Project Report

1.Target Action: Clapping hands

2.Dataset for training:

I assemble two datasets from existing datasets. I collected the videos of clapping hands from HMDB dataset, STAIR action dataset and Action Database. The first dataset I assembled contain 18 actions and 3,032 labeled video files in total. The second dataset I assembled contain 2 actions (clapping hands and other actions). The second dataset has 2,473 video files with 1,163 video files as clapping hands.

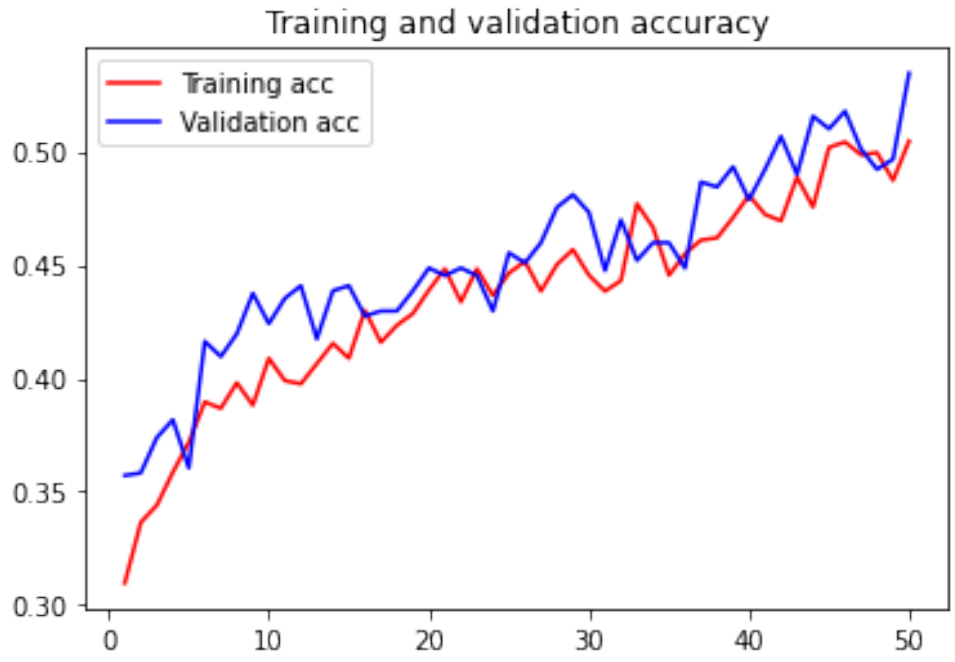
3. Model

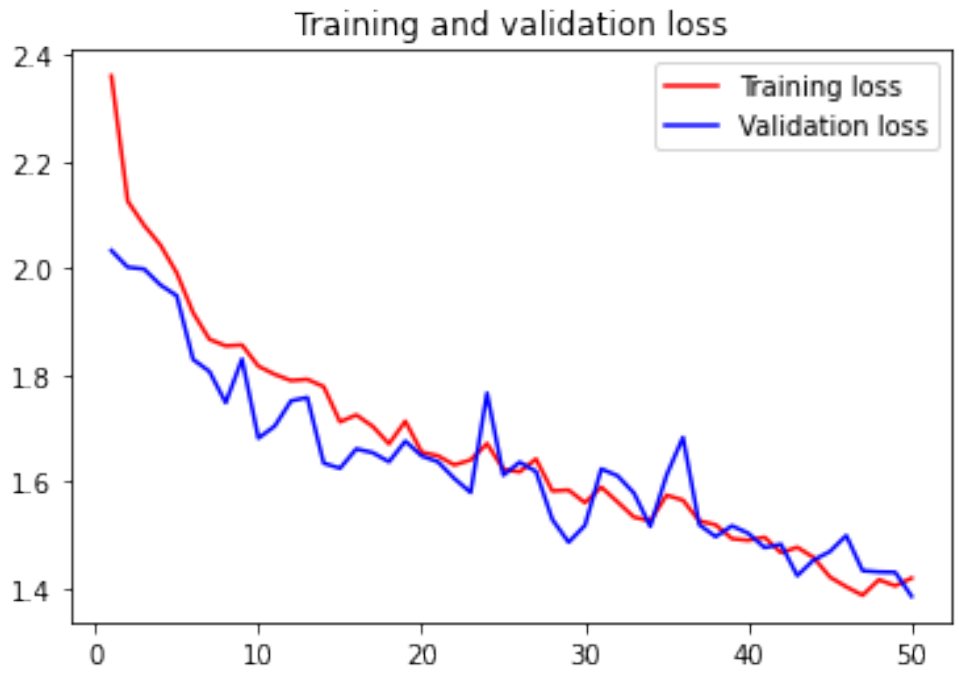
I tried many different settings of the model and decided to use a CNN+RNN+Classifier type model. I used CNN to extract the features from video frames and used RNN to extract the sequence information from videos. The CNN model is trained by a transfer learning approach. I used pre-trained Neural Network, MobileNetV2 and fine-tuned its last 12 layers. The RNN model used GRU as the units.

