# Report on project submission 3 for CSCE 636- Spring 2021

**Bahareh Alizadeh Kharazi**

**March 11, 2021**

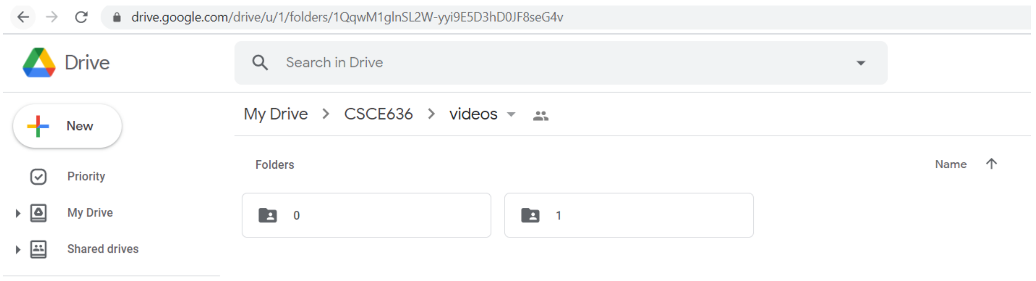
## Abstract

In this project, the goal is to detect the action of shaking head in video clips. A transfer learning approach is proposed in this project which was developed based on using ResNet50 and replacing its first layer with one LSTM layer and three other dens layers. A dataset is created by the author which 50% of it is labeled as shaking head. The test accuracy of this model on the 30% of the entire dataset was achieved as 77% which is acceptable. However, the limited amount of videoclips in the dataset, still questions the generalizability of the model.

## Collecting training dataset

The topic selected for this project was "Shaking-head". To create a dataset for training the model, I had to search for videoclips that have the action "shaking-head" in one clip. Since most videos available on the web have different actions in one clip, I decided to find GIF files (The Graphics Interchange Format) that are commonly used file types in social media that focuses on one action at a time. This trick helped me to train my model easily on the dataset with limited amount of data for my first submission.

For the first submission, I collected 50 photos containing shaking-head action and 50 photos containing other actions. To organize the data I named all vidoes with shaking-head action and videos with non-shaking-head action as "shakinghead_number" and "other_number", respectively. All these folders were moved to Google Drive folders to be used in Google Collab.


drive.google.com/drive/u/1/folders/1QqwM1glnSL2W-yyi9E5D3hD0JF8seG4v

My Drive > CSCE636 > videos

Folders

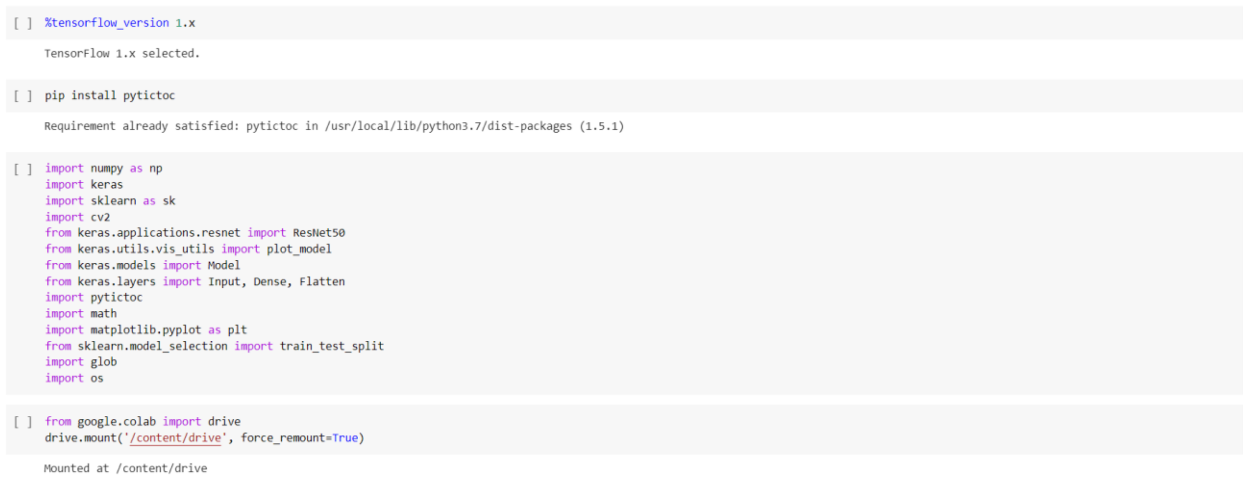
0

1

Folder 0 contains all videos with other actions and folder 1 contains videos with shaking-head action.

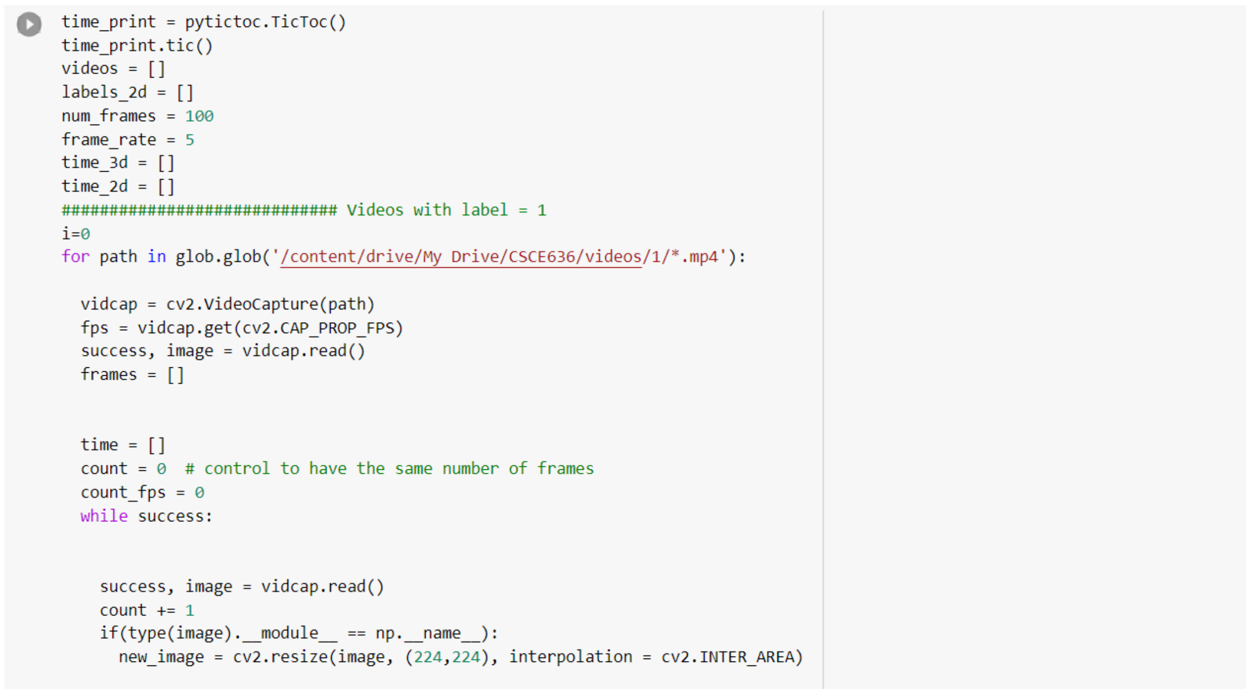
## Preprocessing of the dataset

To prepare videoclips for my model, first I load the modules that I need. I also, mount the Google Drive in the Google Collab:

