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:黃宇平 Dep.:電信碩一 Student ID:R06942065

[Problem1]

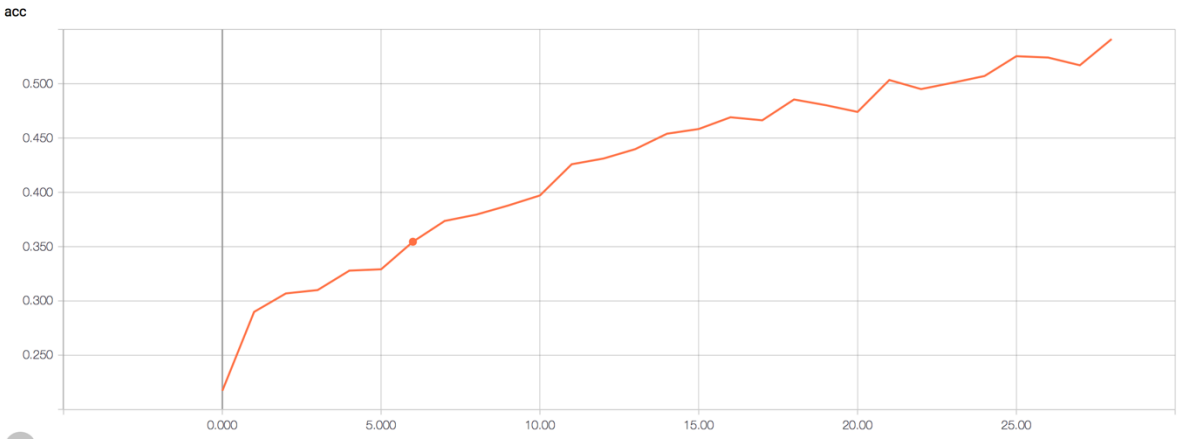
1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.
   我使用在ImageNet上Pre-trained的ResNet50 model, 去除最後一層Dense(1000)和Softmax後，對於每個224*224*3的畫格能輸出一個2048維的feature vector。對於每部影片，我取首、尾以及中間的三個1/5等分點，共5個畫格，通入Pre-trained model後能輸出一個5*2048維的feature。為了讓其中每個2048維的frame feature經過同樣的處理、共用同樣的weight，我讓這個5*2048維的input經過兩個Stride為2的Conv1D(32) layer，分別接一個Dropout(0.3)，最後Flatten再經由Dense(11)和Softmax activation得到class預測值。

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 3, 32)	131104
dropout_1 (Dropout)	(None, 3, 32)	0
conv1d_2 (Conv1D)	(None, 2, 32)	2080
dropout_2 (Dropout)	(None, 2, 32)	0
flatten_1 (Flatten)	(None, 64)	0
dense_1 (Dense)	(None, 11)	715

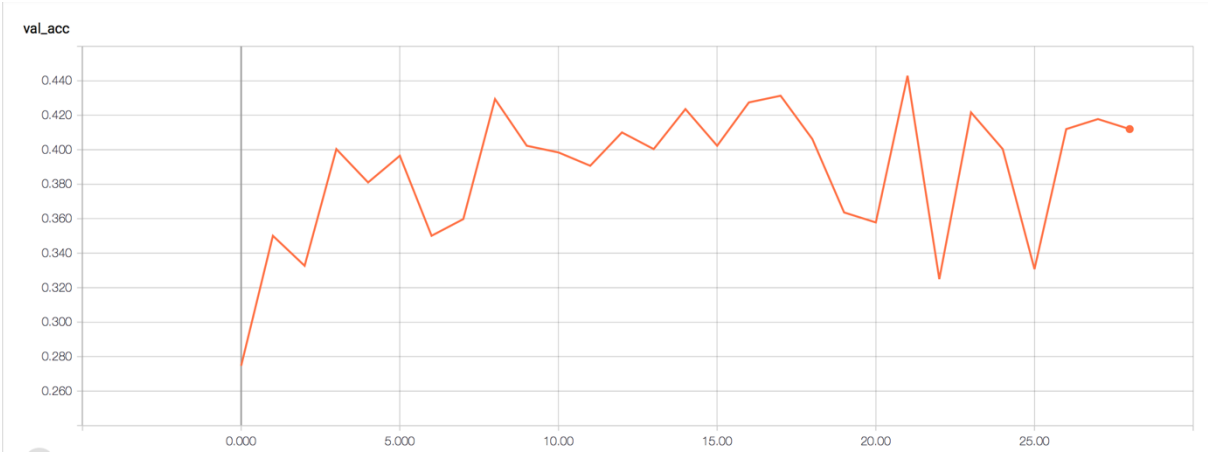
Total params: 133,899
Trainable params: 133,899
Non-trainable params: 0

