Human Activity Recognition System

Priyanshu Jain
Information Technology,
Medi-Caps University
Indore, India
en17it301073@medicaps.ac.in

Samyak Jain
Information Technology,
Medi-Caps University
Indore, India
en17it301083@medicaps.ac.in

Sarthak Ranka
Information Technology,
Medi-Caps University
Indore, India
en17it301085@medicaps.ac.in

Shruti Dhanotiya
Information Technology,
Medi-Caps University
Indore, India
shruti.dhanotiya@medicaps.ac.in

Abstract:Human activity recognition (HAR) has become a popular topic of research because of its wide application. Computers are getting better at solving some very complex problems (like understanding an image) due to the advances in computer vision. Models are being made wherein, if an image is given to the model, it can predict what the image is about, or it can detect whether a particular object is present in the image or not. Here, a deep network architecture using residual bidirectional long short-term memory (LSTM) cells is proposed.Here we uses deep learning for Video Recognition - given a set of labelled videos, train a model so that it can give a label/prediction for a new video.Generally, the proposed network shows improvements on both the temporal (using bidirectional cells) and the spatial (residual connections stacked deeply) dimensions, aiming to enhance the recognition rate.Finally, the confusion matrix of the public domain UCI data set was analyzed.

Key words: human activity recognition; bidirectional LSTM; residual network;K-Nearest Neighbor,convolutional neural network

1. Introduction

Human activity recognition (HAR) is of value in both theoretical research and actual practice. It can be used widely, including in health monitoring [1], smart homes [2], and human–computer interactions [3]; for example, LSTM cells are a good choice for solving HAR problems.

The aim of this project is to create a model that can identify the basic human actions like running, walking, standing, sitting,walking_upstairs,walking_downstairs The model will be given a set of videos where in each video, a person will be performing an action. The label of a video will be the action that is being performed in that particular video. The model will have to learn this relationship, and then it should be able to predict the label of an input (video) that it has never seen. Technically, the model would have to learn to differentiate between various human actions, given some examples of these actions.

Unlike traditional algorithms, LSTM can catch relationships in data on the temporal

dimension without having to mix the time steps together as a 1D CNN would do.

LSTM architecture can offer great performance and many potential applications. A public domain benchmark of HAR has been introduced, and different methods of recognition have been analyzed [4]. The results showed that the K-Nearest Neighbor (KNN) algorithm outperforms other algorithms in most recognition tasks. Unlike the manual filtering features in previous algorithms, a systematic feature learning method that combines feature extraction with CNN training has also been proposed. Subsequently, DeepConvLSTM networks outperformed previous algorithms in the Opportunity Challenge by an average of 4% of the F1 score; the effects of parameters on the final result were also analyzed. Although researchers have made great strides in HAR, room for improvement remains. Inspired by previous neural networks' architectures, we describe a novel Deep Residual Bidirectional Long Short-term Memory LSTM (Deep-Res-Bidir-LSTM) network. The deep LSTM has improved learning ability and, despite the time required to reach maximum accuracy, shows good accuracy early in training; it is especially suitable for complex, large-scale HAR problems where sensor fusion would be required. Residual connections and bidirectional communication through time are available to ensure the integrity of information flowing deeply through the neural network. In recent years, deep learning has shown applicability to many fields, such as image processing [5], speech recognition [6], and natural language processing [7].

2. Related Work:

In ILSVRC 2012, AlexNet [8], proposed by Alex Krizhevsky, took first place, and, since then, deep learning has been considered to be applicable to solving real problems and has done so with impressive accuracy. Indeed, deep learning has become a popular area for scientists and engineers. Another event in 2016 that drew considerable attention was the century man–machine war at the end of the game in which AlphaGo achieved victory. This event also demonstrated that deep learning, based on big data, is a feasible way to solve the non-deterministic polynomial problem. LSTM cells, which were first proposed by Juergen Schmidhuber in 1997 [9], are variants of recurrent neural networks (RNNs). They have special inner gates that allow for consistently better performance than RNN on a time series. Compared with those of other networks, such as CNN, restricted Boltzmann machines (RBM), and auto-encoder (AE), the structure of the LSTM renders it especially good at solving problems involving time series, such as those related to natural language processing, speech recognition, and weather prediction, because its design enables gradients to flow through time readily. Several experiments were performed with HAR benchmarks: the public domain UCI data set and the Opportunity data set. We compare the accuracy of recognition of our algorithm with those of other algorithm. Finally, we summarize the research and discuss our future work.