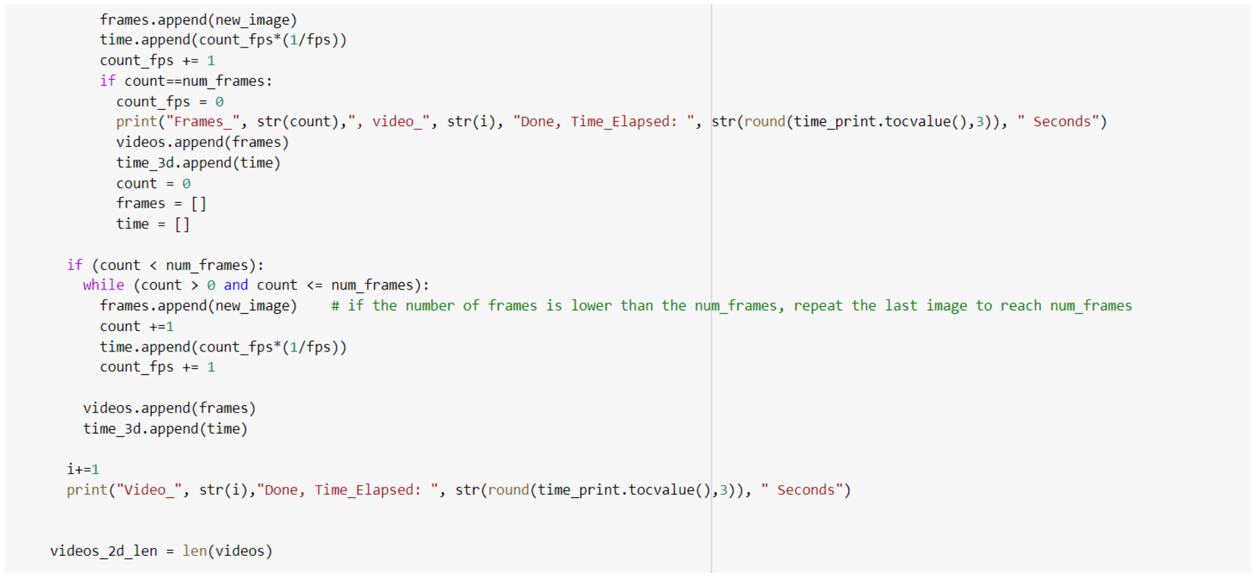


```
[ ] %tensorflow_version 1.x

    TensorFlow 1.x selected.

[ ] pip install pytictoc

    Requirement already satisfied: pytictoc in /usr/local/lib/python3.7/dist-packages (1.5.1)

[ ] import numpy as np
    import keras
    import sklearn as sk
    import cv2
    from keras.applications.resnet import ResNet50
    from keras.utils.vis_utils import plot_model
    from keras.models import Model
    from keras.layers import Input, Dense, Flatten
    import pytictoc
    import math
    import matplotlib.pyplot as plt
    from sklearn.model_selection import train_test_split
    import glob
    import os

[ ] from google.colab import drive
    drive.mount('/content/drive', force_remount=True)

    Mounted at /content/drive
```

Then videos in each folder will be converted to frames (as image) and saved to other folder. The process for videos with label 1 is shown below:

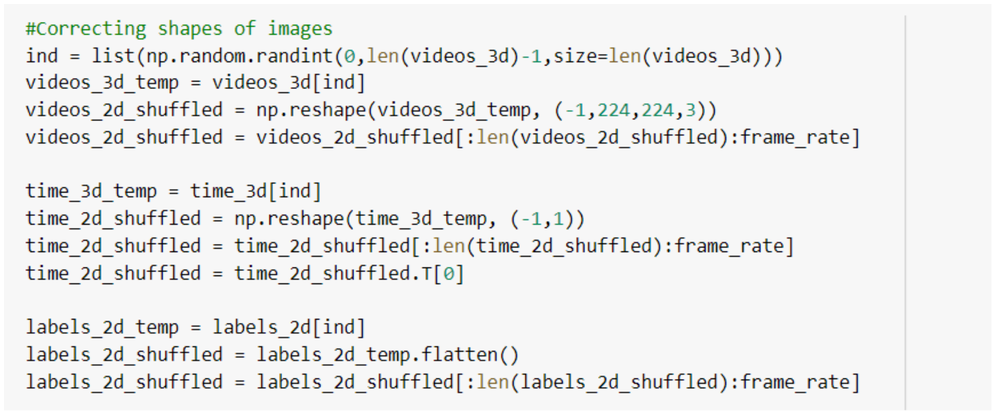


```
time_print = pytictoc.TicToc()
time_print.tic()
videos = []
labels_2d = []
num_frames = 100
frame_rate = 5
time_3d = []
time_2d = []
############################## Videos with label = 1
i=0
for path in glob.glob('/content/drive/My Drive/CSCE636/videos/1/*.mp4'):

  vidcap = cv2.VideoCapture(path)
  fps = vidcap.get(cv2.CAP_PROP_FPS)
  success, image = vidcap.read()
  frames = []


  time = []
  count = 0  # control to have the same number of frames
  count_fps = 0
  while success:


    success, image = vidcap.read()
    count += 1
    if(type(image).__module__ == np.__name__):
      new_image = cv2.resize(image, (224,224), interpolation = cv2.INTER_AREA)
```



```
      frames.append(new_image)
      time.append(count_fps*(1/fps))
      count_fps += 1
      if count==num_frames:
        count_fps = 0
        print("Frames_", str(count),", video_", str(i), "Done, Time_Elapsed: ", str(round(time_print.tocvalue(),3)), " Seconds")
        videos.append(frames)
        time_3d.append(time)
        count = 0
        frames = []
        time = []

  if (count < num_frames):
    while (count > 0 and count <= num_frames):
      frames.append(new_image)    # if the number of frames is lower than the num_frames, repeat the last image to reach num_frames
      count +=1
      time.append(count_fps*(1/fps))
      count_fps += 1

    videos.append(frames)
    time_3d.append(time)

  i+=1
  print("Video_", str(i),"Done, Time_Elapsed: ", str(round(time_print.tocvalue(),3)), " Seconds")


videos_2d_len = len(videos)
```

Then all images will be reshaped to a single shape size.



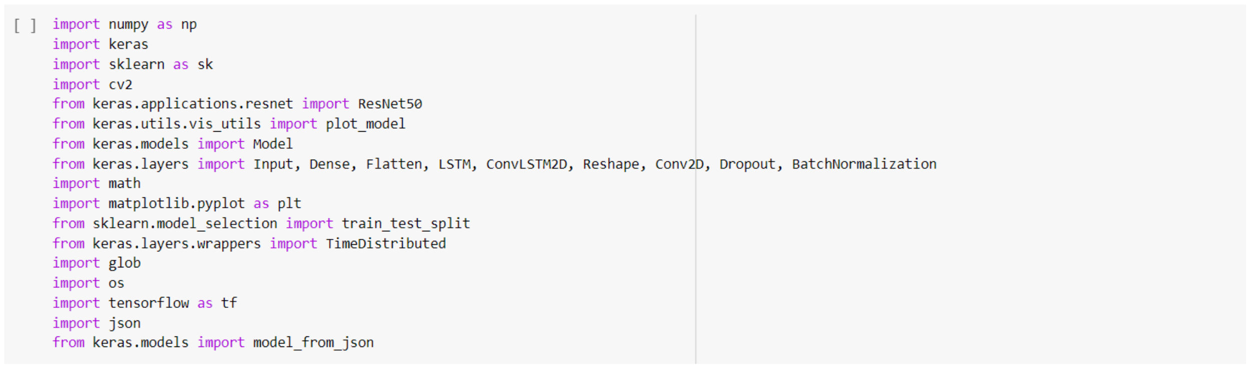
```
#Correcting shapes of images
ind = list(np.random.randint(0,len(videos_3d)-1,size=len(videos_3d)))
videos_3d_temp = videos_3d[ind]
videos_2d_shuffled = np.reshape(videos_3d_temp, (-1,224,224,3))
videos_2d_shuffled = videos_2d_shuffled[:len(videos_2d_shuffled):frame_rate]

time_3d_temp = time_3d[ind]
time_2d_shuffled = np.reshape(time_3d_temp, (-1,1))
time_2d_shuffled = time_2d_shuffled[:len(time_2d_shuffled):frame_rate]
time_2d_shuffled = time_2d_shuffled.T[0]

labels_2d_temp = labels_2d[ind]
labels_2d_shuffled = labels_2d_temp.flatten()
labels_2d_shuffled = labels_2d_shuffled[:len(labels_2d_shuffled):frame_rate]
```

Since processing all images every time is time consuming, I saved all the frames as numpy file to speed up the process for further work. Although currently the number of training image in our dataset is limited, but this approach will save time when my dataset gets larger later.

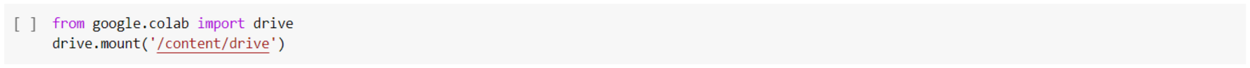
## Selecting a model

I used transfer learning on ResNet50 by excluding the first layer and adding four other layers to it (one ConvLSTM2D and three dense layers).

