How to Cite:
Borikar, D. A., Kumar, S., Kathawate, R. B., Baiswar, S. P. ., & Roy, S. J. (2022). An approach to temporal concept localization in videos. *International Journal of Health Sciences*, *6*(S1), 10473–10483. https://doi.org/10.53730/ijhs.v6nS1.7524

An approach to temporal concept localization in videos

Dilipkumar A. Borikar
Shri Ramdeobaba Colllge of Engineering and Management, Nagpur, Maharashtra, India
Corresponding author email: borikarda@rknec.edu

Sushant Kumar
Ola Cabs, Bengaluru, Karnataka, India

Rakshit Bhagwat Kathawate
VMware India Pvt. Ltd., Bengaluru, Karnataka, India

Sarthak Prakash Baiswar
Compass Inc., Hyderabad, Telangana, India

Sourav Jagannath Roy
p360 Solutions, Mumbai, Maharashtra, India

***Abstract*---**Localizing moments in the videos has been a new challenging task in the field of Computer Science to provide faster search time for video retrieval, query processing and also behavioral analysis. The process involves stages such as video understanding, video segmentation, query processing using NLP and generation of localization of the concepts in the video. Though there have been many attempts to Video understanding in the field of NLP and Computer Vision in past years, they lack to cover the large untrimmed videos in current real-life scenarios. We propose the deep learning-based solution with the use of Random Forest and Bi-LSTM approach to localize the labels in segments and also the time at which they pertain to the particular segments. We used the YouTube 8M dataset provided by YouTube in Kaggle's challenge to train our frame-based model and use it to classify the segments using sliding windows of size 5. Our approach tries to provide a naïve and robust approach to model this concept and provide a way to tackle this large problem. Further improvements in the Bi-LSTM based models and Random Forest models with VLAD would lead to better results.

***Keywords*---**Video Understanding, NLP, Video Segmentation, Bi-LSTM, Random Forest, Concept Localization.

Manuscript submitted: 27 March 2022, Manuscript revised: 18 April 2022, Accepted for publication: 9 May 2022

Introduction

Localizing moments in long videos via natural language labels has always been a difficult task. It involves video understanding and video segmentation, with a proper sliding window which consists of optimal granularity in order to identify cross labelled segments across the untrimmed long videos. Labeling the respective segments of videos and providing this data in the metadata of video will boost the current 'query processing time' and 'optimal results time' for Google search queries and help the user find better results quickly and more efficiently.

Consider the image in Fig. 1, for the given query "The little girl talks after bending down", there is a need to understand, the objects and actions involved in the videos. This means, that apart from just video understanding, understanding the relationship between these actions should go hand in hand to process the natural language query to properly localize the moments in the videos. Current studies in Hendricks et al. and Gao et al. show that there is a wide application in the field of video search and retrieval [1], [2]. The model accepts the video input in mp3 format, convert it into TFRecord format using Google's pre-trained "Inception" and "PCA" (Principal component Analysis) to reduce the dimensionality of the datasets.

TFRecord is processed to extract video and audio level features. Extracted features are fed to the model which outputs the label for each frame. With the use of an optimal sliding window of size k (5), we divide frames into segments and classify it as the most occurring and consistent label among all the frames. The Output will also provide the start and end time for each label in the video which is used as the pointer for the search queries during video retrieval using natural language.

Fig. 1. Representative images for actions/movements

Motivation

In most web searches, video retrieval and ranking are performed by matching query terms to meta-data and other video-level signals. However, knowing that videos can contain an array of topics that aren't always characterized by the uploader and many of these miss localizations to brief but important moments within the video. Video understanding is similar to image understanding, which

has a wide variety of applications in the field of medical sciences, forensic, computer vision, etc. Video understanding also finds its use in real-life scenarios like search query on Google, YouTube, Behavioral analysis, Semantic segmentation for self-driving cars.

Our main aim is to target the current query processing time for optimal search for video retrieval on YouTube and the project was designed to provide the way for YouTube's 8M Video understanding challenge. It is intended to design a machine learning model for Temporal Concept Localization within the video to improve video understanding. The system will be identifying the topics substantially pertinent to a video and also pinpointing where exactly in the video they appear. When provided with video as the input, the model will detect the presence of an action which depicts a notable topic and localizes video-level labels to the precise time in the video. Traditional methods aim to manually annotate the videos by users on the web and use that metadata for video retrieval. Our aim is to automate the annotation of the videos and its timings.

