Please use this report template, and upload it in the PDF format. Reports in other format will result in ZERO point. Reports written in either Chinese or English is acceptable. The length of your report should NOT exceed 8 pages.

Name:曾柏偉　Dep.:電信碩一　　Student ID:R06942098

[Problem1]

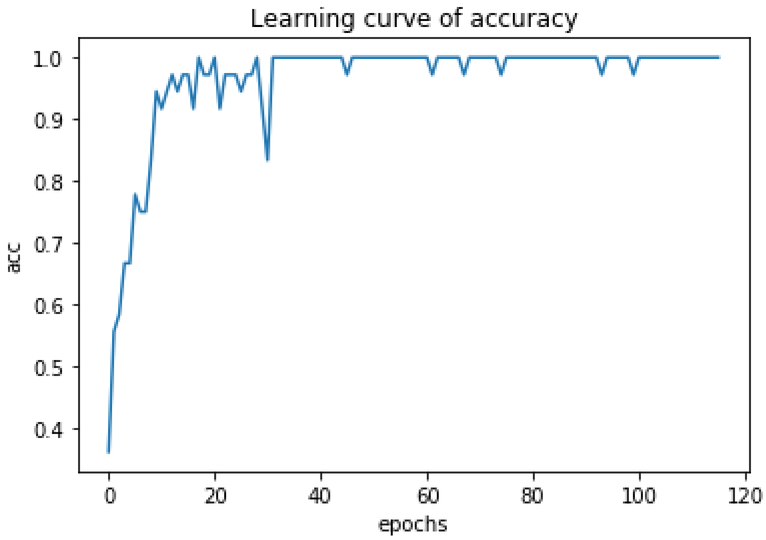
1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.

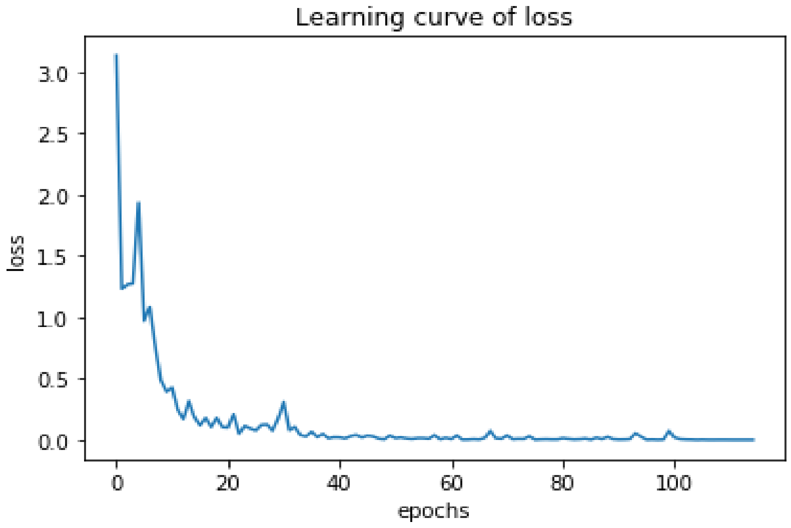
   The pre-trained model I choose is VGG-16 , and I get the CNN features from fully connected layer named “fc6” . Finally , I take the first , middle and the last frame of each video. If the length of video is below three , I will repeat the last time stamp CNN feature and concatenate to the CNN-based features matrix that length is less 3.

2. (15%) Report your video recognition performance using CNN-based video features　and plot the learning curve of your model.

   When I use the CNN-based video features to train a classification model that only include two fully connected layers , it can get the training accuracy : 1.0 , but the validation accuracy is about 0.46808510638297873.

