Recognize Flu-like Symptoms with Deep Learning

Alan Lou
Stanford University
alanlou@stanford.edu

Yechao Zhang
Stanford University
yechaoz@stanford.edu

Abstract

Human action recognition has always been one of the most active research fields, especially with the recent developments of Deep Learning. In this paper, we present a variety of models to detect coughing and sneezing motions from video data. Previous work [3] exists in this subject area, but is limited and not based on Deep Learning approaches. Here we first explored two baseline models: CNN + LSTM model and 3D-Conv model. These two baseline models perform worse than the previous work, which attained a test accuracy of 44.4%. Then we improved the Baseline Model 1 to a VGG-16 Features + LSTM model and achieved a test accuracy of 40.3%. Finally, we proposed an HRNet Features + LSTM network model. Using pre-trained HRNet to extract human skeleton features from video frames, our proposed model outperforms the previous work significantly, achieving a test accuracy of 79.4%.

1. Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic has reached almost every country in the world, infecting millions of people and plunging the global economy into recession as governments imposed tight restrictions to tackle the spread of the virus. Vision-based surveillance such as monitoring and analyzing video footage from densely populated areas to predict flu-like symptoms can be utilized to detect respiratory viral infections such as COVID-19 early.

The previous work [3] predicts flu-like symptoms from videos with traditional feature extraction methods such as histogram of oriented gradient (HOG) and histogram of optical flow (HOF). For this project, we are going to improve the classification with adoption of the recent deep learning models leveraged for computer vision.

2. Problem statement

Recognizing flu-like symptoms from videos is an action recognition problem. The inputs of our model are video frames with human performing different actions and the outputs are action labels corresponding to the videos.

The goal of our model is to recognize flu-like symptoms (sneeze and cough) among a variety of actions, which makes it a binary classification task. We will convert the predicted actions into binary label of coughing and sneezing actions. Since the class distributions are unbalanced, we are going to use precision and recall to evaluate the model performance quantitatively. The previous work [3] additionally proposed an accuracy measure of $TP/(TP+FP+FN)$ for the binary classification, which can be regarded as a lower-bounding summary of the (precision, recall) pair. We will calculate this measure for our models as well and use it to compare against the previous work.

3. Dataset

We use the BII Sneeze-Cough Human Action Video Dataset (BIISC) [4] created by the authors of the paper mentioned above. The dataset consists of 20 subjects. Each subject performs 8 different actions such as coughing and stretching arms. Within each action category, the subject performs the action from 3 different angles (face to the camera, left or right) in 2 poses (stand or walk). For each of the videos, a horizontally flipped version is also created. This in total provides 1920 videos, each of which has a frame rate of 10 fps and a resolution of 480x290 pixels. The video ranges from 3 second to 10 second. We clipped videos to the first 6 second if the video length is longer.

For CNN-based models, we extract 2 frames from each second of video, so each video converts to 12 or fewer images. Then, we crop out the subject from the frame pictures to reduce background noise. To do so, we use Mask R-CNN [1] with Detectron2 and pre-trained MS COCO model. We also resize the images to (224, 224) for faster computation and easier integration with VGG16 feature extract later. For HRNet-based models, we extract 10 frames from each second of video, so each video converts to 60 or fewer images.

To evaluate model performance on the same basis, we split the data following the previous work, where the videos from subjects S002-S006 are used for testing and the remaining subjects are used for training.

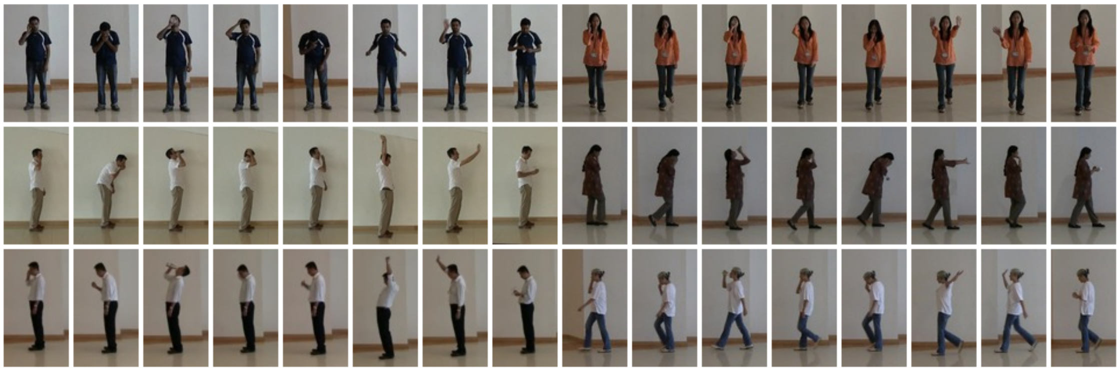

Figure 1. From top to down shows eight actions: answer phone call, cough, drink, scratch head, sneeze, stretch arm, wave hand and wipe glasses. From left to right shows six pose-and-view variations: stand-front, stand-left, stand-right, walk-front, walk-left, and walk-right.