I trained a multi-label classification model on the first dataset and trained a binary-label classification model on the second data set.

4. Performance

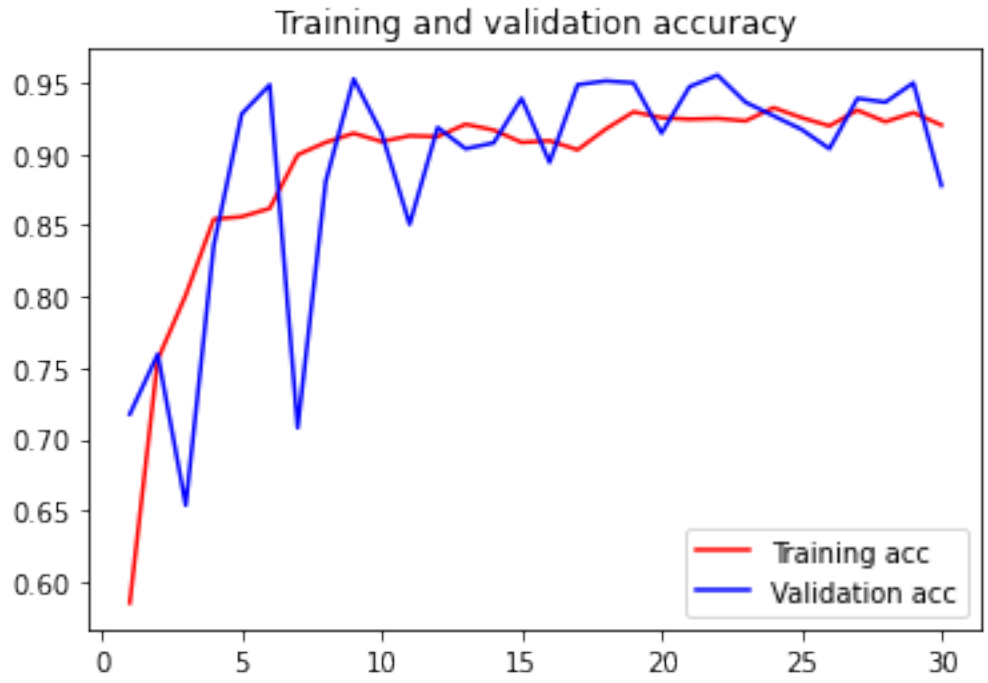
Multi-label classification model:

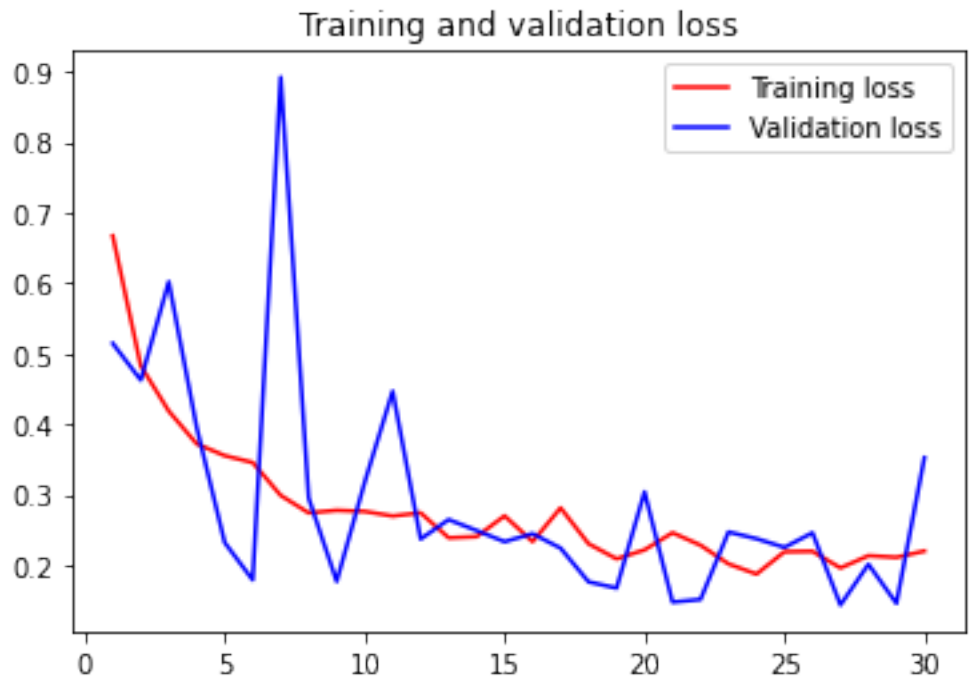




The performance is quite poor. The training and validation accuracy is around 50%.

Binary-label classification model:





The binary-classification model works quite well. The training and validation accuracy is around 90% after 15 epochs of training. The false positives are around 70. So I chose this model to do the action detection.