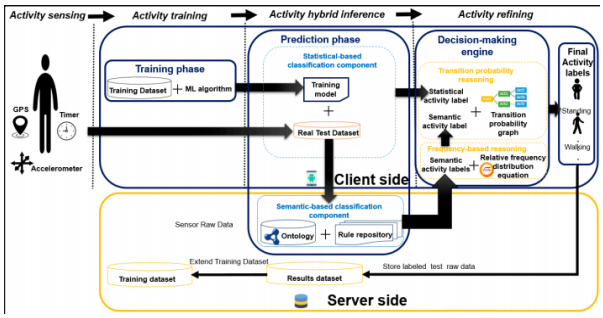

Fig1:Client Server Architecture

3. Methodology

Data pre-processing

1. Reading in the video frame-by-frame.

2. The videos were captured at a frame rate of 25fps. This means that for each second of the video, there will be 25 frames. We know that within a second, a human body does not perform very significant movement. This implies that most of the frames (per second) in our video will be redundant. Therefore, only a subset of all the frames in a video needs to be extracted. This will also reduce the size of the input data which will in turn help the model train faster and can also prevent over-fitting. Different strategies would be used for frame extraction like:

Extracting a fixed number of frames from the total frames in the video – say only the first 200 frames (i.e., first 8 seconds of the video).

Extracting a fixed number of frames each second from the video – say we need only 5 frames per second from a video whose duration is of 10 seconds. This would return a total of 50 frames from the video. This approach is better in the sense that we are extracting the frames sparsely and uniformly from the entire video.

3.Each frame needs to have the same spatial dimensions (height and width). Hence each frame in a video will have to be resized to the required size.

4.In order to simplify the computations, the frames are converted to grayscale.

5.Normalization – The pixel values ranges from 0 to 255. These values would have to be normalized in order to help our model converge faster and get a better performance. Different normalization techniques can be applied such as:

Min-max Normalization – Get the values of the pixels in a given range (say 0 to 1)

Z-score Normalization – This basically determines the number of standard deviations from the mean a data point is.

Implementation:

One of the most important parts of the project was to load the live video dataset and perform the necessary pre-processing steps. So, we developed a class (Videos) that had a function called (read_videos()) that can be used for reading and processing videos. Creating this was very challenging as we concentrated on generalizing this function for any kind of videos (not specific to this project).We have used NumPy (wherever)

for storage and processing of the videos (much faster than in-built python lists with a ton of extra functionalities).The neural network was implemented using Keras.

Experiments

We have tested the LSTM network with the public domain UCI data..Then, we compared it with the outcomes of other methods and analyzed the results. The computer for testing had an i5 CPU with 8 GB RAM as well as an NVIDIA GTX 960m GPU, which has 640 CUDA cores and 8 GB RAM. The GPU and CPU were used alternatively depending on the size of the neural network, which sometimes exceeded the available amount of memory on the graphics card during training.

Data Sets

The research objects of recognition were activities in daily life. Thus, the benchmark for HAR should meet two conditions: first, it should contain most behavioral classes so it reflects real life. Second, it should abstract features and labels for modeling and calculations. Human actions can be divided into several layers such as WALKING, WALKING_UPSTAIRS,WALKING_DOW--NSTAIRS , SITTING , STANDING, JOGGING.A good HAR benchmark should include a clear understanding of the hierarchy.We chose the public domain UCI and the Opportunity data sets for our experiments. The neural network should be readily adaptable to a new data set with an architecture module and a changeable configuration file that also loads the data set. Public domain UCI data set.Each person performed six activities WALKING, WALKING_UPSTAIRS,WALKING_DOWNSTAIRS,SITTING,STANDING, JOGGING The experiments were video-recorded to label the data manually and obtain balanced classes.The dataset obtained was partitioned randomly into two sets: 70% of the dataset were selected for generating the training data, and 30% were selected for generating the test data. Each sample had 561 linear (time-independent) hand-made, preprocessed features from signal analysis (e.g., window's peak frequency), but only six features were used in our study: triaxial gravity acceleration from the accelerometer (from a 0.3 Hz Butterworth low-pass filter) and triaxial body acceleration and triaxial angular velocity from the gyroscope. These are raw signals with a time component and do not fall in the frequency domain but rather in the time domain.

The tasks involved are the following:

Downloading, extracting and pre-processing a video dataset

Dividing the dataset into training and testing data

Create a neural network and train it on the training data.

Test the model on the test data

Compare the performance of the model with some pre-existing models

Metrics: Once the model has been trained on the training data, its performance will be evaluated using the test data. The following metrics will be used:

Accuracy – will be used for evaluating the performance of the model on the test data.

Confusion Matrix - will be used in order to compare the model with the Benchmark model. A confusion matrix is used to describe the performance of a classification model.