   Two figures are the training learning curve :



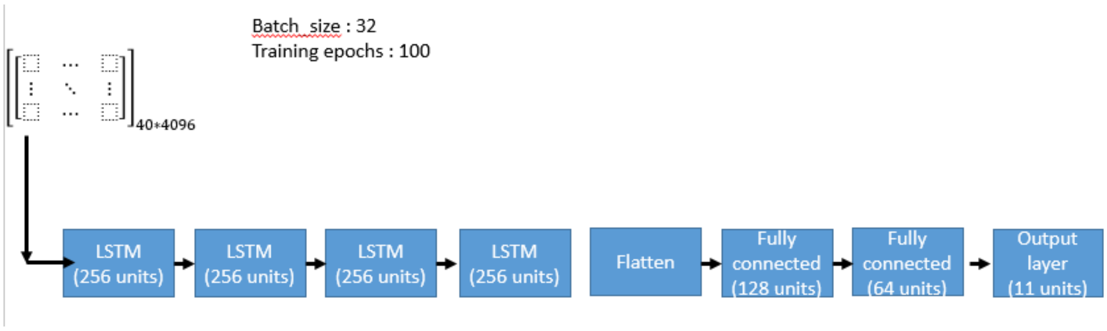


[Problem2]

1. (5%) Describe your RNN models and implementation details for action recognition.

   Because I use the reader.py to load video feature , so the maximum frame of all the video is 261 . Zero-padding is not a good way to fulfill this task. Then I randomly choose continuous 40 frames of each video as training data. If the frame of video is below 40, I will pad zeros to frames of 40. And the CNN I use is VGG-16.

   Model structure :

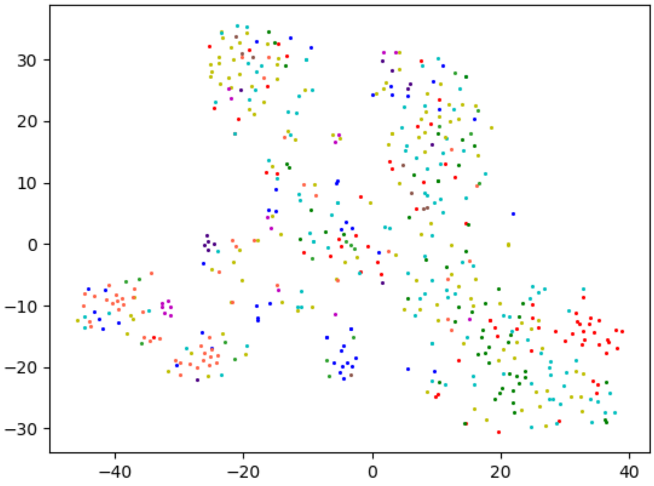


Batch_size : 32
Training epochs : 100

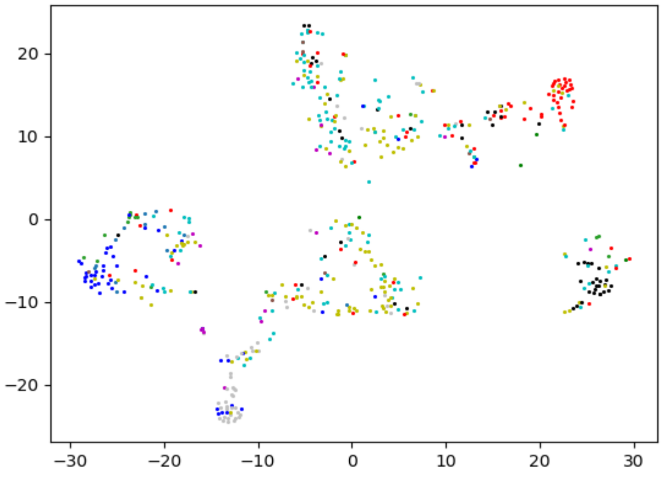
2. (15%) Visualize CNN-based video features and RNN-based video features to 2D space (with tSNE). You need to generate two separate graphs and color them with respect to different action labels. Do you see any improvement for action recognition? Please explain your observation.

