# Human Behavior Prediction for Smart Homes Using Deep Learning

Jingwei Guo, Jinlong Hong, Zhihe Zhao
Department of Computer Science
Supervisor: Dr. Wenjin Lu, Dr. Bailing Zhang

SURF
Summer Undergraduate Research Fellowship

**Abstract --** This project studied the application of attention based Long Short-Term Memory neural network model, first introduced via the application on Image Caption, on the recognition of moving action in videos. The essence of this model lays on determining the "action-label" of the each frame of videos and selecting the merged label with the highest frequency of occurrence as the final outcome. In this study, we implemented the application based on TensorFlow platform, and the empirical accuracy of 60% was found, indicating the success of this application.

## Introduction

Because of the aged dependency ratio, a ratio of the number of persons aged over 65 to the number of persons aged between 15 to 64, is projected to rise significantly due to low fertility rates , smart homes have caused a great interest for many researchers with the aim of assisting occupants for increasing the quality of life. The resulting need for development of such technologies is underscored by the aging of the population and the importance that individuals place on remaining independent in their own homes.

Before smart homes technologies can be deployed for occupants, several challenges should be solved, including the data collection, activity recognition algorithms, etc. In the future, this technology can be used wildly if the accuracy is higher enough.

## Methdology

In the ILSVRC competition in 2014, several successful architectures were proposed for object recognition, such as GoogLeNet and VGGNet. To create the annotations used by our decoder, we used the Oxford VGG19, the pre-trained model, on the dataset. The VGG19 was originally developed for object recognition and detection. It has very deep convolutional architectures with smaller sizes of convolutional kernel, stride and pooling window that have turned out to be effective for object recognition in still images.
For feeding the images, the video frames, to the VGGNet, we resized these images to the fixed size of 224*224.

| ConvNet Configuration | | | | | |
|---|---|---|---|---|---|
| A | A-LRN | B | C | D | E |
| 11 weight layers | 11 weight layers | 13 weight layers | 16 weight layers | 16 weight layers | 19 weight layers |
| input (224 × 224 RGB image) | | | | | |
| conv3-64 | conv3-64 **LRN** | conv3-64 **conv3-64** | conv3-64 conv3-64 | conv3-64 conv3-64 | conv3-64 conv3-64 |
| maxpool | | | | | |
| conv3-128 | conv3-128 | conv3-128 **conv3-128** | conv3-128 conv3-128 | conv3-128 conv3-128 | conv3-128 conv3-128 |
| maxpool | | | | | |
| conv3-256 conv3-256 | conv3-256 conv3-256 | conv3-256 conv3-256 | conv3-256 conv3-256 **conv1-256** | conv3-256 **conv3-256** | conv3-256 conv3-256 conv3-256 **conv3-256** |
| maxpool | | | | | |
| conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 **conv1-512** | conv3-512 **conv3-512** | conv3-512 conv3-512 conv3-512 **conv3-512** |
| maxpool | | | | | |
| conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 | conv3-512 conv3-512 **conv1-512** | conv3-512 **conv3-512** | conv3-512 conv3-512 conv3-512 **conv3-512** |
| maxpool | | | | | |
| FC-4096 | | | | | |
| FC-4096 | | | | | |
| FC-1000 | | | | | |
| soft-max | | | | | |

Table 1: ConvNet configuration. The depth of the configurations increases from the left (A) to the right (E) as more layers are added. The D (VGG16) and E (VGG19) show better performance than others.

### LSTM model

To keep the persistence of learned information and connect information from further back. The Long Short Term Memory networks (LSTMs) is designed by scientists to solve this problem. Old information cloud flow through the cells with only minor linear interactions. New information is dropped or added by several gates by the previous condition Ct-1. Each step of LSTM cell is selected by tanh function to output correct related information.

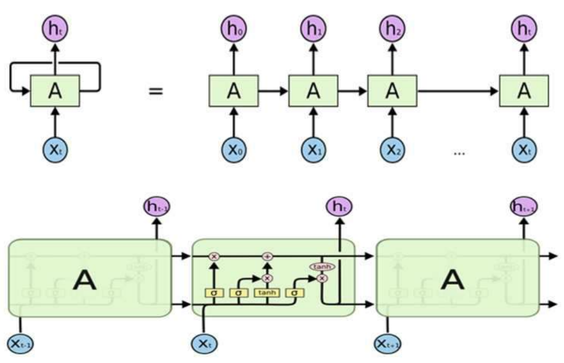


fig1. structure of LSTM

### Attention model

Our project also applied attention model for more accurate results. The attention mechanism helps our model more focuses on the relevant places in each frame. At each time-step t, our model predicts the next location probabilities Lt+1, a softmax over K ×K locations by equation (1).