First we load the modules:

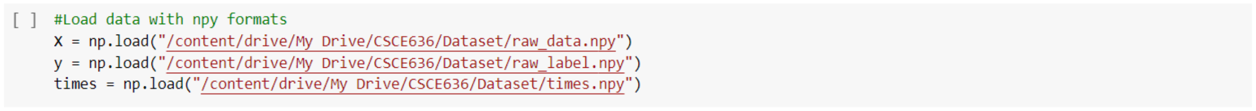


```python
import numpy as np
import keras
import sklearn as sk
import cv2
from keras.applications.resnet import ResNet50
from keras.utils.vis_utils import plot_model
from keras.models import Model
from keras.layers import Input, Dense, Flatten, LSTM, ConvLSTM2D, Reshape, Conv2D, Dropout, BatchNormalization
import math
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from keras.layers.wrappers import TimeDistributed
import glob
import os
import tensorflow as tf
import json
from keras.models import model_from_json
```



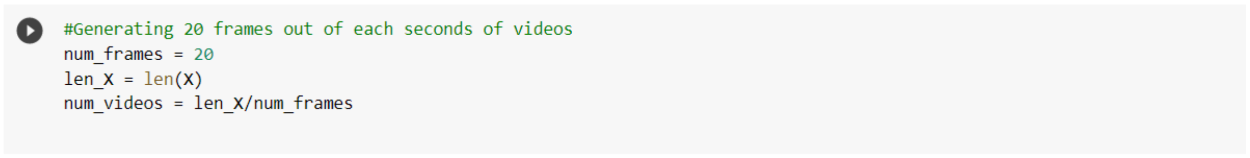
```python
from google.colab import drive
drive.mount('/content/drive')
```

Then we load the numpy data that we created in the previous section:



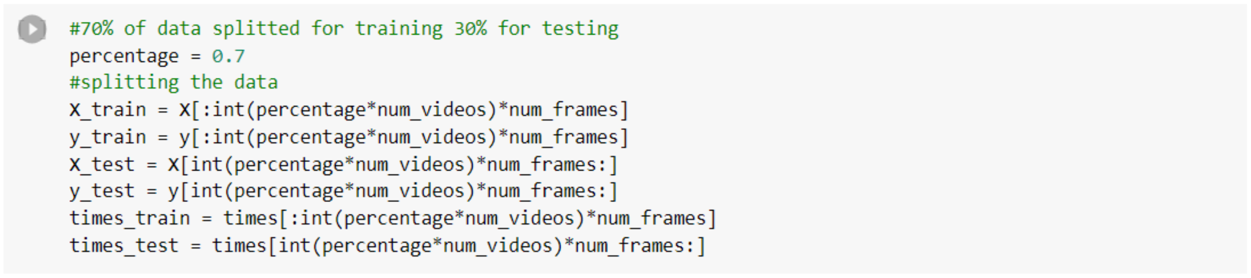
```python
#Load data with npy formats
X = np.load("/content/drive/My Drive/CSCE636/Dataset/raw_data.npy")
y = np.load("/content/drive/My Drive/CSCE636/Dataset/raw_label.npy")
times = np.load("/content/drive/My Drive/CSCE636/Dataset/times.npy")
```

Each second in the videoclips is going to be divided to 20 frames:



```python
#Generating 20 frames out of each seconds of videos
num_frames = 20
len_X = len(X)
num_videos = len_X/num_frames
```

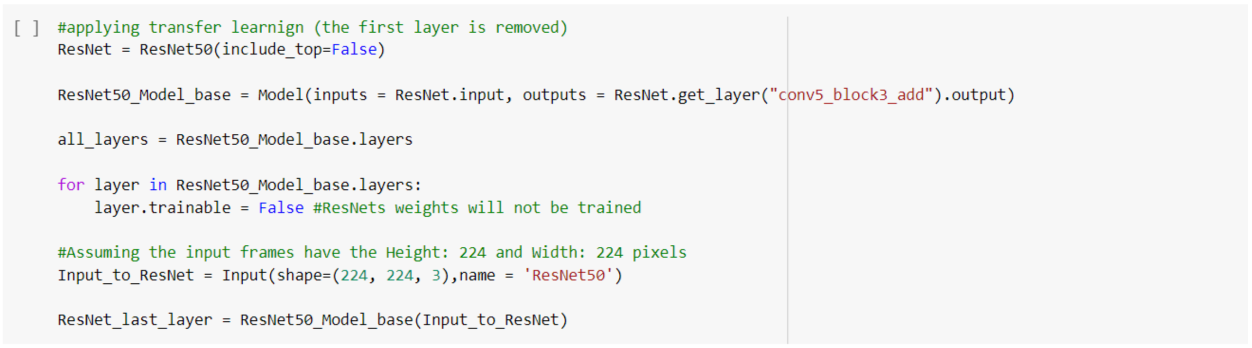
Also, we took 70% of the dataset as the training set and the rest for the test set.



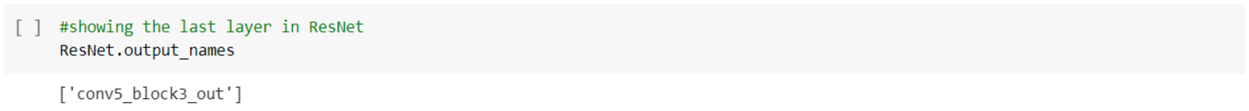
```python
#70% of data splitted for training 30% for testing
percentage = 0.7
#splitting the data
X_train = X[:int(percentage*num_videos)*num_frames]
y_train = y[:int(percentage*num_videos)*num_frames]
X_test = X[int(percentage*num_videos)*num_frames:]
y_test = y[int(percentage*num_videos)*num_frames:]
times_train = times[:int(percentage*num_videos)*num_frames]
times_test = times[int(percentage*num_videos)*num_frames:]
```

## Training and testing the model on the dataset

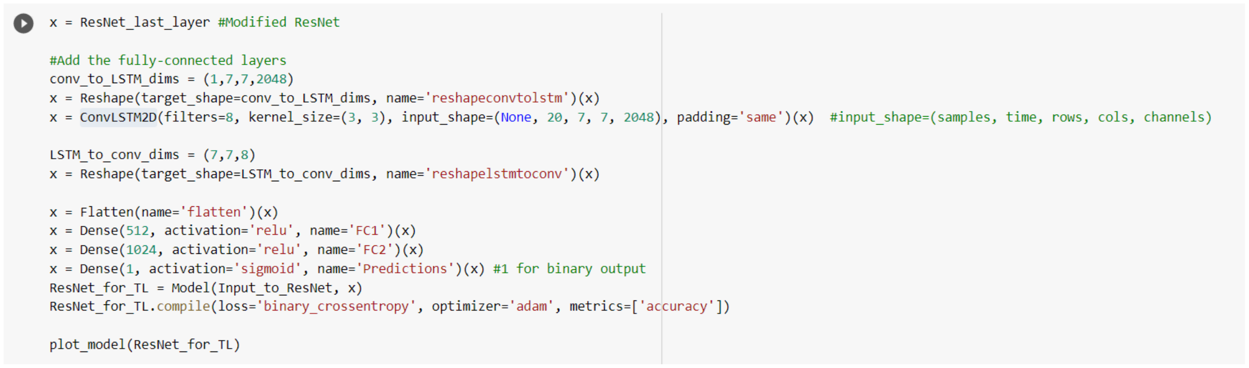
Then we apply transfer learning as stated before, without training the wights of 49 layers of the ResNet model. We only train the weights for the added four layers.


```
[ ] #applying transfer learnign (the first layer is removed)
    ResNet = ResNet50(include_top=False)

    ResNet50_Model_base = Model(inputs = ResNet.input, outputs = ResNet.get_layer("conv5_block3_add").output)

    all_layers = ResNet50_Model_base.layers

    for layer in ResNet50_Model_base.layers:
        layer.trainable = False #ResNets weights will not be trained

    #Assuming the input frames have the Height: 224 and Width: 224 pixels
    Input_to_ResNet = Input(shape=(224, 224, 3),name = 'ResNet50')

    ResNet_last_layer = ResNet50_Model_base(Input_to_ResNet)
```