   It’s obvious that the RNN-based video feature is better than CNN-based. The main reason may be that I have done the end-to-end training on the whole model, so this model learn the feature of each label . The other may be that I have not fine-tune the VGG-16 model..

CNN based features:



RNN based features :



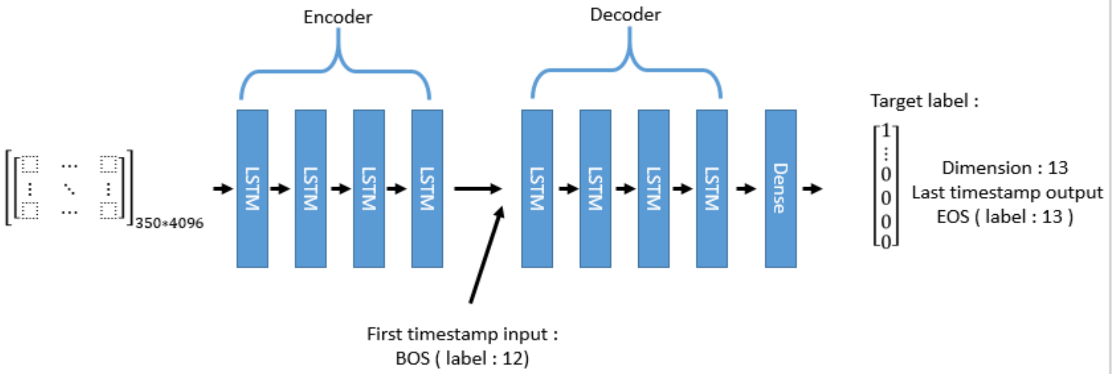
[Problem3]

1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.

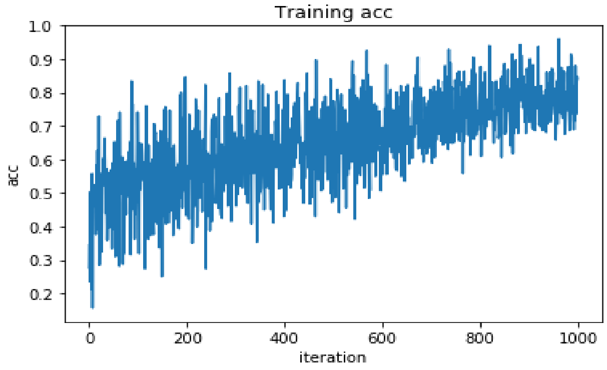
Ans:

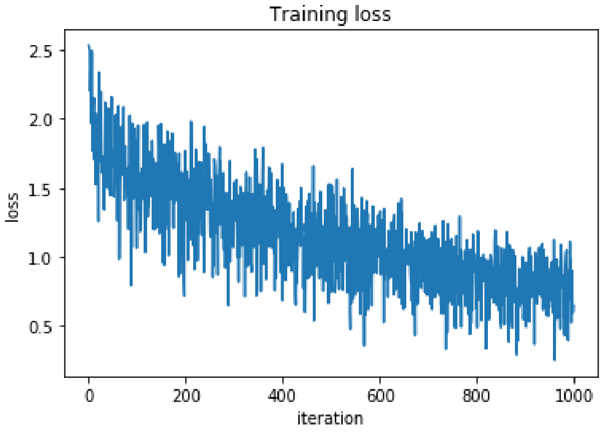
Use VGG-16 to get the CNN features of each images. I try to use random sample to deal with the very long sequences. So each updates I sample 350 continuous frames of each video. Then pass it to seq2seq attention-based model (AttentionWrapper).