$$l_{t,i} = p(\mathbf{L}_t = i|\mathbf{h}_{t-1}) = \frac{\exp(W_i^\top \mathbf{h}_{t-1})}{\sum_{j=1}^{K\times K}\exp(W_j^\top \mathbf{h}_{t-1})}$$

In general, we add the Lt to the feature slice Xt generated from VGG model as input xt according to equation(2).

$$\mathbf{x}_t = \mathbb{E}_{p(\mathbf{L}_t|\mathbf{h}_{t-1})}[\mathbf{X}_t] = \sum_{i=1}^{K^2} l_{t,i}\mathbf{X}_{t,i},$$

The model propagates xt through LSTMs and predicts the next location probabilities Lt+1 and the class label yt, a softmax over the label classes with an additional hidden layer. The mechanism for defining a class of a video is to count the most common label of each frame.

## Experiment and Result:

The data, involved in this study, is derived from the benchmark on the Internet and includes 1248 videos forming 10 categories of indoor actions including ‘carry’, ‘pick up with one hand’, ‘pick up with two hands’, ‘drop trash’, ‘walk around’, ‘sit down’, ‘stand up’, ‘donning’, ‘doffing’, and ‘throw’. Additionally, the videos was separated into three data set including train, validation, and test following the proportion of 2:1:1.
In this study, we empirically set the step in the LSTM model as 17, so 17 frames for each video are in need. Considering the heterogeneous distribution of the number of frames videos contain, we proposed an efficient and disinterested approach unitizing the frames of videos intended for the further experimental procedures.

fig2. Example of video frames.

First, we extract the frames of each videos following the proposed method.
Then, the features of images will be extracted based on CNN-vgg19 model.
Third, the training process will be conducted based on the extracted features and fixed labels, and models will be saved during the optimization.
Fourth, we will select the model with highest performance based on the validation data set.As showed in fig4, the performance of the models along epochs surged first, fluctuated during a period, and showed the declining tendency in the end of the epoch.
Besides, the models with highest performance achieving the general accuracy of 63.16%, after 430th epoch, is tested on the test data and attained the general accuracy of 59.99%, indicating the success of this application.
Last, the performance of the selected model will be tested on the test data set.
Besides, in this study, the batch size and the learning rate is set as 15 and 0.006 empirically.

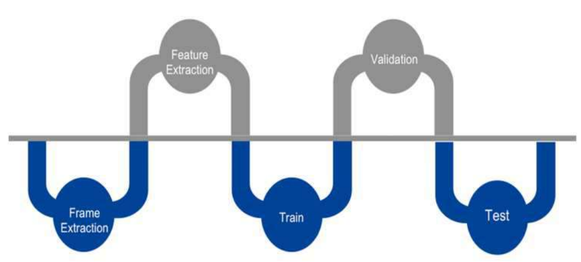


fig3. The procedure of processing data

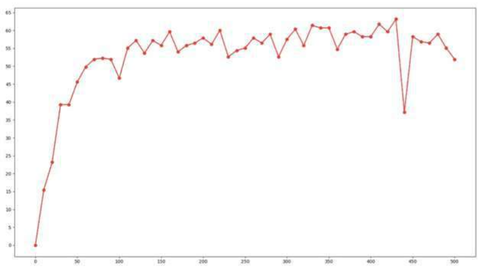

fig4. Results of our model

## Conclusion

In this study, we successfully applied the attention based LSTM model on the recognition of actions in videos. However, there are still some issues blocking the accuracy of the test result. In order to locate the possible issues and improve the efficiency of the designed model, the future work will focus on first increasing the 'time step' of LSTM, second changing the single-layer LSTM into multilayer LSTM and third changing the tradition optimization algorithms to improve the speed and accuracy of optimization procedure.

## Acknowage and Reference

Ackowage:
The authors would like to thank Xi’ an Jiaotong-Liverpool University for the support of Summer Undergraduate Research Fellowship (SURF) and the laboratory for the experiment.

Reference :
Xu, K, Ba, J, Kiros, R, Cho, K, Courville, A, Salakhutdinov, R, Zemel, R, & Bengio, Y 2015, 'Show, Attend and Tell: Neural Image Caption Generation with Visual Attention', arXiv, EBSCOhost, viewed 31 August 2017.
Sharma, S, Kiros, R, & Salakhutdinov, R 2015, 'Action Recognition using Visual Attention', arXiv, EBSCOhost, viewed 31 August 2017.
https://github.com/yunjey/show-attend-and-tell
http://users.eecs.northwestern.edu/~jwa368/my_data.html