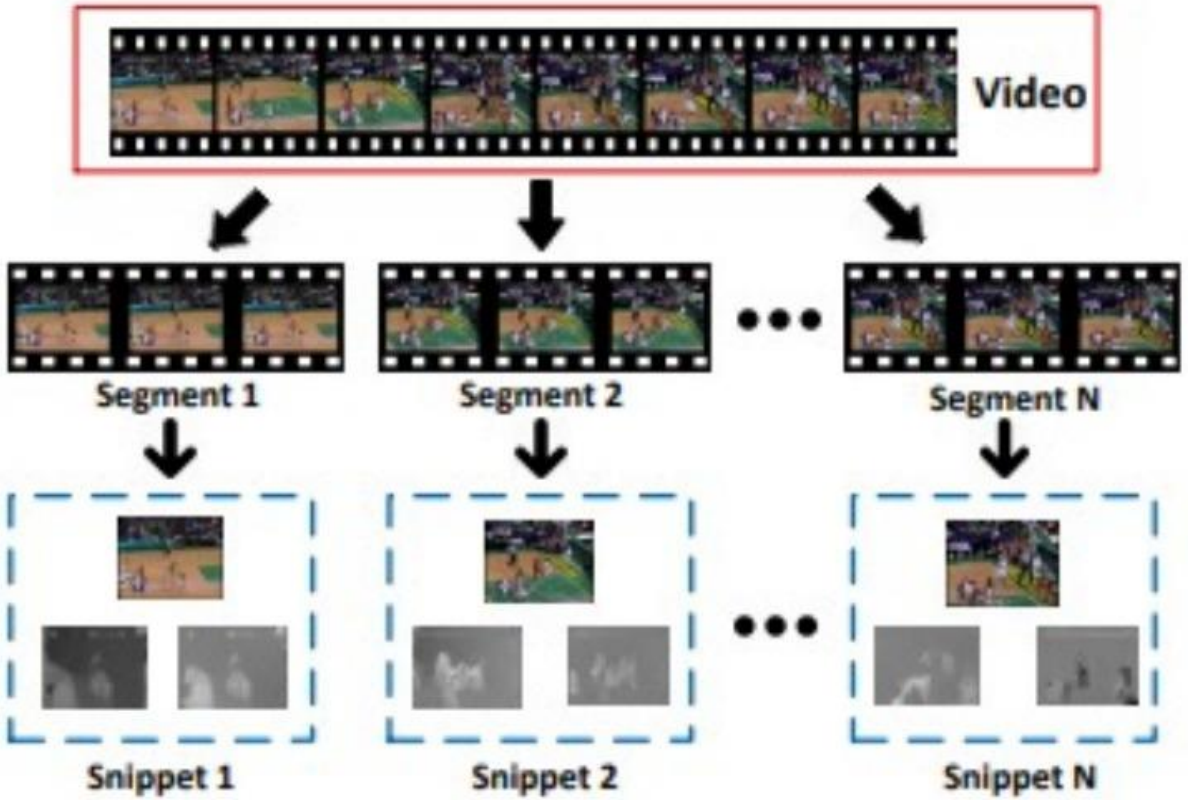


Data Flow

Segmentation

- Videos are nothing but a pile of frames, so we have captured 40 frames for every video in our dataset. While extracting frames from the videos we took the frames in such away that the sequence of frames describes the whole video. In the dataset for every category of videos there are different sub-categories, like a particular sub-category has some ‘n’ number of videos, so to train the model well we have divided the frames extracted into training and testing as for every sequence of 5 videos 4 videos goes to training and 1 video goes to testing data so that the model will be exposed to every kind of video in the dataset.