Literarture Survey

There have been tremendous explorations about action classification in videos using deep convolutional neural networks (ConvNets). Representative methods include two-stream ConvNets, C3D (3D ConvNets) and 2D ConvNets with temporal LSTM or mean pooling. Specifically, Simonyan and Zisserman [3] modelled the appearance and motion information in two separate ConvNets and combined the scores by late fusion. Tran et al. [4] used 3D convolutional filters to capture motion information in neighbouring frames. The 2D ConvNets has been used to extract deep features for one frame and use temporal mean pooling or LSTM to model temporal information.

Most existing work on activity recognition focuses on classifying the pre - segmented clips. Past work has studied this topic under computer vision [3] as part of action recognition and object detection in images and then sequencing output to form the output for video. Also activity type range available in HMDB 51 [4] and UCF 101 datasets provides only 100 to 200 primitive human actions, sports or broad level categories; moreover the length of videos was also fixed and short, up to 5 to 10 seconds only.

Karpathy et al. proposed the Sports1M data set with more than 1 million untrimmed videos of Youtube to process the videos using CNNs [5]. Current computer vision approaches had failed to process such large videos and hence this topic was reviewed under the field of Deep learning in past years.. Action recognition typically involves two basic steps: feature extraction and classifier training. As a generally excepted approach the low level features are extracted and further aggregated into a fixed length vector for classification. Oneata et al. showed that a combination of visual features and audio features represented using Fisher Vectors produced state-of-the-art activity and event classification performance [6].

Shou et al. has proposed the PCA-based Convolutional Neural Network (PCN) approach to extract features using CNN [7]. This has been accepted as the most

efficient approach and has been extended using fine CNN based architectures [8]. Youtube uses Inception, VGGNET, imagenet to extract the visual and audial features from large youtube videos. Other extensions have seen the use of Double Sliding window [9] for classification of segments in videos by monitoring the actions of object for period in the window.

The modified approach to use single sliding window with optimal window size of k, to put forth the most appropriate label occurring recently in window to classify the given segment has been implemented successfully.

The long short memory network (LSTM) was also proposed by Hochreirter as an improvement over traditional CNN based architectures and RNNs for classification and prediction of labels since it simulates to the time series analysis [10]. Specifically, LSTM can remember the past weights and also has the ability to forget them unlike RNN where error gradient decays exponentially with more time lag between the two windows. This approach was also modified to use Bi-LSTM which is discussed in the next sections. More Recently, LSTM has been used for Video level classification [11][12] and also for generating the image descriptions [13].

Terminologies

Temporal Concept Localization

Temporal Concept localization refers to the identification of actions persistent in the image or video and locates the presence of it, in the videos. It also includes the use of Natural Language Processing to process concept given in the query, tokenize it and locate those concepts in the videos.

Video Understanding

Video understanding is one of the fundamental problems in computer vision [14]. Videos in addition to the image features like colour, texture include the temporal components that facilitate detecting motion in the image recognition task. Fig. 2 shows the sample of video understanding along with natural language statement to describe the actions in the video segment.

Fig. 2. Video Understanding Sample

VGG Net Model

The convolutional neural network based architecture, VGG was proposed in 2014 by Simonyan and Zisserman as an entry to Large Scale Visual Recognition Challenge 2014 (ILSVRC2014). The model achieves 92.7% top-5 test accuracy in ImageNet that constitutes of over 14 million hand-annotated images.