4. Technical Approach

4.1. CNN + LSTM (Baseline 1)

The first baseline model runs every image of the sequenced video frames through a Convolutional Neural Network (Feature Extractor). Next, we feed the flattened outputs as a sequence into a Long Short Term Memory Recurrent Neural Network (LSTM). We then connect the LSTM to Dense layers with Relu activation, a dropout layer, and finally a Dense layer with Softmax activation function to produce the probabilities of every class. The class with maximum probability is the predicted action. The whole network structure is described in Figure 3.

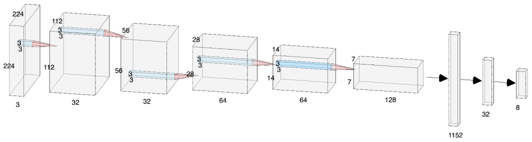

Figure 2. CNN Feature Extractor for CNN + LSTM model

The Convolutional Neural Network feature extractor (Figure 2) is composed of a series of Conv 3 and Max-Pooling 2 blocks.

4.2. Conv-3D (Baseline 2)

The second baseline model uses a 3D convolution layer with kernel size 3 connected with a 3D max-pooling layer as the feature extractor. Then the flattened vector is fed into the LSTM + Dense layers described in Figure 3.

4.3. VGG-16 Features + LSTM

We modified the CNN + LSTM (Baseline 1) model by utilizing transfer learning from a pre-trained VGG-16 [2] network. VGG is a classical Convolutional Neural Network architecture that is characterized by its simplicity and efficiency. The only other components of VGG network are pooling layers and dense layers.

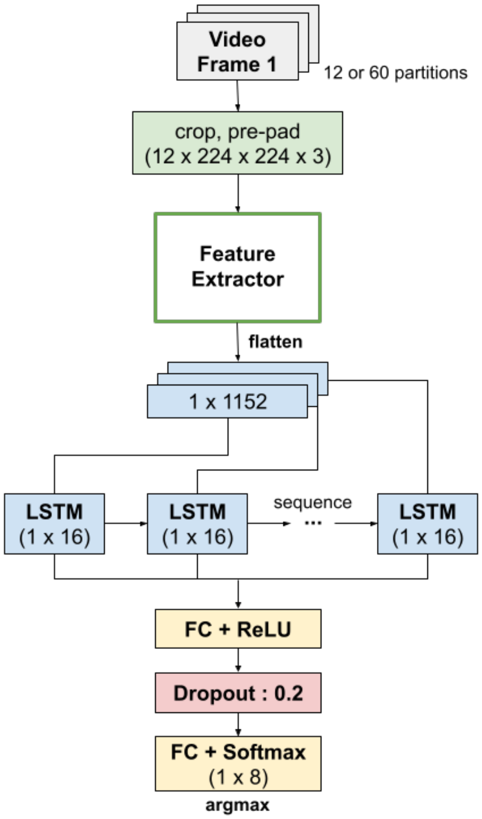

Figure 3. Network architecture

The VGG-16 feature extractor helps the network to learn more intricate features of the input video frames. The last Dense layer of the VGG-16 model is fed into the LSTM + Dense layers described in Figure 3.

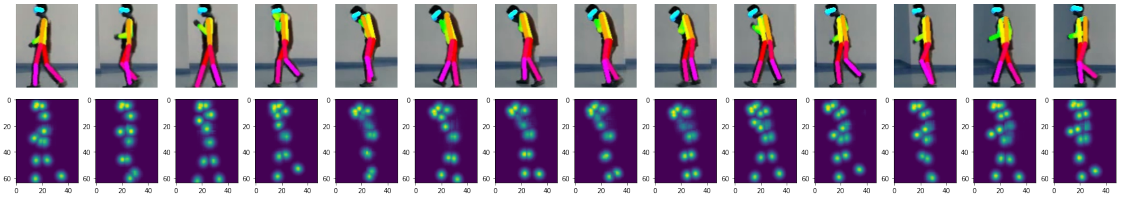

Figure 4. Top row is extracted video frames with HRNet visualization. Bottom row is the overlap of heatmaps from 17 channels.

