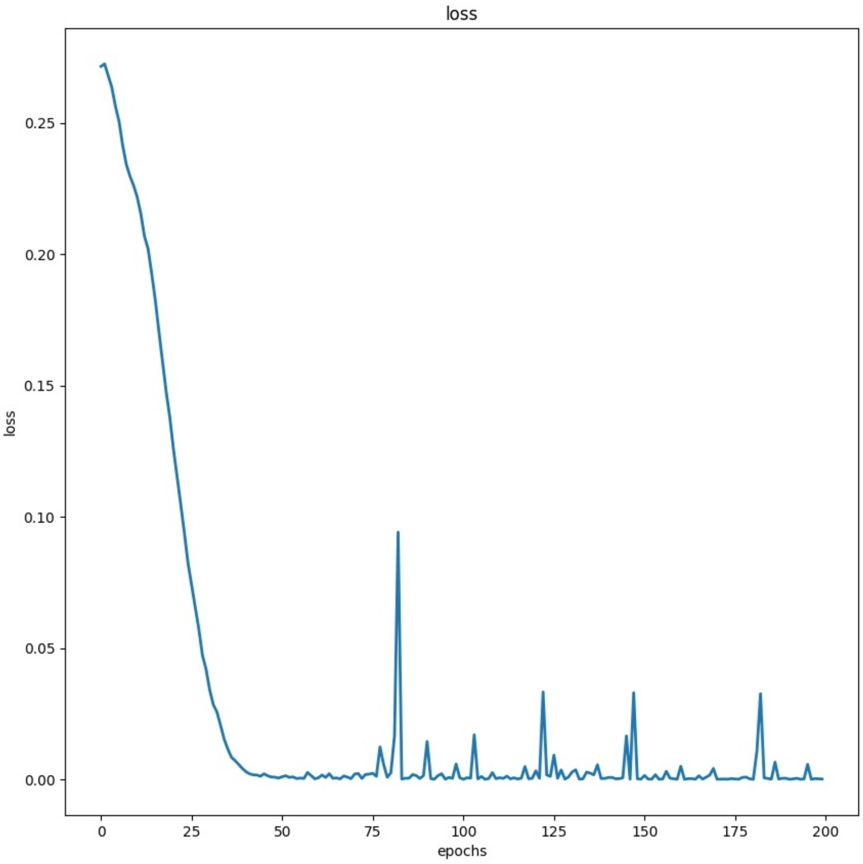
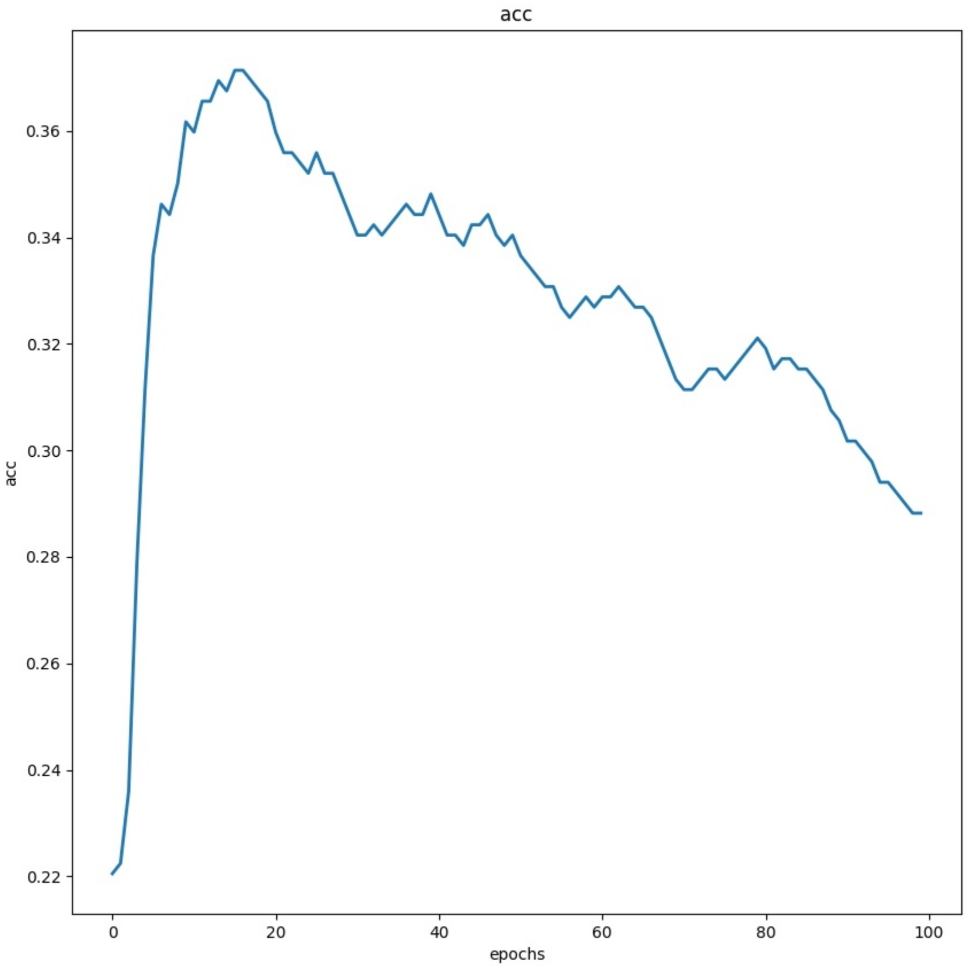
Please use this report template, and upload it in the PDF format. Reports in other format will result in ZERO point. Reports written in either Chinese or English is acceptable. The length of your report should NOT exceed 8 pages.