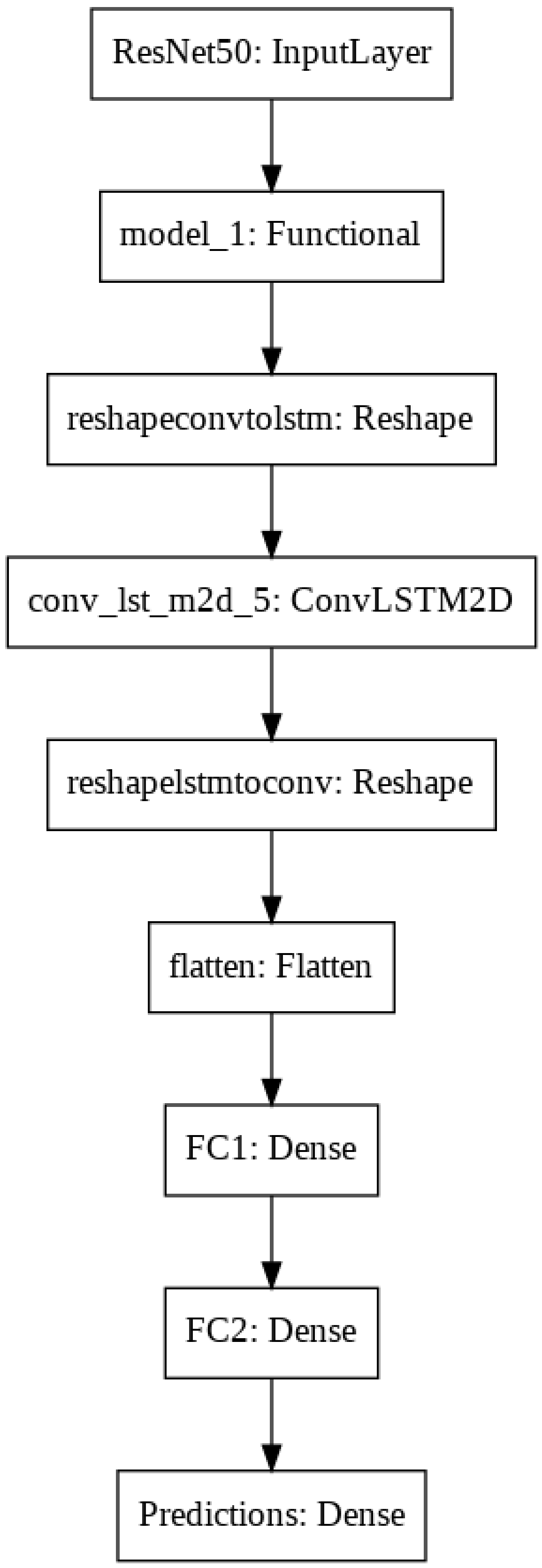
We check the name of the last layer in ResNet (below) and add it as the output in the code above:


```
[ ] #showing the last layer in ResNet
    ResNet.output_names

    ['conv5_block3_out']
```

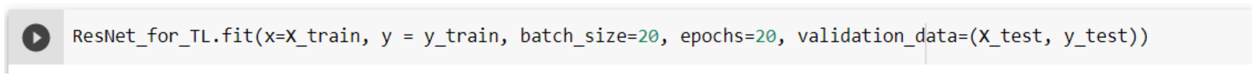
Here, four layers are added to the ResNet model. The first one is ConvLSTM2D and the other three are three dens layers (two with relu active function and one with sigmoid active function). The last layer has one node because our output is binary (0 or 1).


```
x = ResNet_last_layer #Modified ResNet

#Add the fully-connected layers
conv_to_LSTM_dims = (1,7,7,2048)
x = Reshape(target_shape=conv_to_LSTM_dims, name='reshapeconvtolstm')(x)
x = ConvLSTM2D(filters=8, kernel_size=(3, 3), input_shape=(None, 20, 7, 7, 2048), padding='same')(x)  #input_shape=(samples, time, rows, cols, channels)

LSTM_to_conv_dims = (7,7,8)
x = Reshape(target_shape=LSTM_to_conv_dims, name='reshapelstmtoconv')(x)

x = Flatten(name='flatten')(x)
x = Dense(512, activation='relu', name='FC1')(x)
x = Dense(1024, activation='relu', name='FC2')(x)
x = Dense(1, activation='sigmoid', name='Predictions')(x) #1 for binary output
ResNet_for_TL = Model(Input_to_ResNet, x)
ResNet_for_TL.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

plot_model(ResNet_for_TL)
```

The plot of the model is shown below:


ResNet50: InputLayer
model_1: Functional
reshapeconvtolstm: Reshape
conv_lst_m2d_5: ConvLSTM2D
reshapelstmtoconv: Reshape
flatten: Flatten
FC1: Dense
FC2: Dense
Predictions: Dense

Now we fit the ResNet model on our dataset with batch_size=20 and epoch=20. The training accuracy is achieved 99.75% and the test accuracy is obtained 82.64%. Because we have limited amount of data for the first submission, the accuracy varies a lot. By adding more data, our model will be more robust in terms of its accuracy.



```
ResNet_for_TL.fit(x=X_train, y = y_train, batch_size=20, epochs=20, validation_data=(X_test, y_test))
```

```
Epoch 1/20
81/81 [==============================] - 8s 74ms/step - loss: 0.1359 - accuracy: 0.9370 - val_loss:
0.6610 - val_accuracy: 0.8681
Epoch 2/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0189 - accuracy: 0.9938 - val_loss:
0.8774 - val_accuracy: 0.8403
Epoch 3/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0245 - accuracy: 0.9895 - val_loss:
1.4699 - val_accuracy: 0.7264
Epoch 4/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0369 - accuracy: 0.9883 - val_loss:
0.8230 - val_accuracy: 0.8153
Epoch 5/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0167 - accuracy: 0.9938 - val_loss:
0.9994 - val_accuracy: 0.8611
Epoch 6/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0408 - accuracy: 0.9883 - val_loss:
1.0026 - val_accuracy: 0.8389
Epoch 7/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0230 - accuracy: 0.9901 - val_loss:
0.9920 - val_accuracy: 0.8306
Epoch 8/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0157 - accuracy: 0.9957 - val_loss:
1.0627 - val_accuracy: 0.8347
Epoch 9/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0093 - accuracy: 0.9975 - val_loss:
1.2372 - val_accuracy: 0.8361
Epoch 10/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0067 - accuracy: 0.9975 - val_loss:
1.4526 - val_accuracy: 0.8347
Epoch 11/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0105 - accuracy: 0.9975 - val_loss:
1.0312 - val_accuracy: 0.8236
Epoch 12/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0086 - accuracy: 0.9975 - val_loss:
1.4540 - val_accuracy: 0.8417
Epoch 13/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0099 - accuracy: 0.9951 - val_loss:
2.5456 - val_accuracy: 0.7361
Epoch 14/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0678 - accuracy: 0.9846 - val_loss:
0.8886 - val_accuracy: 0.8069
Epoch 15/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0336 - accuracy: 0.9895 - val_loss:
1.1932 - val_accuracy: 0.8417
Epoch 16/20
```

```
81/81 [==============================] - 5s 59ms/step - loss: 0.0709 - accuracy: 0.9753 - val_loss:
0.7298 - val_accuracy: 0.8153
Epoch 17/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0135 - accuracy: 0.9957 - val_loss:
0.6954 - val_accuracy: 0.8500
Epoch 18/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0096 - accuracy: 0.9969 - val_loss:
0.9194 - val_accuracy: 0.8389
Epoch 19/20
81/81 [==============================] - 5s 59ms/step - loss: 0.0102 - accuracy: 0.9963 - val_loss:
1.2448 - val_accuracy: 0.8208
Epoch 20/20
81/81 [==============================] - 5s 60ms/step - loss: 0.0091 - accuracy: 0.9975 - val_loss:
0.8940 - val_accuracy: 0.8264
<tensorflow.python.keras.callbacks.History at 0x7f33ebea6390>
```

The plot of the ResNet model's base is shown below (you need to zoom in to the picture):

## Testing the model on video clips on youtube

Now that we have our model trained on our dataset, we can test our model on new videoclips. The process of testing the model is elaborated in the following.

We download YouTube videoclips and move them into the folder in the Google Drive and set the test path to its directory:

```
path_test = "/content/drive/My Drive/CSCE636/Youtube_test/test2.mp4"
```

Again, we convert the video into frames and save them as image files. We randomly select frames out of each video.

