Human Activity Recognition Using LSTM/BiLSTM

Article *in* International Journal of Advanced Science and Technology · January 2020

CITATIONS
5

READS
18

5 authors, including:

Rajiv Vincent
VIT University Chennai
30 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Arun Kumar Sivaraman
VIT University
34 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Rajesh M
VIT University
18 PUBLICATIONS **35** CITATIONS
SEE PROFILE

International Journal of Advanced Science and Technology
Vol. 29, No. 04, (2020), pp. 7468 – 7474

Human Activity Recognition Using LSTM/BiLSTM

*Rajiv Vincent[1], Akshat Wagadre[2], Arun Kumar Sivaraman[3], M Rajesh[4],
[5]Arun Rajesh
[1,2,3,4]*SCOPE, Vellore Institute of Technology, Chennai, Tamilnadu, India*
[5]*Faculty of IT, Higher College of Technology, Muscat, Sultanate of Oman*
**[1]rajiv.vincent@vit.ac.in, [2]akshat.wagadre2016@vitstudent.ac.in,*
[3]*arunkumar.sivaraman@vit.ac.in, [4]rajesh.m@vit.ac.in, [5]arunrajesh@hct.edu.om*

Abstract

The tracking and understanding the behavior of human beings is an important issue in a number of industrial applications. At the very best level, the system capable of resolving this problem has to be able to recognize the human behavior and understand the motive from that observation alone. This is a difficult task, as people can have different ways of performing certain tasks, so differentiating them properly can be challenging. In this work, a method for the design of a Deep Learning based intelligent human activity monitoring system is proposed, that can detect and track suspicious activity in any surveillance environment. This method makes use of features extracted using modern day image classification algorithms and passing them to sequence based neural network. Being able to perform classification of human activities on a live feed can be helpful in health sectors as well as surveillance systems and prevent disasters.

Keywords: *Long Short Term Memory (LSTM), Bidirectional LSTM, Convolutional Neural Network (CNN), InceptionV3, Feature Extraction*

1. Introduction

In the past few years, there have been massive developments in the field of Artificial Intelligence (AI) and Machine Learning and there are a multiple components in these technologies that are being used to solve many real world problems. AI is like a simulation of human intelligence by the machines which when implemented properly can help millions because of its great scalability. An important component of AI is Computer Vision which can classify and generate essential data from images and videos. One of the tasks that it can perform is Human Activity Recognition (HAR). This is a way to analyze human activities by processing the time series data. In today's world, crimes and illegal activities are prevalent which go unnoticed. There are times when such activities cause a lot of issues. But if there is a detection system which can notify the authorities of a suspicious event, it can save a lot of time and create opportunities for responding to them using a response system. With the help of HAR, we can make it possible. This can be used for tasks like video surveillance systems and medical monitoring and tracking of patients. Most existing activity recognition systems make use of depth sensors to detect motions but in this work, the experiments demonstrate use of LSTM models for RGB only videos to detect human activities. Finally, a system to detect suspicious activities is created.

2. Literature Review

A. Dargazany and M. Nicolescu talks about a real-time human body tracking system using video sequences [1]. The body parts model is generally constituted by

body components like torso, head, legs and arms. The body components are modeled using the location of torso and the size which are obtained through a torso tracking technique in every frame. To track the torso, a blob tracking module is used to seek out the approximate location and size of the torso in each frame. By tracking the torso, it will be able to track other body parts based on their location with respect to the torso on the silhouette that is detected using background separation.

The technique for designing an intelligent human activity monitoring system based on artificial neural networks is proposed by M. K. Fiaz and B. Ijaz that can detect and track unusual activities in any surveillance environment [2]. This method makes use of the silhouette pattern of the human blob obtained from segmentation of the scene using some background separation technique, captured by the camera. This model recognizes anomalies based on the silhouette pattern observed by the Artificial Neural Network (ANN) model. This gives us an idea of extracting features from images and using them.