Accuracy is the most common evaluation metric used for classification problems. In particular, accuracy is very useful when there are an equal number of samples in each class. Since our dataset have similar characteristics, accuracy would be a suitable metric to evaluate the model.

Benchmark:

The existing models use the notion of local features in space-time to capture and describe local events in a video. The general idea is to describe such events is to define several types of image descriptors over local spatio-temporal neighbourhoods and evaluate these descriptors in the context of recognizing human activities. These points have stable locations in space-time and provide a potential basis for part-based representations of complex motions in video.

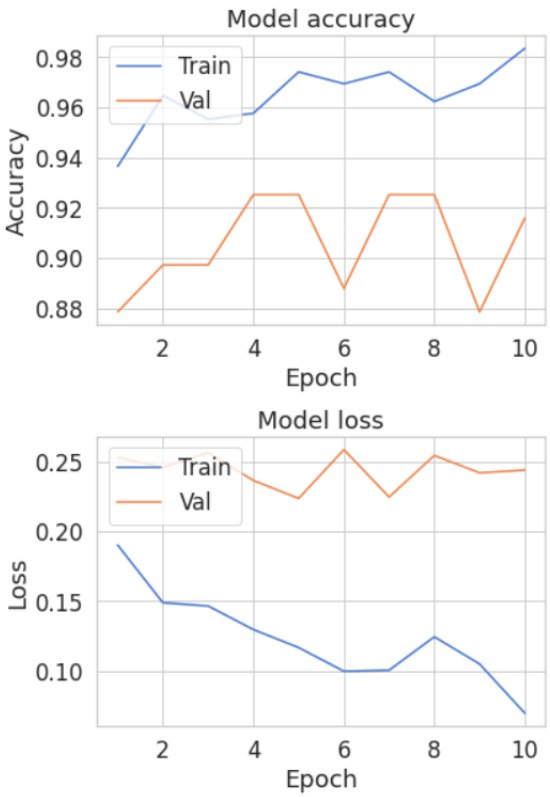

Fig2:Plot curve showing model accuracy and model loss

4.Results

The six different activities were recognized successfully and the accuracy of model is calculated .

5. Conclusions & Future Scope

In this paper, the significance of HAR research is analyzed, and an overview of emerging methods in the field is provided. LSTM neural networks have been used in many innovations in natural language processing, speech recognition, and weather prediction. This technology was adapted to the HAR task. We proposed the novel framework of the Deep-Res-Bidir-LSTM network. This deep network can enhance learning ability for faster learning in early training. In our experiments, the proposed

network was able to improve the accuracy, by 4.78%, for the public domain UCI data set and increase the F1 score, by 3.68%, for the Opportunity data set in comparison with previous work. We also found that window size was a key parameter. It will be important to find an adaptive way to automatically adjust the searching process and also make the neural network's architecture evolve, such as automatically reshaping, adding, and removing various layers. Also, exploring the effect of mixing 1D time-based convolutions at one or some points in the LSTM cells might improve results. Finally, applying the Deep-Res-Bidir-LSTM network to other fields could be revealing. A good model should have outstanding generalization.

6. Acknowledgement

This work was supported by Medicaps University and our guide Ms. Shruti Dhanotiya, Assistant Professor, Medicaps University. We are thankful to our guide and our college for helping us in this journey to reach our ultimate goal.

7.References

[1] Pantelopoulos A, Bourbakis N G. A survey on wearable sensor-based systems for health monitoring and prognosis, IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews. 40(1) (2015) 1–12.

[2] Ahmed S H, Kim D. Named data networking-based smart home, ICT Express. 2(3) (2016) 130–134.

[3] Rautaray S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey, Artificial Intelligence Review. 43(1) (2017) 1–54.

[4] Chavarriaga R, Sagha H, Calatroni A, et al. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognition Letters. 34(15) (2016) 2033–2042.

[5] Dong C, Loy C, He K, et al. Learning a deep convolutional network for image super-resolution, in: European Conference on Computer Vision, 2018.

[6]Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing. 2017.

[7]Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional Neural Networks for Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing. 22(22) (2014) 1533–1545.

[8]Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, 2012.

[9]Hochreiter, Sepp, and Jürgen Schmidhuber. Long short-term memory, Neural computation. 9(8) (1997) 1735–1780.

[10]Anguita, Davide, et al. A Public Domain Dataset for Human Activity Recognition using Smartphones, in: ESANN, 2016.

[11]Maclean W J. Spatial Coherence for Visual Motion Analysis, Springer Berlin Heidelberg, 2016.