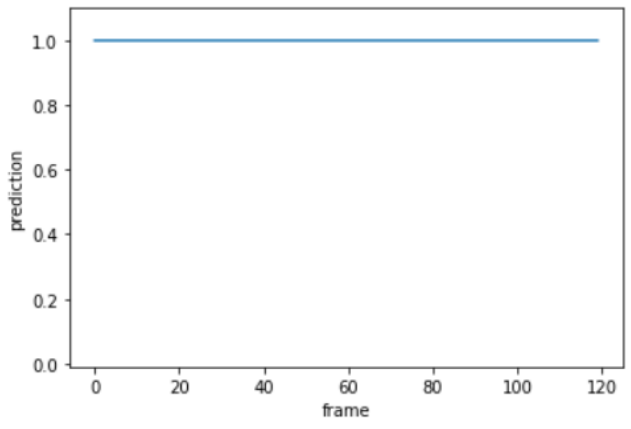
Here we call the model to predict the label of frames in the test data:

```
[ ]  u_test = ResNet50_loaded.predict(videos_2d_shuffled)
```

And then we plot the predicted label for each frame in the video. For this sample video, the shaking-head action is detected which was correct. For each video lips in the test set, we repeat the same process and then we upload true positive samples in the ilab website.

```
plt.plot(u_test)
plt.ylim(-.01,1.1)
plt.ylabel('prediction')
plt.xlabel('frame number')
```

```
Text(0.5, 0, 'frame')
```



## Project submission 3 (Date: 03/11/2021)

### Collecting training dataset

In the third submission of the project I changed my topic to Cooking/Cutting. I collected 18 long video files (total size was 57.3 MB) that contained cutting activity. Some video clips were collected from a publicly available dataset and other videos were collected from YouTube.

### Preprocessing of the dataset

The preprocessing section was not changed comparing to the previous submission. All videos were uploaded to the Google Colab frames were extracted from videos and saved as numpy files.

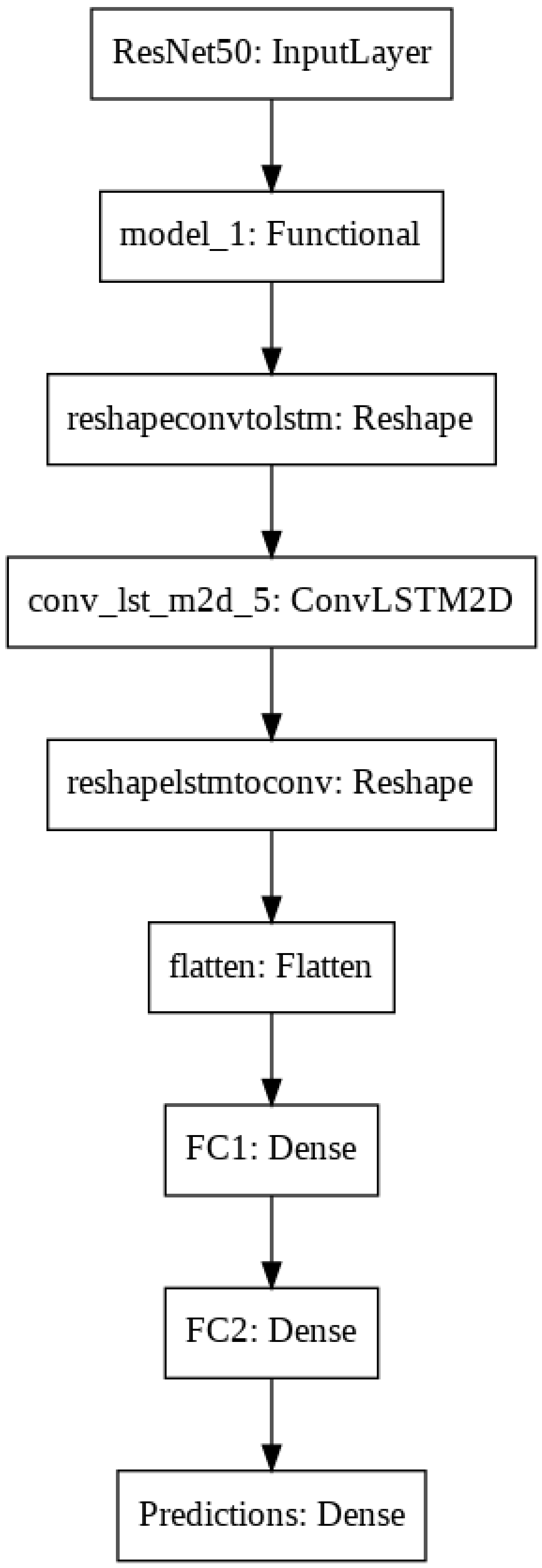
### Selecting a model

A transfer learning is used on ResNet50 by excluding the first layer and adding four other layers to it (one ConvLSTM2D and three dense layers).

Each second in the videoclips is going to be divided to 20 frames:

Also, we took 70% of the dataset as the training set and the rest for the test set.Then we apply transfer learning as stated before, without training the wights of 49 layers of the ResNet model. We only train the weights for the added four layers.

Here, four layers are added to the ResNet model. The first one is ConvLSTM2D and the other three are three dens layers (two with relu active function and one with sigmoid active function). The last layer has one node because our output is binary (0 or 1).


ResNet50: InputLayer
model_1: Functional
reshapeconvtolstm: Reshape
conv_lst_m2d_5: ConvLSTM2D
reshapelstmtoconv: Reshape
flatten: Flatten
FC1: Dense
FC2: Dense
Predictions: Dense

```
Train on 4120 samples, validate on 1780 samples
Epoch 1/20
4120/4120 [==============================] - 14s 3ms/step - loss: 0.0046 - accuracy: 0.9988 - val_loss:
0.5636 - val_accuracy: 0.8994
Epoch 2/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0655 - accuracy: 0.9789 - val_loss:
1.1785 - val_accuracy: 0.7573
Epoch 3/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0646 - accuracy: 0.9748 - val_loss:
0.3715 - val_accuracy: 0.8646
Epoch 4/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0318 - accuracy: 0.9905 - val_loss:
0.7867 - val_accuracy: 0.8399
Epoch 5/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0208 - accuracy: 0.9934 - val_loss:
0.9298 - val_accuracy: 0.8449
Epoch 6/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0120 - accuracy: 0.9959 - val_loss:
1.8405 - val_accuracy: 0.7854
Epoch 7/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0110 - accuracy: 0.9961 - val_loss:
1.3531 - val_accuracy: 0.8067
Epoch 8/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0234 - accuracy: 0.9910 - val_loss:
1.9183 - val_accuracy: 0.8000
Epoch 9/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0300 - accuracy: 0.9913 - val_loss:
0.9928 - val_accuracy: 0.6657
Epoch 10/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0319 - accuracy: 0.9915 - val_loss:
0.6522 - val_accuracy: 0.8034
Epoch 11/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0392 - accuracy: 0.9874 - val_loss:
1.1651 - val_accuracy: 0.7264
Epoch 12/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0122 - accuracy: 0.9951 - val_loss:
1.4307 - val_accuracy: 0.6056
Epoch 13/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0073 - accuracy: 0.9973 - val_loss:
1.5495 - val_accuracy: 0.5876
Epoch 14/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0207 - accuracy: 0.9910 - val_loss:
0.7260 - val_accuracy: 0.7449
Epoch 15/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0112 - accuracy: 0.9968 - val_loss:
1.2824 - val_accuracy: 0.8157
Epoch 16/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0024 - accuracy: 0.9995 - val_loss:
1.0874 - val_accuracy: 0.8393
Epoch 17/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0026 - accuracy: 0.9988 - val_loss:
1.8460 - val_accuracy: 0.8202
Epoch 18/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0034 - accuracy: 0.9990 - val_loss:
1.4922 - val_accuracy: 0.8213
Epoch 19/20
```

```
4120/4120 [==============================] - 13s 3ms/step - loss: 8.0434e-04 - accuracy: 1.0000 -
val_loss: 1.6618 - val_accuracy: 0.8118
Epoch 20/20
4120/4120 [==============================] - 13s 3ms/step - loss: 0.0215 - accuracy: 0.9915 - val_loss:
0.2597 - val_accuracy: 0.9107
<keras.callbacks.callbacks.History at 0x7f75664d90d0>
```

The accuracy on the training set is 99.15% and the accuracy on the test set is 91.07%.