3. Methodology

The main goal of Activity Recognition system is to identify certain activities from videos. Here a need for capturing the visual information from the activity videos arises. A video is a sequence of frames combined together, so the information extracted needs to have spatial as well as temporal characteristics. This information extracted is called features that we use to feed our model to make predictions. In this work, we study how to extract that information and use them to train out model.

3.1 Model used: As video has temporal characteristics, a model that can retain the extracted features sequentially is required. Here LSTM model [3] is chosen which is a type of Recurrent Neural Network (RNN) that can learn the temporal characteristics of a video from the extracted features. A standard RNN model has the tendency to forget the information that occurred at a long duration of time and this issue is called the vanishing gradients problem. This problem is solved by the LSTM model as it selectively adds and removes information from its memory and becomes capable of learning long-term dependencies.

3.2 Feature Modeling: The extraction of features from the videos is the basis for activity recognition. These features encode the motions of humans in a reduced form. This work uses CNN models to extract the features from each frame of the videos. The need for the model to be deployed on a live setup makes it important that the feature extraction model used is fast and efficient. So a widely used CNN based networks Inception-V3 and Xception, developed by Google are used. These models are under 100 MB and generate features fast enough to capture the temporal characteristics of a video. Both models give a feature vector of dimension (,2048) at the average pooling layer. [Fig. 1] depicts recognition system architecture.

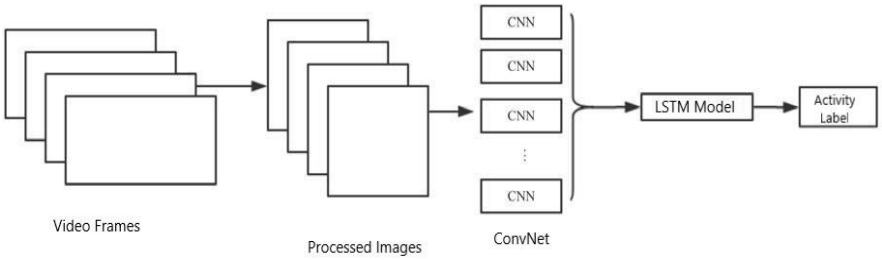

Figure 1. Recognition System Architecture

4. Datasets Used

This work adopts activity datasets from UCF101 and HMDB51. UCF101 is a dataset of 101 human activities collected from YouTube videos created by University of Central Florida. HMDB51 is a large human motion database of 51 activities created by Serre Lab research group from Brown University. [Fig 2a, 2b] shows sample images from the datasets UCF101 and HMDB51.

Figure 2a. UCF101 Data Sample

Figure 2b. HMDB51 Data Sample

5. Implementation

In this section, the preprocessing, optimization used, model creation and deployment is discussed.

5.1 Preprocessing

Five activities from HMDB51 and UCF101 are selected. The activities are Sitting, Waving, Hitting and Kicking from HMDB51 and Walking from UCF101. These are the 5 classes that are used to train the model. Videos from each class are sent to the feature extraction module. Each frame of the videos is converted to 299*299 dimensions and

passed to Inception-V3 and Xception CNN networks. Videos in the datasets have frame rate of 30 fps and each video is trimmed down to 90 frames during feature extraction which gives us up to 3 seconds of features of each video. The produced feature for each class is stored in numpy arrays of size (n, 2048) where n is the total extracted features. After the extraction, numpy arrays of each class are selected such that all the rows except the last 500 are taken and concatenated for the training set. The remaining 500 rows in each class are concatenated and used for the testing set. Simultaneously, output array is created using the video labels.

5.2 Optimization

To an LSTM model, data is fed using batches. Each batch of data is mapped to a certain class. In this work, batches are made of 90 frames and passed as input to the LSTM. Training Data shape: X = (n, 90, 2048), y = (n, num class) Where 'n' is the number of batches.

To improve the training of model, the number of batches is increased by repeating the features in batches multiple times. This is achieved by taking strides of 15 (selected using trial and error) while creating a batch. So a sliding window keeps sliding 15 units and adding 90 feature vectors to the input data and simultaneously adding category vector to the output data.