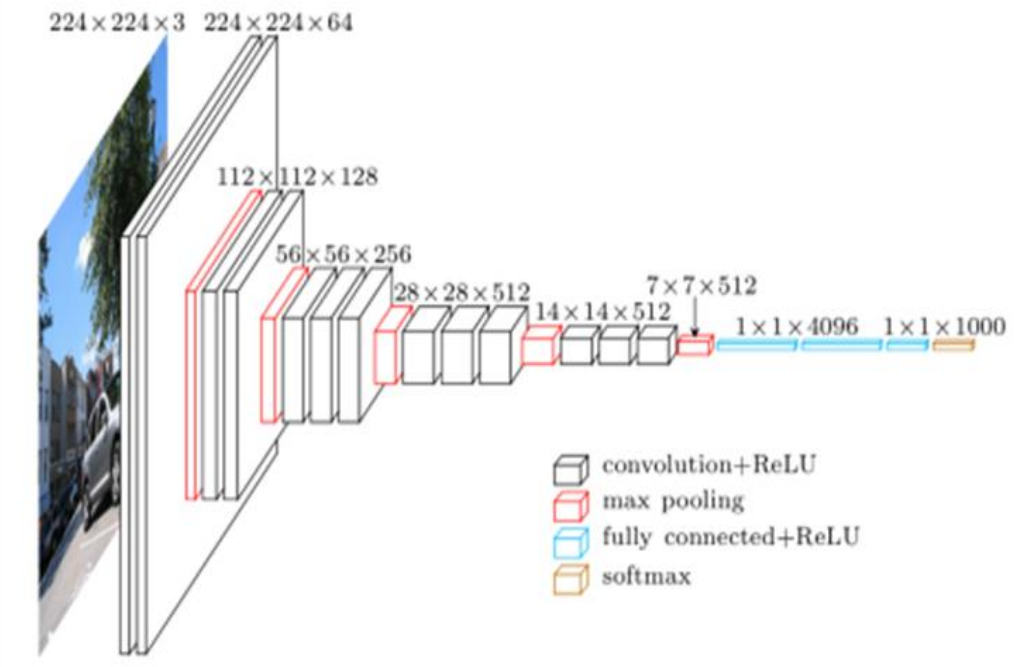

Fig. 3. *VGG Architecture*

TFR Record

TF Record is file format owned by Tensorflow organization. To read data efficiently it can be helpful to serialize your data and store it in a set of files (100 - 200 MB each) that can each be read linearly. This is especially true if the data is being streamed over a network. It can be used effectively for caching in data pre-processing task. TF Record stores a sequence of binary records. The cross-platform, cross-lingual library, the Protocol buffer provides for efficient serialization of the structured data. Protocol messages are defined using .proto files and are designed for use with Tensor Flows.

Dataset

Our proposed solution was designed for YouTube 8M Videos challenge 3rd edition, so we used the datasets that has been provided by YouTube on [8M dataset]. Dataset description used for Training is as follows:

- The dataset to be used is YouTube-8M dataset which consists of 237k Human-verified segment labels, spread across 3862 different classes. It has

over 2.6 billion audio &video features with average 3 labels per video havinga total size of more than 1.5TB.
- Dataset is divided into number of shards and stored in the TFRecord format for faster processing of the data.

For the training purpose the data will have 8 parameters:
- Video ID (Identifier for every video)
- Video-level labels (A list of labels for the whole video)
- Frame wise information for RGB per frame (1024 8bit quantized features)
- Frame wise information for AUDIO per second (128 8bit quantized features)

For validation, segment level dataset was provided which was extension of 8M Frame level dataset with human verified segment annotation for each segment. Dataset has human-verified labels on about 237K segments on 1000 classes from the validation set of the YouTube-8M dataset. The dataset description is as follows:
- Segment start time (A list of start time for the segments)
- Segment end time (A list of end time for the segments)
- Segment labels (A list of labels for each segment)
- Segment Score (A value to indicate the presence of a particular label in segment.

The analysis was performed to review the data received. All the dataset shards were initially in the TFRecord format which is binary format storing serial data. It was extracted with the help of various scripts provided on the official documentation of TFRecord site. It was converted into 1152 Column vector for each row. 1024 VIDEO features, 128 AUDIO features, LABEL ID present in video, VIDEO ID of video on YouTube.

Approaches
LSTM based approach

Video Classification, seen frame by frame is actually image classification. The Research Papers on Image Classification show a (second) approach of using a bidirectional LSTM. Bidirectional Long Short Term Memory is a form of Recurrent Neural Network used in Sequence Modelling.

Long Short-term Memory (LSTM) is a type of recurrent neural network (RNN) that solves the vanishing and exploding gradients problem of previous RNN architectures when trained using back-propagation. Standard LSTM architecture includes an input layer, a recurrent LSTM layer and an output layer. The recurrent LSTM layer stores real-valued state information from the previous state observations in a set of memory cells. This recurrent information flow, from previous observations, is particularly useful for capturing temporal evolution in videos, which we hypothesize is useful in distinguishing between fine-grained sports activities. The LSTM memory cells through input gates and forget gates allows them to maintain long-term memory and ability to reset when required.