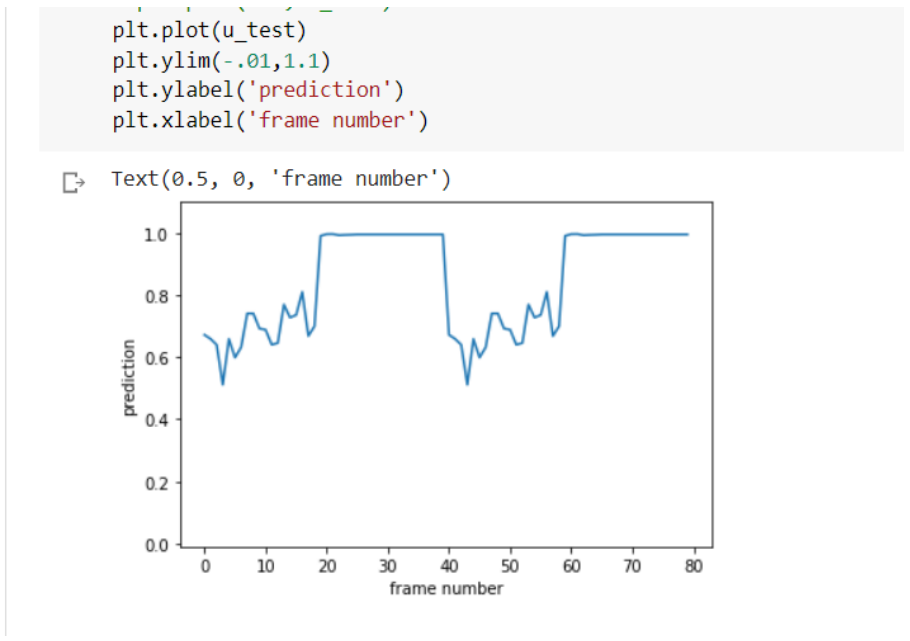
**Testing the model on YouTube videos.**

Again, we convert the video into frames and save them as image files.

Here we call the model to predict the label of frames in the test data:

And then we plot the predicted label for each frame in the video. For this sample video, the shaking-head action is detected which was correct. For each video lips in the test set, we repeat the same process and then we upload true positive samples in the ilab website by a JSON file named CSCE636_baharealik_V3_JSONfile.

An example of a labeled Youtube video by the model is shown below:



```
plt.plot(u_test)
plt.ylim(-.01,1.1)
plt.ylabel('prediction')
plt.xlabel('frame number')
```

```
Text(0.5, 0, 'frame number')
```

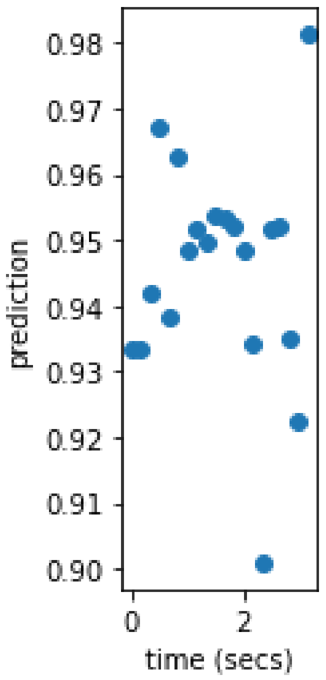
**Discussion:**

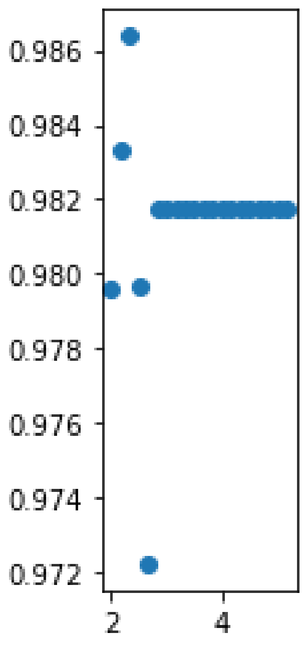
When I was assessing detected positives of test set, I found out that the model cannot detect cutting when the knife is not visible (when the knife is held by a person right in front of the camera. But when the person holds the knife in sideways, the cutting action is recognized. Moreover, when the camera is zoomed on the cutting action, the action is not recognized anymore. In the next submission, I need to collect more data for cutting activity and reduce the false negative of the model.

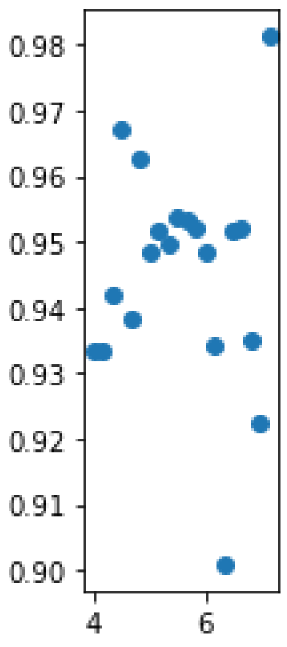
## Project submission 5 (Date: 03/25/2021)

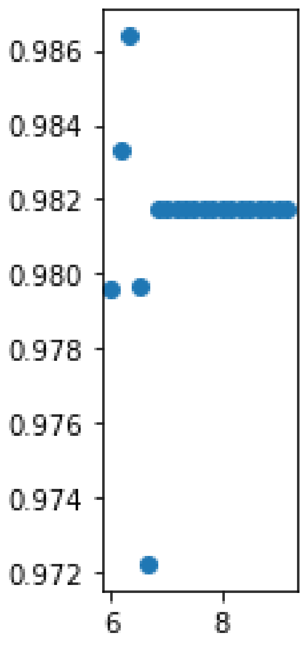
I tried to use different epochs and batch size in this model. I also automated some parts of the model. I also added some new data to my training set. I also added some new data to my test set. I changed how the model was drawing diagrams (previously the diagram was prediction over frame. But now it is prediction over time).

New diagrams are added in this submission:




0.986
0.984
0.982
0.980
0.978
0.976
0.974
0.972
2
4


0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
4
6



Videos are now searched directly from Youtube.

```
import pytube
from pytube import YouTube
import urllib.request
# yt = YouTube("https://www.youtube.com/watch?v=n06H7OcPd-g")
# yt = yt.get('mp4', '720p')
# yt.download('/path/to/download/directory')
searchIDs=['slice fruit','vegetables cutting skill', 'fruit cutting', 'cooking cut']
# os.chdir("/content/drive/My Drive/CSCE636/3_test_Youtube")
downVid=[]
for searchID in searchIDs:
   vidSearch = VideosSearch(searchID, limit=2)
   print(searchID)
   id=vidSearch.result()['result'][0]['id']
   url='https://www.youtube.com/watch?v='+id
   print(url)
   youtube = pytube.YouTube(url)
   video = youtube.streams.first()
   video.download('/content/drive/My Drive/CSCE636/3_test_Youtube')
```

Detected action with an average confidence interval > 95% are reported in the JSON file.

## Project submission 6 (Date: 04/01/2021)

Submission 6

I added several videos to my test dataset, and training dataset. But based on the recent update from the instructor, I removed those that humans are not seen in the videos.

## Project submission 8 (Date: 04/15/2021)

I realized that the format of my JSON file was incorrect. I changed the format of JSON file in my code. The code now generate only integer numbers for time and also each video is 2-seconds long.

I also changed the topic for this submission which is going up and down stairs ladders. For the training set, I asked my family and friends to record some video for me and also I added some video downloaded from the Internet to the training set.

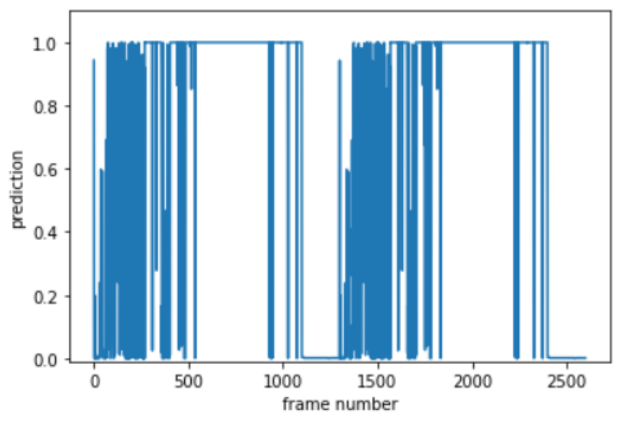
At the beginning the model was sometimes confused by some types of workout (elliptical bikes), and if a human was standing up next to an staircase, and labeled them as 1. For this reason, I added new videos and labeled them as not going up and down stairs and ladders (or label 0). This helped my model to get less confused in the abovementioned situations.