Model structure :

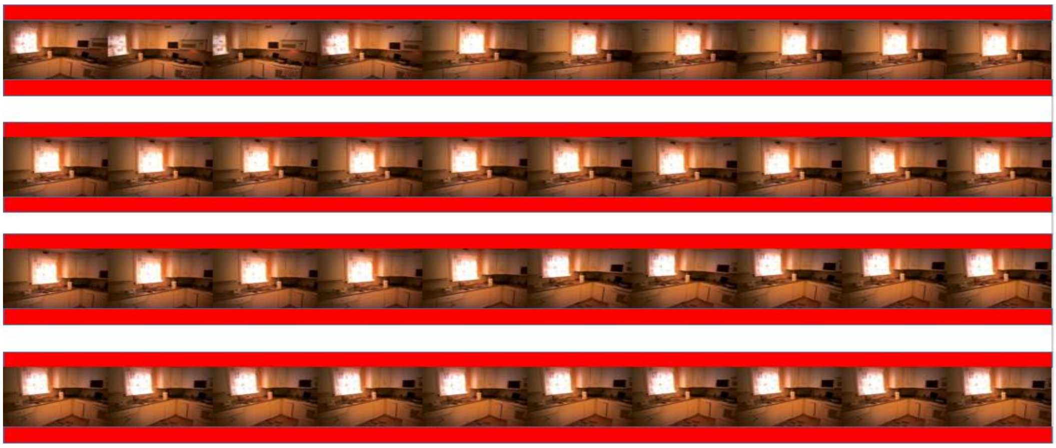


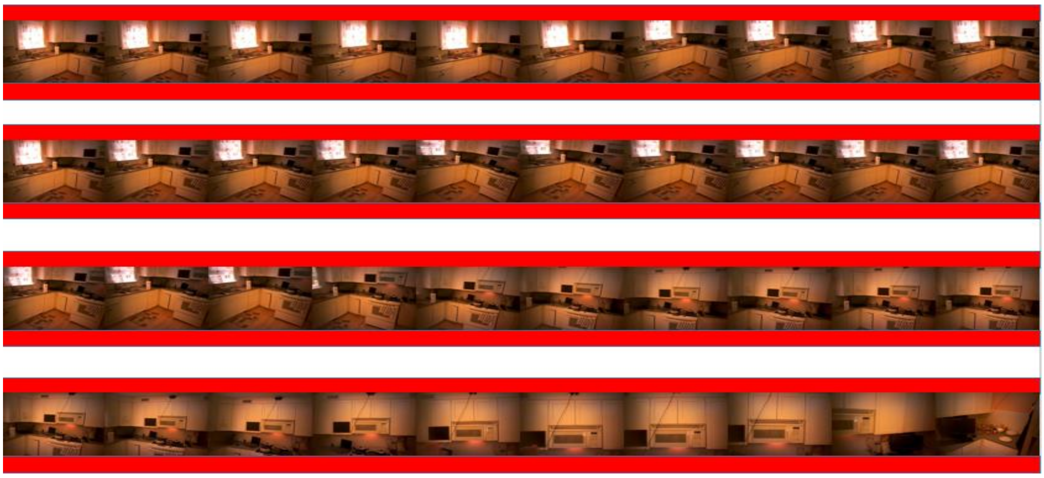
2. (10%) Report validation accuracy and plot the learning curve.
(model only fit training data…)
Validation accuracy : only 0.3627129750982962