Our approach involves fetching of a sequence of frames, process them and then detect proper classification label. This allows us to have a proper slice of video

classified to a particular label. To accomplish this task the Bi-LSTM approach was found most appropriate.

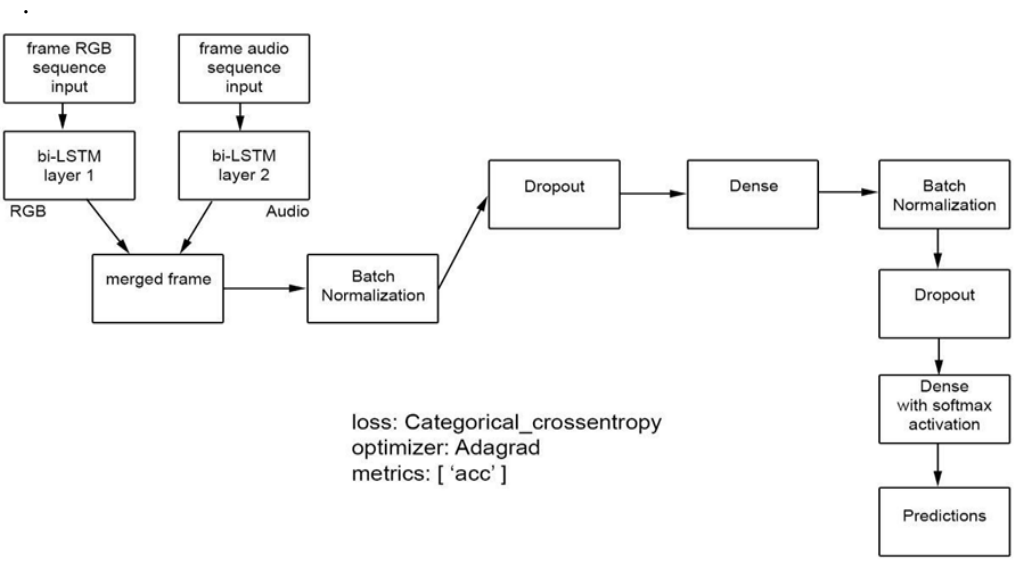

Fig. 4. *LSTM Model*

A LSTM looks into the past to learn the weights and predict the future parameters. In Bi-LSTM, the learning is done both ways learning from past as well as the future attributes. While using Bi-LSTMs the learning algorithm is exposed to original data in forward and reversed directions.

Our Approach for bi-LSTM uses separate 2 bi-LSTM layers for video attributes and audio attributes. To this, we are passing a group of 5 encoded frames to be sent as a batch to learn. This is treated as a sequence in Bi-LSTM and weights are learnt processing the encoding and their changes with each frame. Initially the video and audio bi-LSTM layers are separate but later in the model they are concatenated to be sent as a merged frame for normalization and dropout layers. The model architecture is depicted in Fig. 4.

Random Forest based approach

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges so to a limit as the number of trees in the forest becomes large. Using a random selection of features to split each node yields error rates that compare favourably to Adaboost [30], but are more robust with respect to noise.

Random Forest is considered to be robust in nature because single decision tree may be prone to a noise but aggregate of many uncorrelated decision trees reduce the effect of noise, providing more accurate results over large datasets also [28]. Since the attributes for the dataset were quite large, we used to Random Forest Using Random Input Selection method to decide the root for the decision trees.

It has been seen that random input selection is much faster than adaboosting or baging for set of decision trees in the random forest approach. There are various approaches for root decisions like ID3, Gini index, gain ratio, etc., but GINI index was found to be suitable over the other approaches in current scenario. Tree was allowed to grow to the maximum depth of 30 and prior pruning was applied to limit the model and stop it from over fitting. Following are Model specifications:

- Number of trees: 200
- Number of Jobs: 200
- Max Depth: 30
- Criterion : Gini Index

Segment Classification

With use of LSTM and RF for the Frame level label classification, we used sliding window of size k, to classify the frames into segments. If particular label occurs successively in sequence of frames, then without the loss of generalization it can be assumed that it is correctly classified segments. With this assumption, frames are grouped into group of 5 called as segments and each segment is labelled according to the most probably occurring label in window of size 5. Size of Window was decided after reviewing the dataset and it was pre-processed to contain the average of 5 frames per segments or bunch of multiple of 5 by YouTube.