The architecture of the model is the same as before. The accuracy that I got for the validation set was 85.49%:

```
Epoch 1/20
236/236 [==============================] - 50s 58ms/step - loss: 0.3882 - accuracy: 0.7944 -
val_loss: 0.5554 - val_accuracy: 0.9059
Epoch 2/20
236/236 [==============================] - 13s 54ms/step - loss: 0.0287 - accuracy: 0.9871 -
val_loss: 0.6708 - val_accuracy: 0.8627
Epoch 3/20
236/236 [==============================] - 13s 55ms/step - loss: 0.0730 - accuracy: 0.9803 -
val_loss: 0.7077 - val_accuracy: 0.8696
Epoch 4/20
236/236 [==============================] - 13s 56ms/step - loss: 0.0709 - accuracy: 0.9779 -
val_loss: 0.6862 - val_accuracy: 0.8725
Epoch 5/20
236/236 [==============================] - 13s 57ms/step - loss: 0.0104 - accuracy: 0.9982 -
val_loss: 0.8712 - val_accuracy: 0.8451
Epoch 6/20
236/236 [==============================] - 13s 56ms/step - loss: 0.0611 - accuracy: 0.9865 -
val_loss: 0.5867 - val_accuracy: 0.9078
Epoch 7/20
236/236 [==============================] - 13s 55ms/step - loss: 0.0328 - accuracy: 0.9885 -
val_loss: 1.1216 - val_accuracy: 0.8647
Epoch 8/20
236/236 [==============================] - 13s 55ms/step - loss: 0.0153 - accuracy: 0.9968 -
val_loss: 1.0900 - val_accuracy: 0.8676
Epoch 9/20
```

```
236/236 [==============================] - 13s 55ms/step - loss: 0.0174 - accuracy: 0.9950 -
val_loss: 0.6884 - val_accuracy: 0.8814
Epoch 10/20
236/236 [==============================] - 13s 56ms/step - loss: 0.0059 - accuracy: 0.9977 -
val_loss: 1.0716 - val_accuracy: 0.8676
Epoch 11/20
236/236 [==============================] - 13s 56ms/step - loss: 0.0114 - accuracy: 0.9972 -
val_loss: 0.8610 - val_accuracy: 0.8892
Epoch 12/20
236/236 [==============================] - 13s 56ms/step - loss: 9.0443e-04 - accuracy: 1.0000 -
val_loss: 1.1222 - val_accuracy: 0.8814
Epoch 13/20
236/236 [==============================] - 13s 56ms/step - loss: 1.8037e-04 - accuracy: 1.0000 -
val_loss: 1.3167 - val_accuracy: 0.8588
Epoch 14/20
236/236 [==============================] - 13s 56ms/step - loss: 1.1627e-04 - accuracy: 1.0000 -
val_loss: 1.4240 - val_accuracy: 0.8569
Epoch 15/20
236/236 [==============================] - 13s 55ms/step - loss: 6.7435e-05 - accuracy: 1.0000 -
val_loss: 1.4919 - val_accuracy: 0.8539
Epoch 16/20
236/236 [==============================] - 13s 55ms/step - loss: 4.3231e-05 - accuracy: 1.0000 -
val_loss: 1.5429 - val_accuracy: 0.8539
Epoch 17/20
236/236 [==============================] - 13s 56ms/step - loss: 3.8825e-05 - accuracy: 1.0000 -
val_loss: 1.5887 - val_accuracy: 0.8539
Epoch 18/20
236/236 [==============================] - 13s 55ms/step - loss: 2.0261e-05 - accuracy: 1.0000 -
val_loss: 1.6316 - val_accuracy: 0.8549
Epoch 19/20
236/236 [==============================] - 13s 56ms/step - loss: 1.5567e-05 - accuracy: 1.0000 -
val_loss: 1.6764 - val_accuracy: 0.8549
Epoch 20/20
236/236 [==============================] - 13s 56ms/step - loss: 2.1570e-05 - accuracy: 1.0000 -
val_loss: 1.7100 - val_accuracy: 0.8549
```

One sample diagram for one of the videos is:



For the next submission, I will work more on the accuracy and also increase the variations in my dataset.

# Project submission 10

# Training a deep learning model for activity recognition

Bahareh Alizadeh

January 20, 2021

**Abstract**

For this submission, the action selected was "pick up something from floor". In order to improve the accuracy, a few strategies were tested such diversifying the training set, adding similar actions and labeling them as zero, testing different hyperparameters, splitting videos in 4seconds videoclips, and defining better searching keywords. The false positive rate and the true negative rate for the model were reduced significantly (near zero) after applying the identified the strategies.

## Topic:

For the last submission, I chose the topic "Pick up something from floor". There were zero pre-existing video clips for this topic in the iLab website. The possible challenge for this topic is that the model may detect similar actions as false positives.

## Motivation:

To prevent the model from making this type of error, we have to add these actions to the negative training set (label 0). As seen in the code below, I defined the word "pick up" and also "lift". For negative actions, I defined "drop", "throw" and "workout". As an example, a person might lift a dumbbell and the model assumes that the person is picking up something from flood. Thus, by identifying similar actions, we increase the accuracy of the model by reducing the number of false positives.

## Related work:

In a research by Xia, et al. (2020), combined a deep neural with LSTM models for detecting pattern recognition. In another research by Guan and Plotz (2017), a deep LSTM model was used for activity recognition in wearable sensors. Ibrahim et al. (2016), also focused on the temporal features of video and designed an LSTM based model to capture that. In a research by Hammann et al. (2020) used a model combining LSTM and ResNet for Human Identity Recognition in a car cabin with the accuracy of 79.49 % on a 10-drivers subset of NUDrive and 96.90 % on a 5-drivers subset of UTDrive. Figure 1 and 2 show the architecture of the model used by this paper.

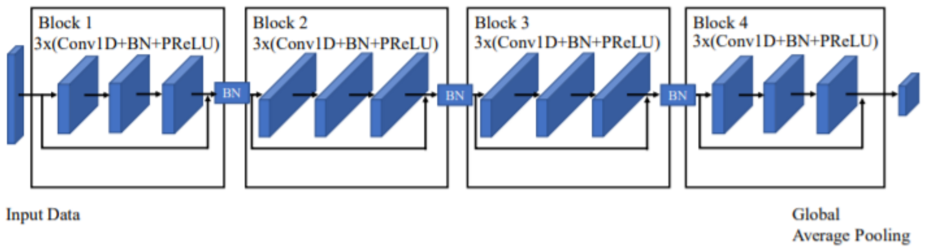


Fig. 1: ResNet Architecture

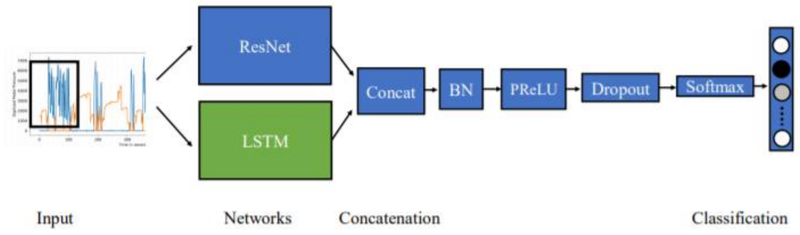


Fig. 2: LSTM-ResNet Architecture

For this project, I chose the combination of ResNet and LSTM to detect human activity.

To measure the change in the accuracy, I examined the model on 5 videoclips from YouTube using the search keywords. After defining negative actions, the number of false positive with regard to the three negative actions (“drop”, “throw”, “workout”) was zero. Thus the model’s performance in detecting false positives was improved 31% (Eq. 1). There are many factors that may affect this number, such as the number of videoclips with negative actions, total number of videoclips, and the model’s threshold for it’s confidence level. However, this number gives us a good estimation on how the model’s performance improves with training on a diversified dataset.

$$\frac{FP}{N} = \frac{FP}{FP + TN} \qquad \text{Eq.1}$$

Here the model load data from the specified folders.