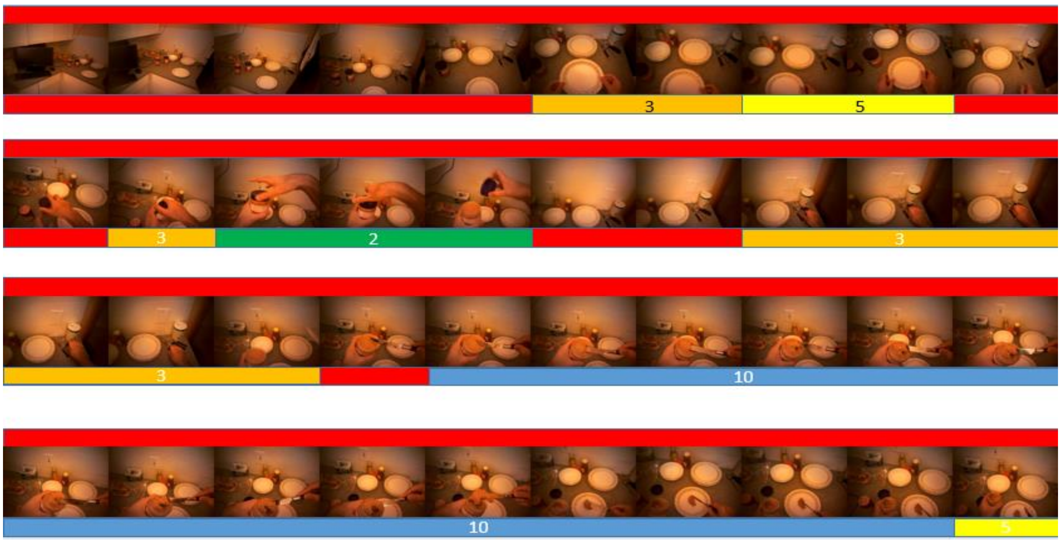




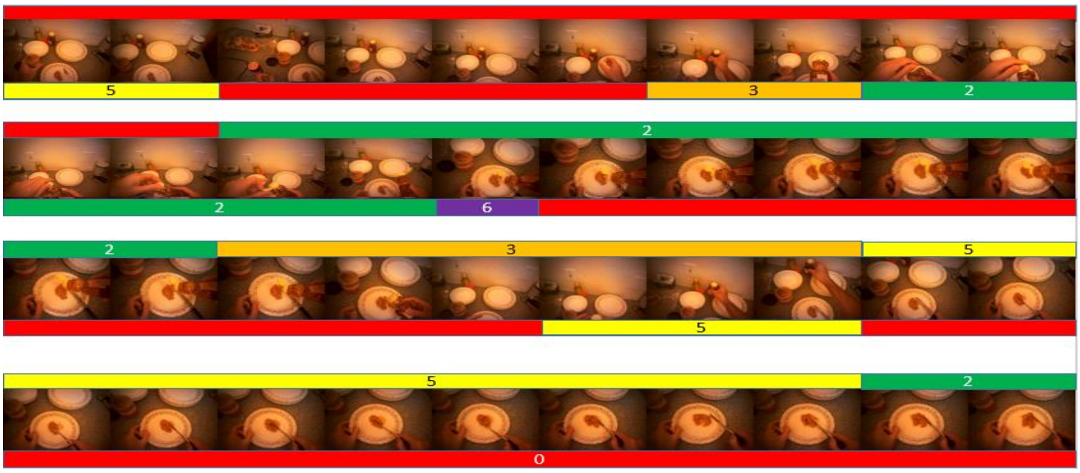
3. (10%) Choose one video from the 5 validation videos to visualize the best prediction result in comparison with the ground-truth scores in your report. Please make your figure clear and explain your visualization results. You need to plot at least 300 continuous frames (2.5 mins).