5.3 Modeling

Two models are built for training. First is the LSTM model which has 4 layers. The first two layers have 1024 and 400 LSTM units respectively and the last two layers are dense layers with 500 and 5 units respectively. Activation used for final layer is softmax. Loss is calculated using the categorical cross entropy function. The second model is similarly built using the BiLSTM layers in the first two layers instead of LSTM layers.

5.4 Model Deployment on Live Feed

The trained LSTM model is saved and exported to a python script. This script makes use of OpenCV library to read video frames from the camera. When the script begins receiving frames, it starts extracting features from each frame and storing it in a numpy array. After the first 90 frames are received, it starts creating batches of 90 features from the feature array. This batch of 90 features is then sent to the LSTM model for classification of activity observed.

6. Results

After training the models on different set of features, the following accuracies were observed. Table 1 shows an example table of shape functions for quadratic line elements.

Table 1. Shape Functions for Quadratic Line Elements

Classification Model	CNN Model	Accuracy
LSTM	Inception − v3	80.12%
BiLSTM	Inception − v3	80.75%
LSTM	Xception	77.02%

6.1 Training/Testing accuracy/epoch plots

[Fig. 3a, 3b] shows plots using Inception-v3 features. [Fig. 3a, 3b] illustrates plots using Xception features.

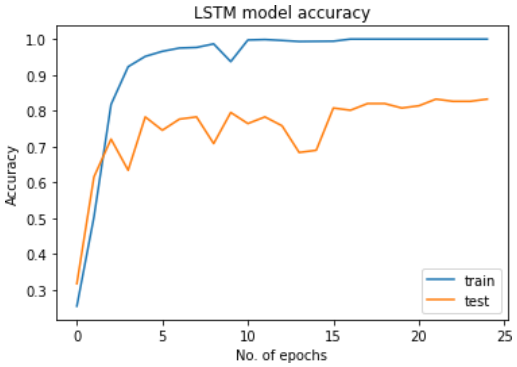

Figure 3a. Label1

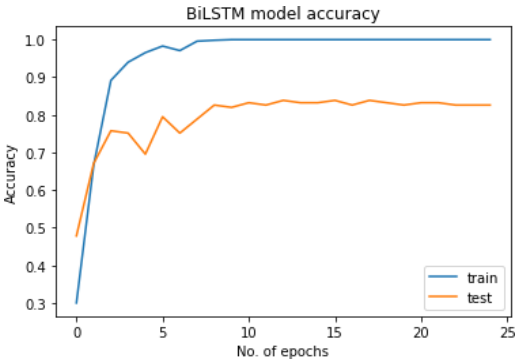

Figure 3b. Label 2

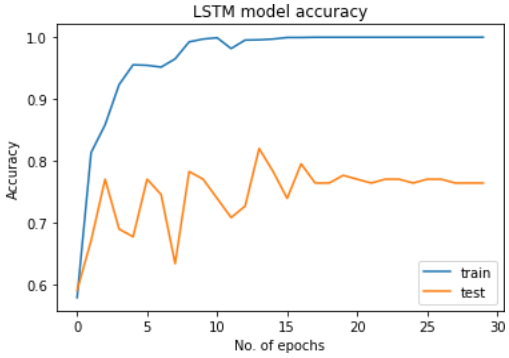

Figure 4a. Label1

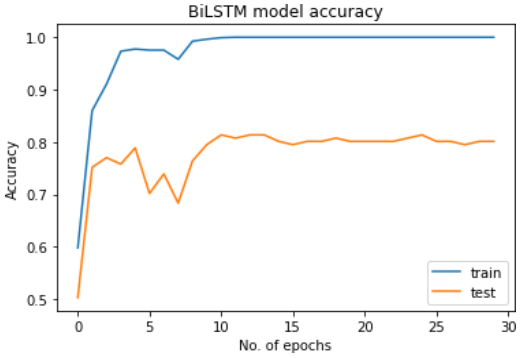

Figure 4b. Label 2

7. Conclusion

In this work, we successfully built a model that automates the task of security personnel and found that the Inception and Xception type models are suitable for activity recognition tasks. The computational complexity for extracting good features and training the models is high and needs powerful machines to be properly built. We also conclude that Bi-LSTMs perform slightly better than LSTMs while training but fall a bit short on faster predictions. For this work, the LSTM model trained using the InceptionV3 features turned out best.