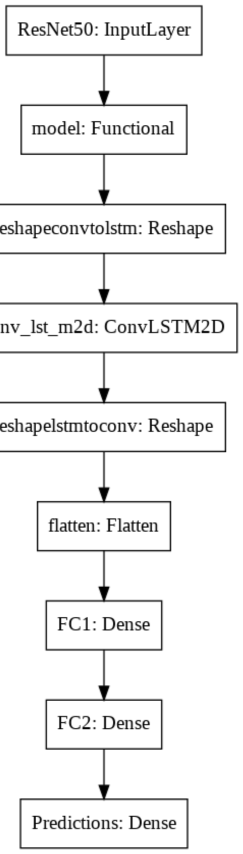
**Video processing:**

Training videos are formatted, and their information are placed in a long list. Features of each video will be extracted. Features of each video will be saved as a numpy file. The reason for saving the features as numpy file is that the size of if is much smaller and can be easily loaded later whenever needed.

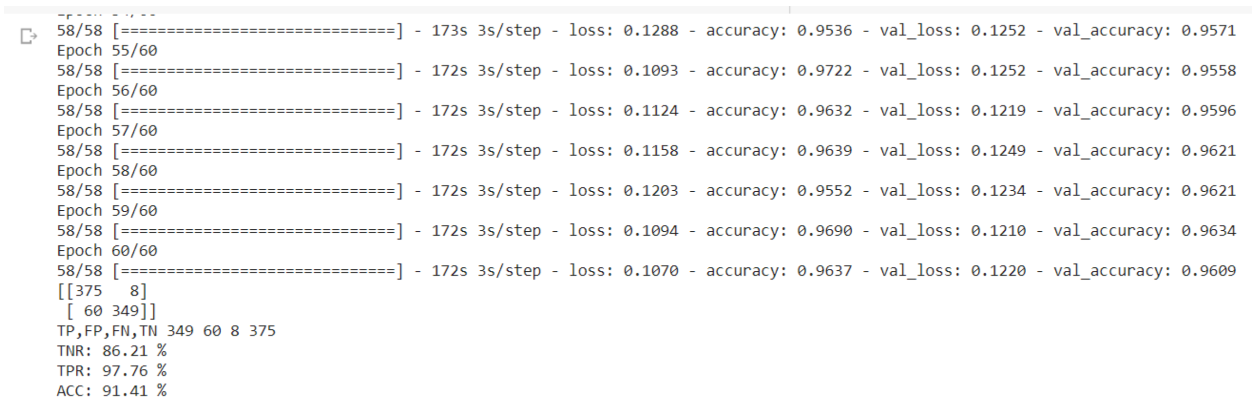
After extracting the features, the dataset will be splitted to training and validation sets. 30% of the training will be selected as validation set.

## Proposed Model

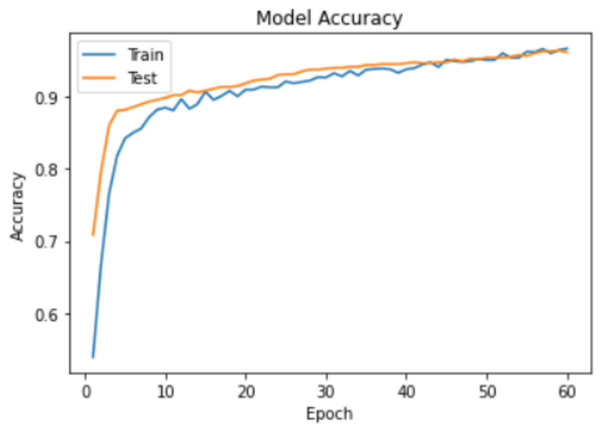
I designed the model based on ResNet and LSTM and I got this idea from the paper by Hammann et al. (2020). The architecture of this model is showen below. For more information on the details of the model please refer to the submission 1.

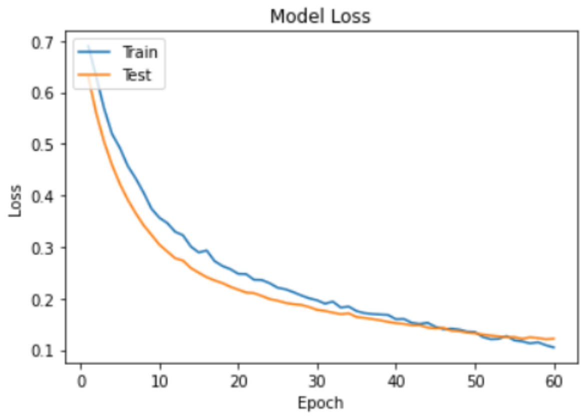


## Model Training and Performance



```
58/58 [==============================] - 173s 3s/step - loss: 0.1288 - accuracy: 0.9536 - val_loss: 0.1252 - val_accuracy: 0.9571
Epoch 55/60
58/58 [==============================] - 172s 3s/step - loss: 0.1093 - accuracy: 0.9722 - val_loss: 0.1252 - val_accuracy: 0.9558
Epoch 56/60
58/58 [==============================] - 172s 3s/step - loss: 0.1124 - accuracy: 0.9632 - val_loss: 0.1219 - val_accuracy: 0.9596
Epoch 57/60
58/58 [==============================] - 172s 3s/step - loss: 0.1158 - accuracy: 0.9639 - val_loss: 0.1249 - val_accuracy: 0.9621
Epoch 58/60
58/58 [==============================] - 172s 3s/step - loss: 0.1203 - accuracy: 0.9552 - val_loss: 0.1234 - val_accuracy: 0.9621
Epoch 59/60
58/58 [==============================] - 172s 3s/step - loss: 0.1094 - accuracy: 0.9690 - val_loss: 0.1210 - val_accuracy: 0.9634
Epoch 60/60
58/58 [==============================] - 172s 3s/step - loss: 0.1070 - accuracy: 0.9637 - val_loss: 0.1220 - val_accuracy: 0.9609
[[375   8]
 [ 60 349]]
TP,FP,FN,TN 349 60 8 375
TNR: 86.21 %
TPR: 97.76 %
ACC: 91.41 %
```





The model's accuracy will be increased on train set and test set to 96.37% and 96.09%, respectively. The accuracy of the model is 91.41%. The true positive rate is 97.76%. The true negative rate 86.21%.

**Dataset**

A pre-existing dataset was downloaded from the Internet which contained a lot of action including pick up. Since the videoclips in the dataset was old, I also recorded some videoclips from myself and friends and family members and added them to the training dataset. In these videos we tried to do actions similar to pick up as well and then labeled them as zero.

**Performance on YouTube Videos**

To prepare a test dataset, videoclips will be downloaded from YouTube using related keywords. The searching keywords were 'pick up something', 'lift something floor', 'messy room', 'house cleaning'. To reason for selecting messy room and cleaning house is that in these people will pick up things from floor while cleaning.

## JSON file

Then the information of video clips (the start time of the detected action in the video clip, the end time of the detected action in the video clip (2 seconds after the start time), the confidence level, the label, and the YouTube video ID was saved in the JSON file. Then the JSON file will be submitted in the specified folder. Then, to check the videos manually, videos were downloaded separately in a folder and the user evaluated if videos uploaded contain the target action.

## Improve Accuracy and Efficiency

As stated in the report in different sections, different strategies were used for improving accuracy. These strategies are listed as below:

- Adding similar actions to the dataset and label them as zero
- Diversifying the videos in the training set
- Because 2seconds was not good enough time to detect an action, I changed the time span to 4 seconds and tried to detect the activity on the video and then trim the 1-3 seconds of the video.
- Processing all frames in videos are not an efficient way. Thus, in each 4seconds of videos, 12 frames were extracted.
- All frames that are extracted from the videoclips are saved in numpy format. This helped the model to save a lot of time and memory in processing them later with the model.

## Reference

Xia, K., Huang, J., & Wang, H. (2020). LSTM-CNN architecture for human activity recognition. *IEEE Access*, *8*, 56855-56866.

Guan, Y., & Plötz, T. (2017). Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1-28.

Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1971-1980).

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access*, *6*, 1155-1166.

Hammann, M., Kraus, M., Shafaei, S., & Knoll, A. (2020). Identity Recognition in Intelligent Cars with Behavioral Data and LSTM-ResNet Classifier. *arXiv preprint arXiv:2003.00770*.