4.4. HRNet Features + LSTM

We find that human poses are essential to the classification of human actions. High Resolution Net (HRNet) [5] is a state of the art neural network for human pose estimation, which finds the configuration of subjects' joints and body parts in images.

This proposed model runs every frame of the input video through a pre-trained HRNet and outputs a heatmap of size 64 x 48 and 17 channels for each image. The 17 channels correspond to the 17 key-points of human body. Sample extracted HRNet features are presented in Figure 4.

Using pre-trained HRNet as the new feature extractor, we pass the flattened heatmap to the LSTM + Dense layers to obtain the action with maximum probability.

5. Intermediate/Preliminary Results

Below we present the validation and test accuracy of multi-class action classification over training epochs. In the test dataset, the number of samples from each class is the same, so accuracy is a balanced performance measure. Comparing different methods, the HRNet + LSTM model achieved best training and validation accuracy.

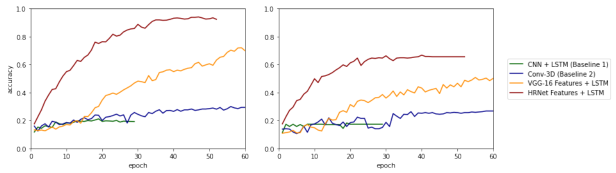

Figure 5. Training and Validation Accuracy Plots for different models

Figure 6 shows the multi-class action recognition confusion matrix based on test dataset. Baseline Model 1 and Baseline Model 2 tend to predict certain classes over the other ones, which resulted in lower accuracy. The VGG-16 Features + LSTM model and HRNet Features + LSTM have more balanced predictions.

We convert the predicted action label into binary label of coughing and sneezing actions. Since the binary label is biased, we evaluate the performance of our models using Precision (Prec.) and Recall (Rec.), which are calculated as $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, respectively. To compare with the previous work, we also adopt a different accuracy measure of $\frac{TP}{TP+FP+FN}$, which can be regarded as a lower-bounding summary of the (precision, recall) pair.

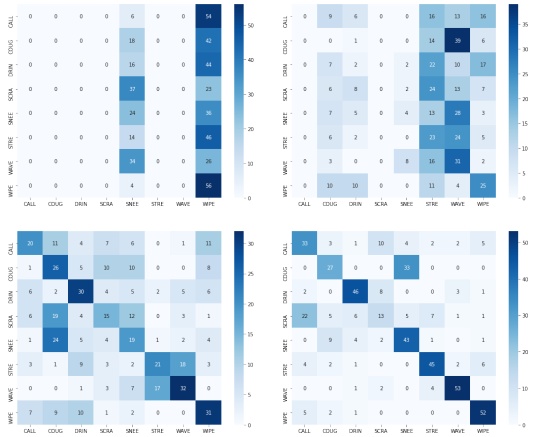

Figure 6. From top-left to bottom-right are confusion matrices of CNN + LSTM (Baseline 1), Conv-3D (Baseline 2), VGG-16 Features + LSTM, and HRNet Features + LSTM, respectively.

The result of binary classification on test dataset is reported in Table 1. The two baseline models perform worse than the handcrafted traditional feature extraction methods. The VGG-16 Features + LSTM has a slightly lower accuracy than the previous work, while the HRNet Features + LSTM model performs significantly better than all other models.

Model	Prec.%	Rec.%	Acc.%
cuboid + AMK II	55.3	62.1	41.3
HOGHOF + AMK II	58.9	64.4	44.4
CNN + LSTM	27.5	35.0	18.2
Conv-3D	17.2	9.2	6.4
VGG-16 Features + LSTM	51.0	65.8	40.3
HRNet Features + LSTM	**84.2**	**93.3**	**79.4**

Table 1. Comparisons of recognition accuracies of Sneeze-Cough actions

Given more time and resources, we would continue to tune the architecture of the proposed models. We believe that the models outlined in this paper could still be improved, especially the baseline models.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 1

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2

[3] Tuan Hue Thi, Li Wang, Ning Ye, Jian Zhang, Sebastian Maurer-Stroh, and Li Cheng. Recognizing flu-like symptoms from videos. *BMC Bioinformatics*, 15(1):300, 2014. 1

[4] Tuan Hue Thi, Li Wang, Ning Ye, Jian Zhang, Sebastian Maurer-Stroh, and Li Cheng. Recognizing flu-like symptoms from videos, 2014. 1

[5] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *CoRR*, abs/1908.07919, 2019. 3