2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.
在有Earlystop的機制下，經過28個Epoch的training，我得到0.412的Validation accuracy。以下分別為Training set和Validation set的Learning Curve:

a. Training Accuracy



b. Validation Accuracy



[Problem2]

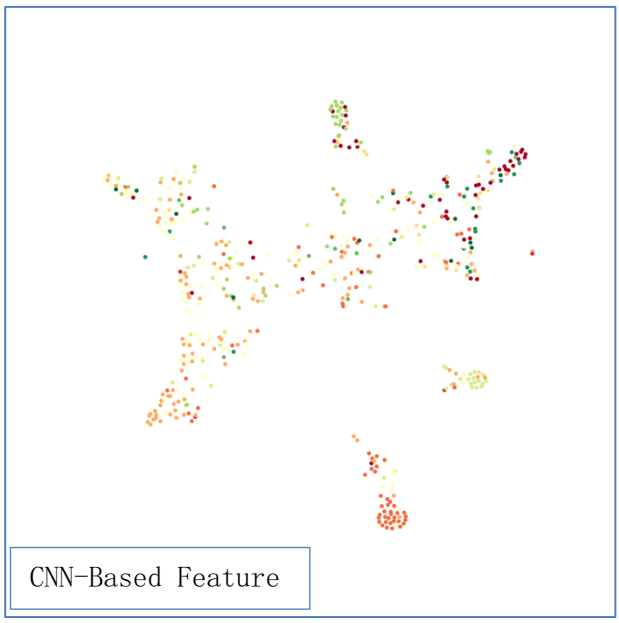
1. (5%) Describe your RNN models and implementation details for action recognition.

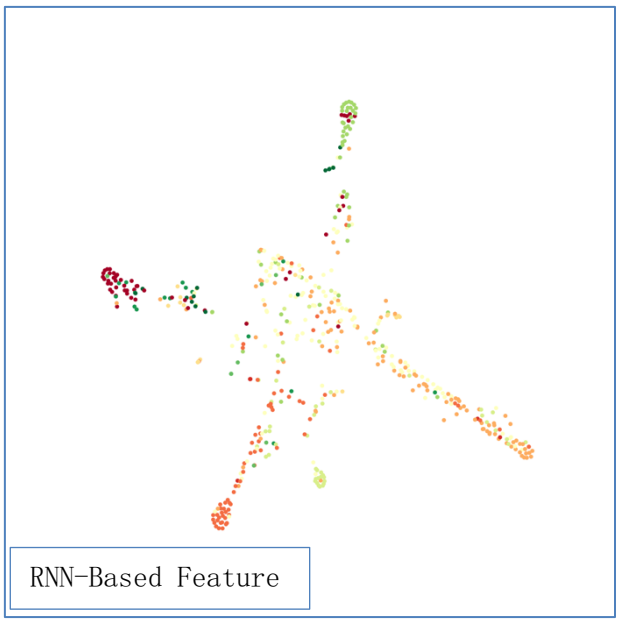
   我使用和Problem1相同的Pre-trained model來將輸入的frames轉為2048維的feature vectors，對於每部影片，我取前10張畫格(2 fps)，若少於10張則作zero-padding。因此每部影片的feature shape會是10*2048。

   Model方面，我使用雙層的Bi-directional LSTM(32)，加上Dropout(0.3)的機制，Train 19個epochs(有Early-stop)下可以得到0.5358的Validation Accuracy

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_1 (Bidirection (None, 10, 64)            532736
_________________________________________________________________
bidirectional_2 (Bidirection (None, 64)                24832
_________________________________________________________________
dense_1 (Dense)              (None, 11)                715
=================================================================
Total params: 558,283
Trainable params: 558,283
Non-trainable params: 0
_________________________________________________________________
```

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.





上圖是作在Validation Set的結果，由此結果我發現雖然兩圖中間都仍有許多難以分開的資料點，但相較於CNN而言，RNN的確可以把不同class的資料點分得較開。此外，在RNN的結果中，同一class的資料呈長條狀分佈，可說明RNN能較好地找出不同class資料間的過渡關係。

[Problem3]

1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.
在這個Task中，我並沒有更改model架構，而是以Sliding window的方式一次取10個frames喂進model做training和validation，而label則取這10個frame labels的眾數。

2. (10%) Report validation accuracy and plot the learning curve.
使用Task2的model為基礎，我可以達到0.499的Validation Accuracy

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).

[BONUS]