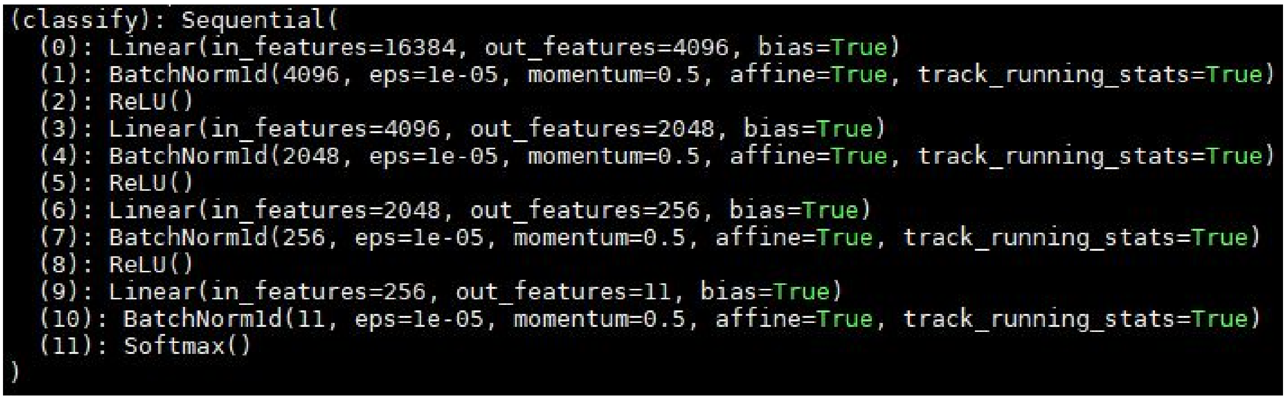
Name: 鄭閎 Dep.:電機三 Student ID:B04901155

[Problem1]

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

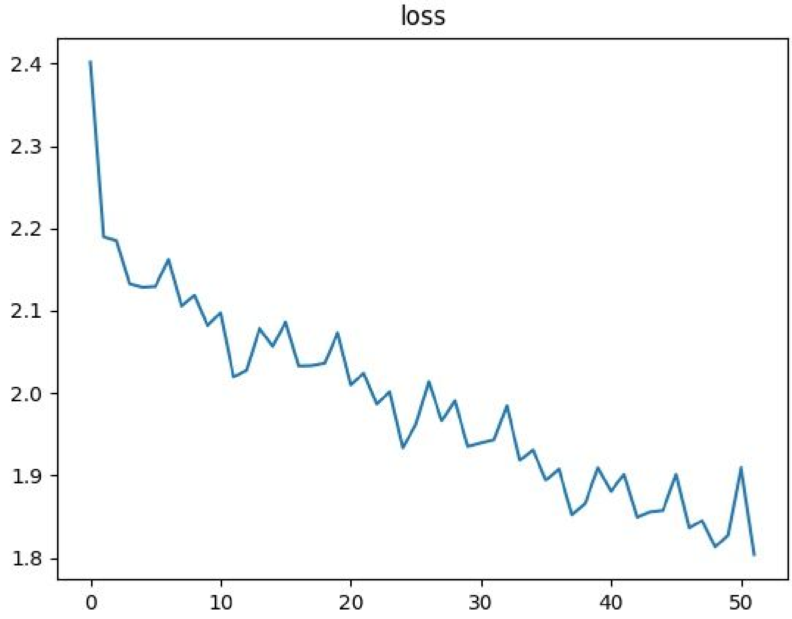
I use ResNet50 to extract the CNN-based video features. Remove the last classification layer to get the feature. Then using the reader function provided by TA, I sample the frame from videos and average the features in same video. Feed the averaged feature vectors to fully connected layers to classify.



```
(classify): Sequential(
  (0): Linear(in_features=16384, out_features=4096, bias=True)
  (1): BatchNorm1d(4096, eps=1e-05, momentum=0.5, affine=True, track_running_stats=True)
  (2): ReLU()
  (3): Linear(in_features=4096, out_features=2048, bias=True)
  (4): BatchNorm1d(2048, eps=1e-05, momentum=0.5, affine=True, track_running_stats=True)
  (5): ReLU()
  (6): Linear(in_features=2048, out_features=256, bias=True)
  (7): BatchNorm1d(256, eps=1e-05, momentum=0.5, affine=True, track_running_stats=True)
  (8): ReLU()
  (9): Linear(in_features=256, out_features=11, bias=True)
  (10): BatchNorm1d(11, eps=1e-05, momentum=0.5, affine=True, track_running_stats=True)
  (11): Softmax()
)
```

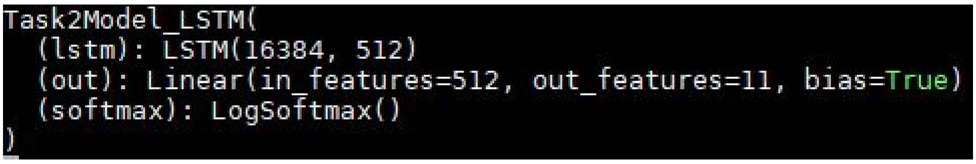
2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.

accuracy on training set: 0.9978
accuracy on validation set: 0.4545



[Problem2]