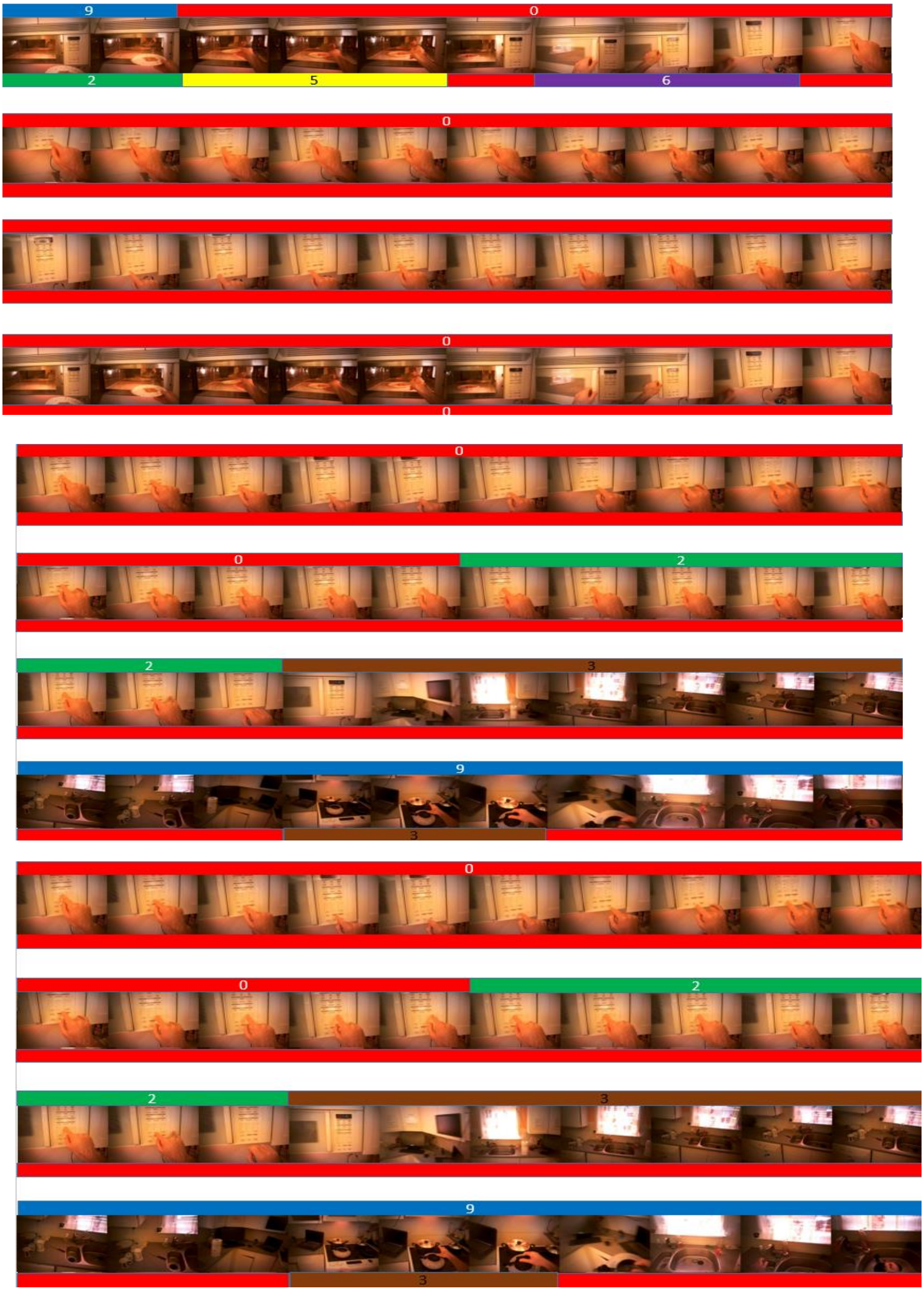


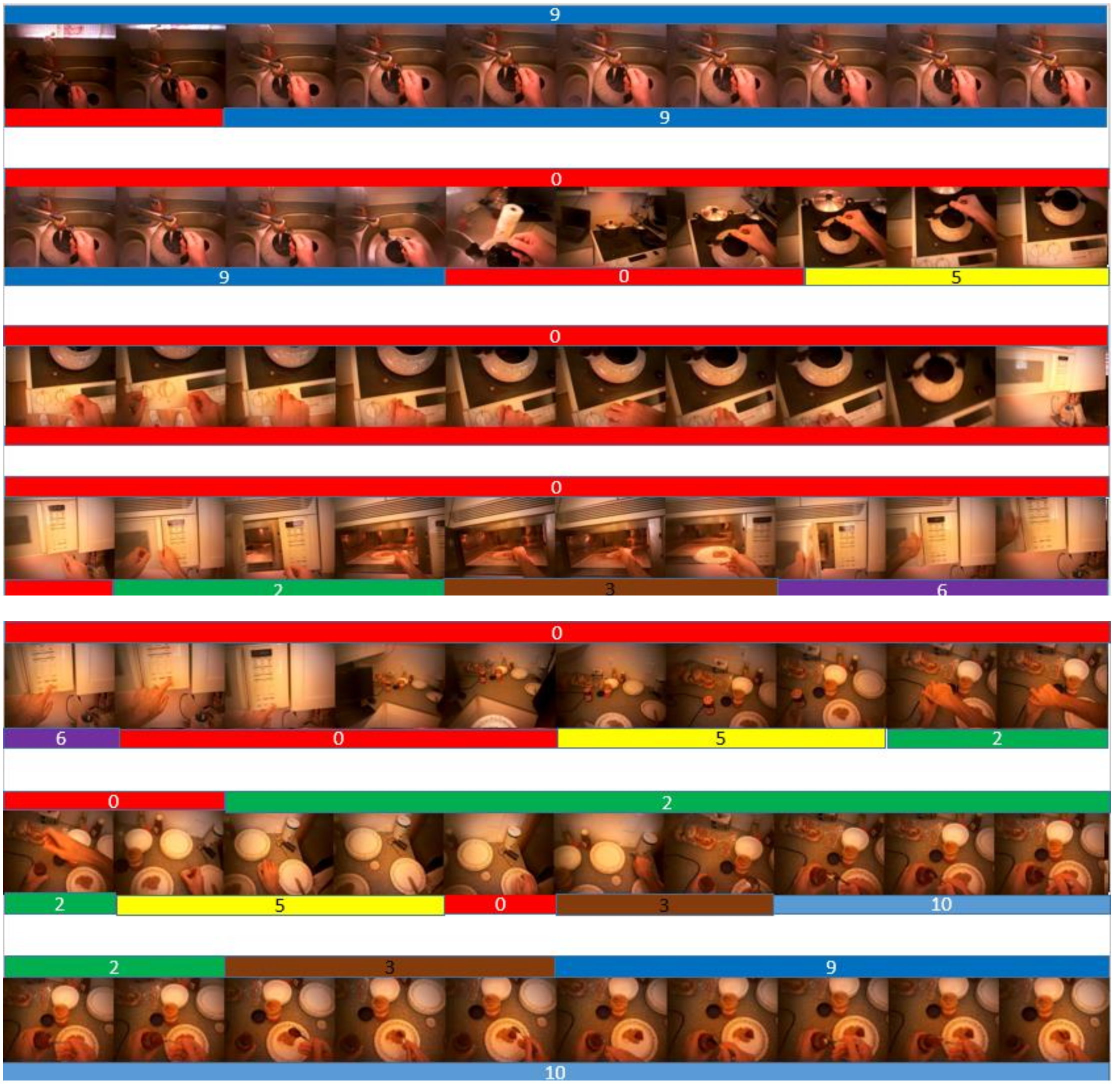



3
5
3
2
3
3
10
10
5

5
3
2
2
2
6
2
3
5
5
5
2
0


5
3
2
2
2
6
2
3
5
5
5
2
0


9
0
2
5
6
0
0
0
0
0
2
2
3
9
3
0
0
2
2
3
9
3



Among all videos, this one can achieve accuracy: 0.43(best) .After plotting every ten frames of this videos about it prediction(above color bar) and ground truth (below color bar) , I observe there are sample problems of machine translation … . If there is any image that model can’t distinguish which class it is, model will always predict label 0. It’ same as machine translation , model only say “ I don’t know “ . Another problem is , if the previous timestamp are wrong , the following timestamp will be wrong too…

[BONUS]