References

[1] A. Dargazany and M. Nicolescu, "Human Body Parts Tracking Using Torso Tracking: Applications to Activity Recognition," *2012 Ninth International Conference on Information Technology - New Generations*, Las Vegas, NV, **(2012)**, pp. 646-651

[2] M. K. Fiaz and B. Ijaz, "Vision based human activity tracking using artificial neural networks," *2010 International Conference on Intelligent and Advanced Systems*, Manila, **(2010)**, pp. 1-5.

[3] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to forget: continual prediction with LSTM," *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, Edinburgh, UK, vol.2, **(1999)**, pp. 850-855

Authors

Prof. Rajiv Vincent is currently working as an Assistant Professor in the School of Computing Science and Engineering, VIT University, Chennai. He has been an academician for the past 10 years and as System administrator for 2 years. He received his Master's in Computer Science and Engineering from College of Engineering Guindy, Anna University, Chennai, India. He is currently pursuing his Ph.D in Computer Science and engineering from VIT University, Chennai, India. His area of interest is Bioinformatics, Machine Learning, Image processing and Web Tecnologies. He has published several scopus indexed research articles.

.

Akshat Wagadre is a B.Tech Graduate from Vellore Institute of Technology. He is a driven student with strong interest in the field of Artificial Intelligence. He has done multiple Machine Learning courses involving Deep Learning specialization from coursera. He has built various ML based projects and is constantly looking to learn and grow in Computer Science research and development.

Dr. Arun Kumar Sivaraman, is a Doctorate holder in Computer Science and Engineering. He has a decade of domestic and international experience in roles such as Assistant Professor, R&D Research Fellow and Information Technologist with excellent knowledge in latest Industry Standard skills. He did his Bachelor Degree (April 2006) in Computer Science & Engineering with first class from Anna University, Chennai and Master Degree (April 2010) in Computer Science & Engineering with first class from College of Engineering, Guindy Campus, Anna University, Chennai. He has been awarded Ph.D (June 2017) in Computer Science & Engineering discipline under Manonmaniam Sundaranor University, Government of India, Tamil Nadu. Furthermore, he has good record in the field of Research and Development by undertaking government granted funding research projects. Moreover, he is a Certified Professional with AMAZON, ALIBABA, GOOGLE, ORACLE, CISCO and IEEE Corporations.

Prof. Rajesh M from VIT University,Chennai, India received his Bachelor of Engineering (B.E) degree in Information Technology from Amrita Institute of Technology and Science, Coimbatore, India and the Master degree (M.Tech.) in Computer Science and Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India. He is currently pursuing his Ph.D in Computer Science and engineering from VIT University, Chennai, India. Prof. Rajesh is having more than fifteen years of experience in academics and research. He is currently working as Assistant Professor

(Sr.), in the School of Computing Science and Engineering,VIT University, Chennai, India. He has published several scopus indexed research articles. His area of interests includes interconnection networks, directed networks, IoT, web technologies, operating systems and programming languages.

Dr. Arun Rajesh, is a Doctorate holder in Computer Science and Engineering. He has 20 Years of domestic and international experience in Academic sector with excellent knowledge in Algorithm Analysis skills. He did his Bachelor Degree (1996) in Computer Science & Engineering with first class from Madras University, Chennai and Master Degree (2002) in Computer Science & Engineering with first class from Annamalai University, India. He has been awarded Ph.D (2017) in Computer Science & Engineering discipline under Manonmaniam Sundaranor University, Government of India, Tamil Nadu. Moreover, he is a Certified Professional with CISCO, RedHat and IEEE Corporations.