The above approach was highly successful in random forest but in Bi-LSTM due to correlation of labels among each other it failed to classify correctly, as explained above. Hence The random forests were found as better alternatives for the task.

Architechure

Training

- Feed the Algorithm with input feature vector [1024 Video + 128 Audio + Label Id + Video ID] 1154 column input vector.
- Train the algorithm on the feature vector which will learn to classify the segments into given classes and generate model for testing.

Testing

- Provide the test video to the feature extractor script which generates [1024 Video + 128 Audio] 1152 column feature vector.
- A feature vector is tested using the trained model to get segments in the video with labels predicted by the model.

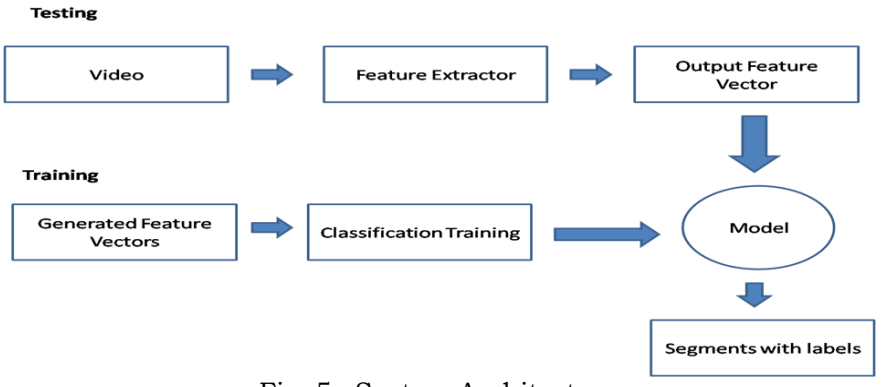

Fig. 5. System Architecture

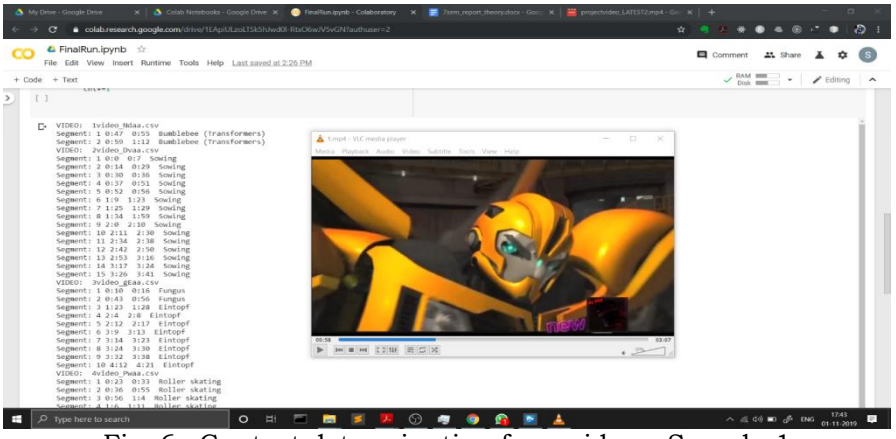

Fig. 6. Context determination from video – Sample-1

Fig. 7 and Fig. 8 show the processing of video file using random forest approach. The localization of the contexts from the images is evident in the output resulting as listing.