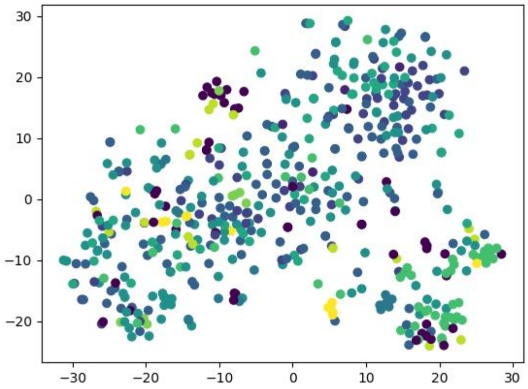
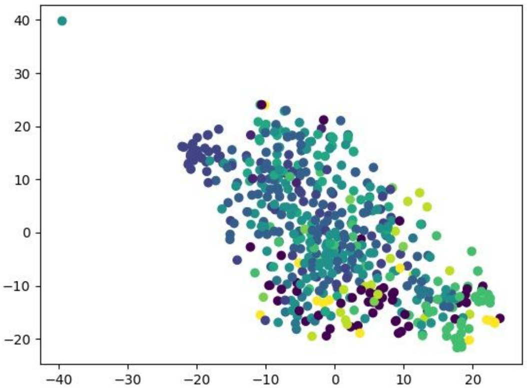
1. (5%) Describe your RNN models and implementation details for action recognition.



```
Task2Model_LSTM(
  (lstm): LSTM(16384, 512)
  (out): Linear(in_features=512, out_features=11, bias=True)
  (softmax): LogSoftmax()
)
```

Use 1 layer of LSTM with hidden state 512, output dimension 11. Use the features get from task 1, instead of averaging the features, I feed these features to RNN one by one and get the output of last timestep as prediction.

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.

	
CNN	RNN

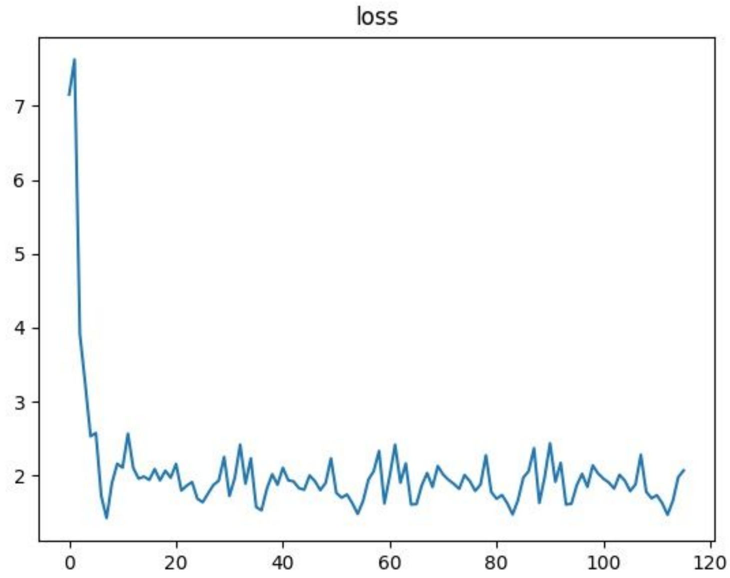
[Problem3]

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

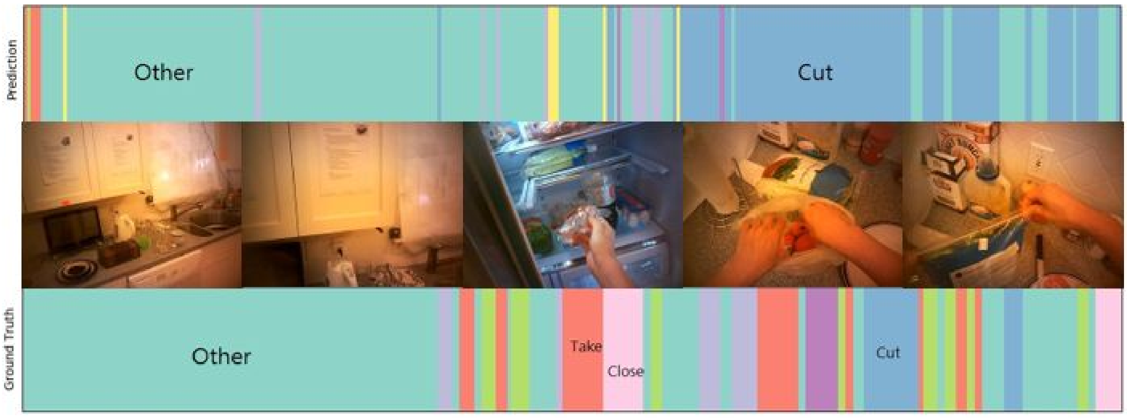
Use the same model as task 2, however this time get the output for every time frame rather than only the final one. I set batch size as 1 to avoid the padding issue. The loss is accumulated for all frames in one video and the gradients are updated after one video is finished.

2. (10%) Report validation accuracy and plot the learning curve.

   validation accuracy: 0.461



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).



For actions that last longer like “Other” and “Cut” in my example, the model has higher chance to predict correctly. However, the model cannot perform well at rapid change of actions.

[BONUS]