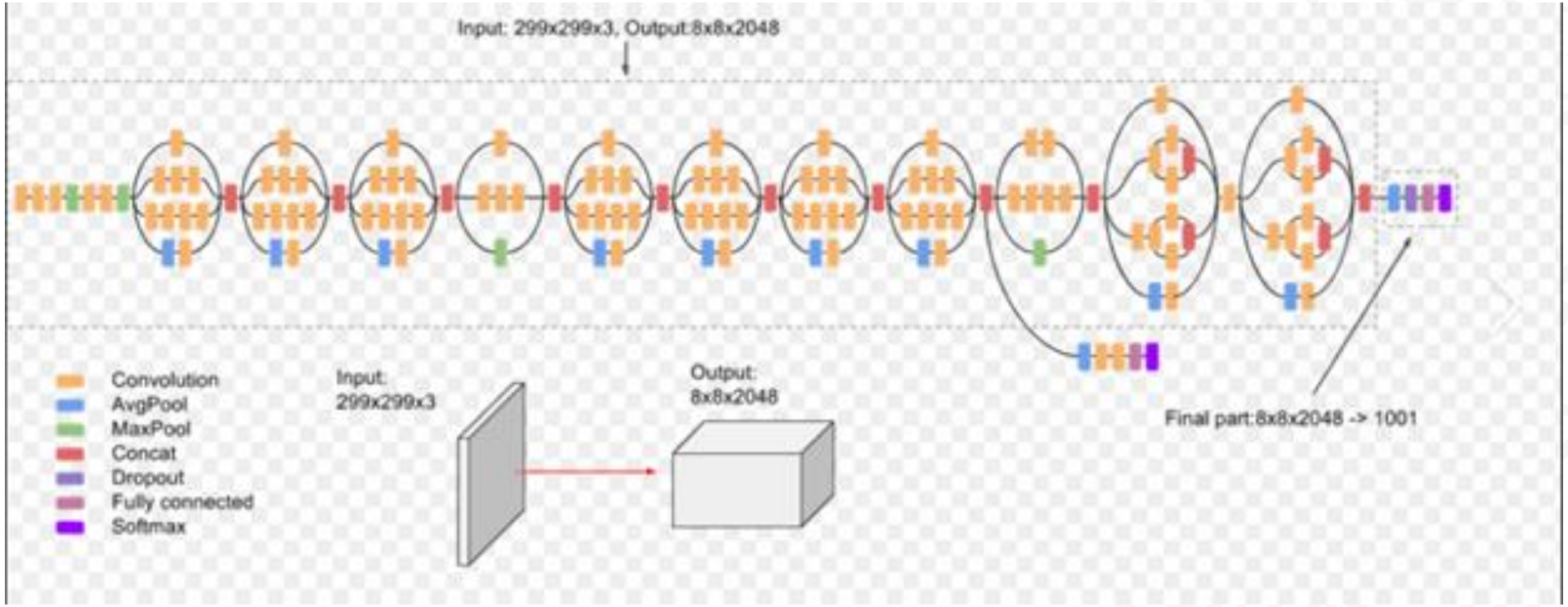
Segmentation



Feature Extraction

- Transfer learning is a machine learning method which utilizes a pre-trained neural network. Here, we used image recognition model, Inceptionv3.
- We have used Transfer learning for feature extraction. Our problem in this case looked similar to Image-net problem (extracting features).
- The frames extracted from the videos are resized to (299,299,3) dimension as the input shape to Inception V3 model is fixed.
- For every video, the extracted feature output from CNN will be of (1,40,2048) shape as it contains 40 frames and 2048 features for every frame.

Feature Extraction



Model training

Although we can use Convolutional Neural Networks to derive the relationship between frames, that was not enough to capture the temporal relationship. We used keras tensorflow package to create RNN's and also experimented with the following networks:

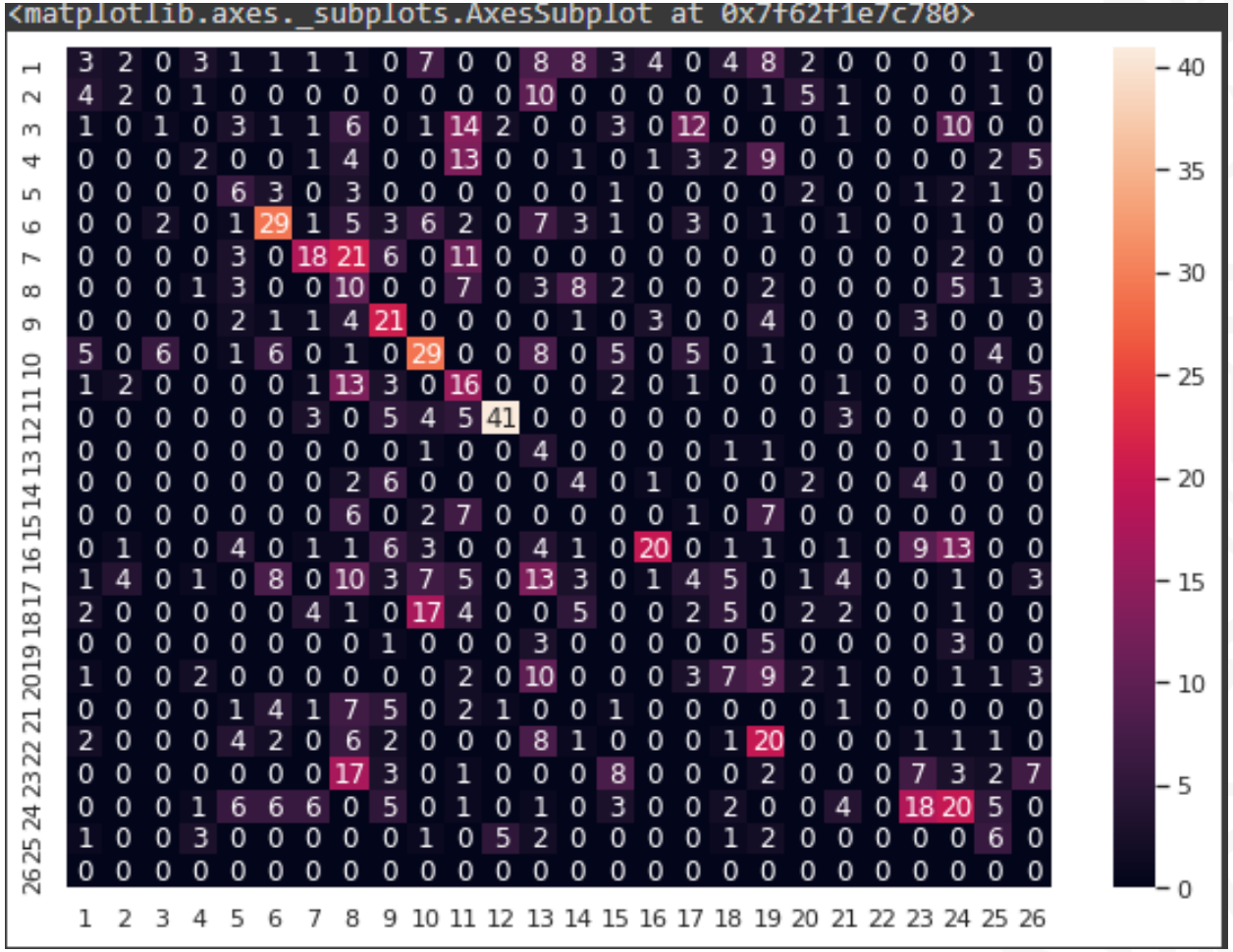
- SimpleRNN layer with tanh activation function
- SimpleRNN layer with relu activation function
- LSTM

Networks other than LSTM are facing vanishing gradient problems when trained on this data. LSTM worked better than the rest for our task. We have used LSTM layer with 2048 nodes and a dropout of 0.5.

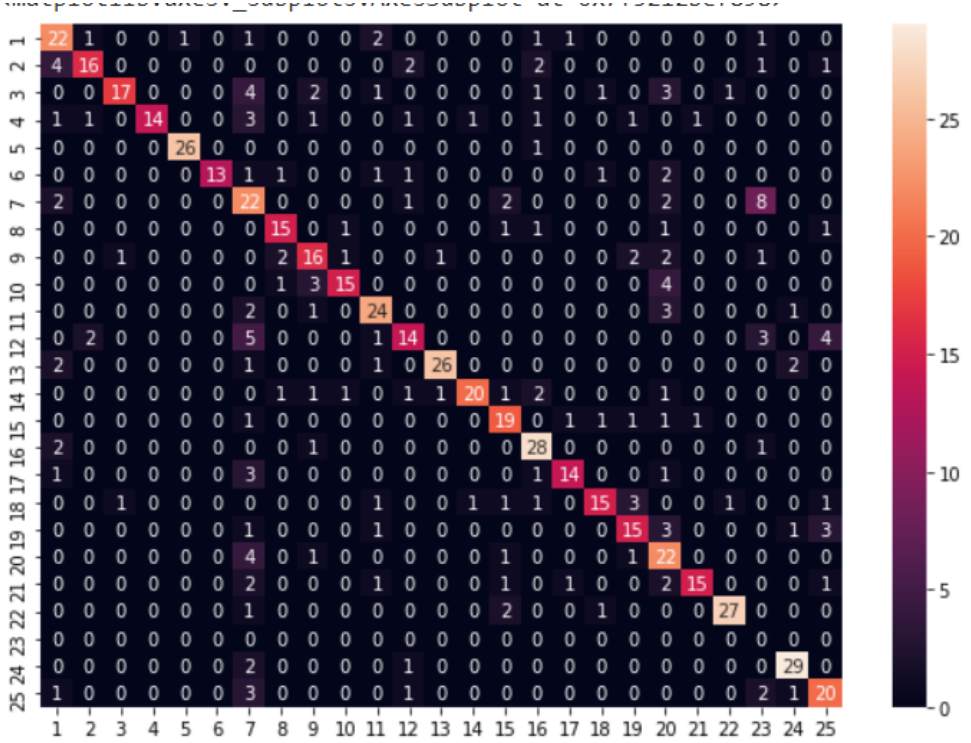
Top K Categorical Accuracy

TopK Categorical Accuracy calculates the percentage of records for which the targets (yTrue) are in the top K predictions (yPred). We rank the yPred predictions in the descending order of probability values. If the rank of the yPred present in the index of the yTrue is less than or equal to K, it is considered accurate. We then calculate TopK Categorical Accuracy by dividing the number of accurately predicted records by the total number of records