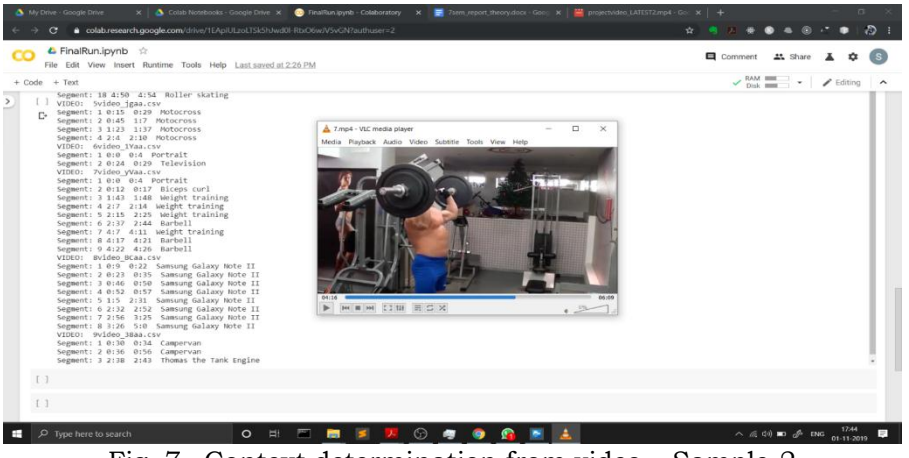

Fig. 7. Context determination from video – Sample-2

As seen in the Fig.7 at 0:58 minutes it indicates the “sowing” context, whereas in Fig. 8 at 04:16 minutes it indicates the “weight lifting” context correctly.

We addressed the problem of Temporal Concept localization in Videos available on the YouTube 8M dataset given for 3rd Edition challenge. Out of the proposed approaches, random forest came out to be working better with testing accuracy of 84.2%. With gained accuracy we can predict the time slices where a particular a label occurs in the segment and use it as metadata during Video search and faster retrieval. The other approach of Bi-LSTM can be further explored with restricted data and more larger deep learning model architecture. Due to computational power of machines, we used a small subset of 1.5 TB dataset but with larger dataset the random forest model will grow better and evolve much better to increase accuracy. The random forest approach is useful in shortening of video search time, GIF creation and may facilitate other future applications.

Conclusion

The sliding window approach for video segment classification using random forest classifier has been implemented for “temporal concept localization” in videos. This approach can be further improved with the use of VLAD, Transformers and CCRL XGB. The improvisation would result in porting this model on candidate pool to ensemble the results using a post-processing filter to increase the accuracy. It can also address the problem of correlated label misclassification. The NLP approaches can be incorporated to use the models to generate the natural language sentences along with segment times to provide better video understanding for a particular video. Bi-LSTM along with Local Action Focus (LAF) method can also be used in order to improve the convergence of models in correlated labels situation.

References

1. JiyangGao, Chen Sun, Zhenheng Yang, Ram Nevatia, “TALL: Temporal Activity Localization via Language Query” ICCV, 2017, pp. 5267-5275
2. Hendricks, Lisa Anne, et al., “Localizing Moments in Video with Temporal Language.” EMNLP (2018).
3. SimonyanK. and Zisserman A., “Two-stream convolutional networks for action recognition in videos”, NIPS, 2014.
4. Tran D., Bourdev L., Fergus R., Torresani L., and Paluri M., “Learning spatiotemporal features with 3d convolutional network”,ICCV, 2015.
5. Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R. and Fei-Fei L., “Large-Scale Video Classification with Convolutional Neural Networks,” 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.
6. Oneata D., Verbeek J., andSchmid C., “Action and Event Recognition with Fisher Vectors on a Compact Feature Set”, ICCV, 2013.
7. Shou Z., Wang D., and Chang S. F., “Temporal action localization in untrimmed videos via multi-stage CNNs” ,CVPR, 2016.
8. Sun Chen, “Temporal Localization of Fine-Grained Actions in Videos by Domain Transfer from Web Images”, MM’15 Proceedings of the 23rd ACM international conference on Multimedia Pages 371-380.

9. Wu J., Yin B., and Qi W., “Video Motion Segmentation Based on Double Sliding Window”, IEEE, November 2011.
10. HochreiterS. and Schmidhuber J., 1997. “Long Short-Term Memory”. Neural Comput. 9, 8 (November 15, 1997), 1735–1780.
11. Donahue, Hendricks L. A., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K. and Darrell T., “Long-term recurrent convolutional networks for visual recognition and description”, CVPR, 2011.
12. Srivastava, Mansimov E. and Salakhutdinov R., “Unsupervised learning of video representations using LSTMs”, ICML, 2015.
13. Wang and Schmid, “Action Recognition with Improved Trajectories”, ICCV, 2013.
14. Joe Yue-Hei Ng, Matthew Hausknecht, SudheendraVijayanarasimhan, OriolVinyals, RajatMonga, George Toderici, “Video Understanding with Deep Networks”, University of Cornell, 2015.