Experimental Results



Final Results



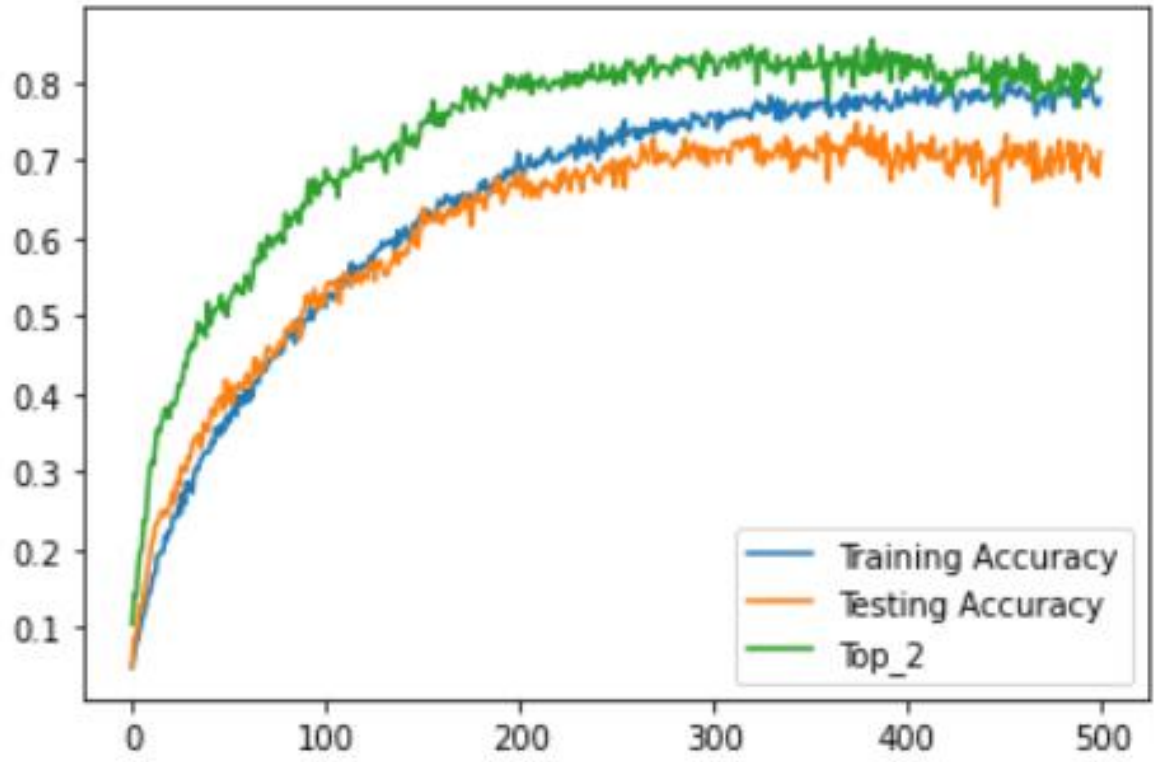
Super Category	Selected video categories.
Human-Object Interaction	'ApplyEyeMakeup','BrushingTeeth', 'HammerThrow', 'HulaHoop', 'Typing'.
Body - Motion	'PullUps', 'RopeClimbind', 'TaiChi', 'Swing', 'TrampolineJumping'.
Human – Human Interaction	'BandMarching','HeadMassage'.
Playing Musical Instruments	'PlayingCello', 'Drumming'.
Sports	'Archery','BasketballDunk', 'BoxingPunchingBag','BreastStroke', 'FieldHockeyPenalty','HorseRace','PoleVault', 'SumoWrestling','IceDancing', 'Kayaking'.

Accuracy

- Accuracy in the model indicates how well the model is able to classify a given video into one of the category.

Accuracy : 75%

Top ‘2’ categorical accuracy: 87%



Conclusion and Future work

- There is a lot of scope in Action recognition and this field is worth exploring. There are have been various attempts to successfully model sequences. Most recent promising results state of art achieved 94 percent. Here are a few ideas which could be explored in future work.
- Actions are predominantly determined by movement of joints of the person involved in the action. A dataset which contains both video and 3D skeletal data was released in CVPR'16. Using skeleton data along with the video would definitely yield better results in action recognition.
- Action detection and understand the activity is of utmost importance as activity recognition finds its application in various fields such as personal healthcare like fitness tracking, fall detection of elderly people, monitoring functional and behavioral health using wearables.

Sample GIF(Basketball Category)

Thank You