Name: 黃聖喻 Dep.:電機三 Student ID:b04901073

[Problem1]

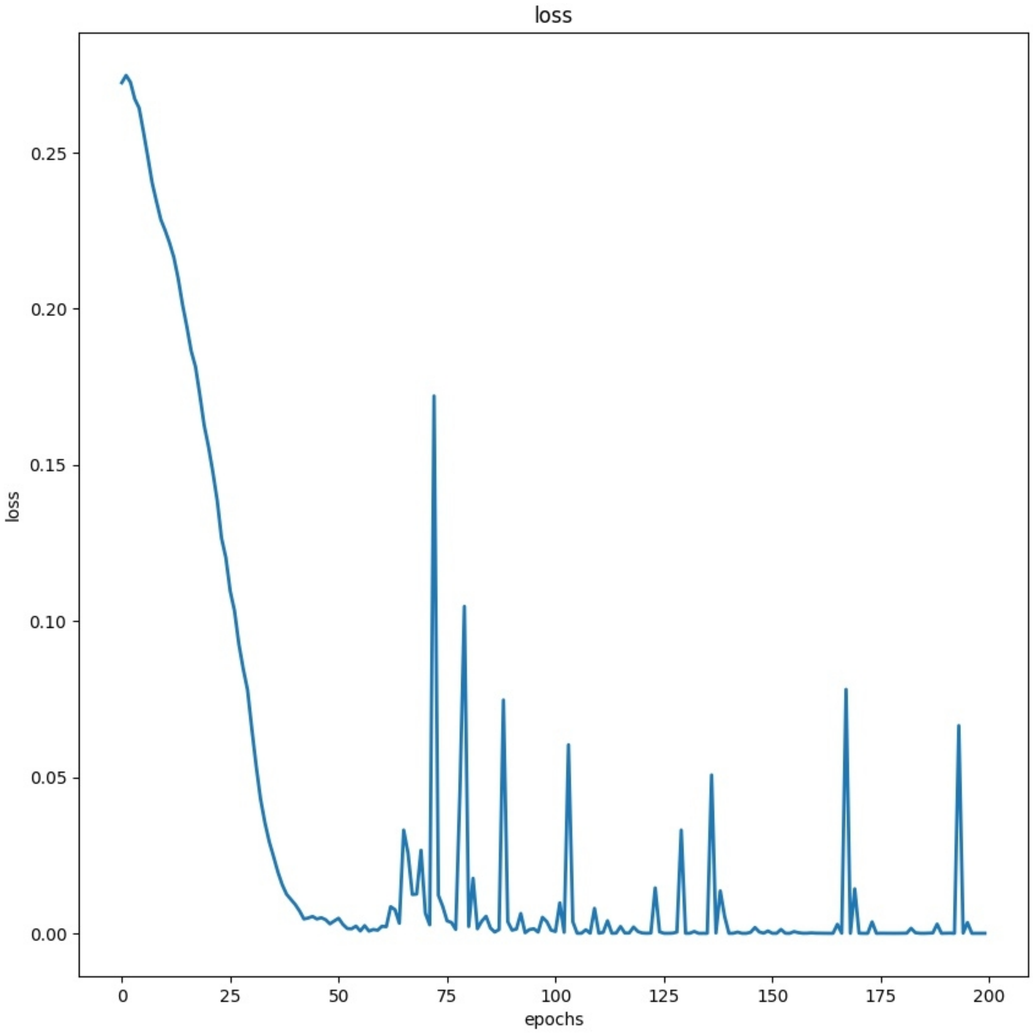
1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.
   先把每部影片取成四個 frame，然後將其取平均之後載入 resnet50，出來後把娶到的 feature 展開成 4096 維，然後用兩個 fc 層降成 1024 再降成 11 維，就變成輸出的 label 了。使用的 optimizer 是 Adam(lr=0.0001)
2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.
   最好的準確度有到 0.37 左右，不過隨著 epoch 次數增加就有明顯的 overfitting，直到 100 個 epoch 結束前都準確度都還在降。

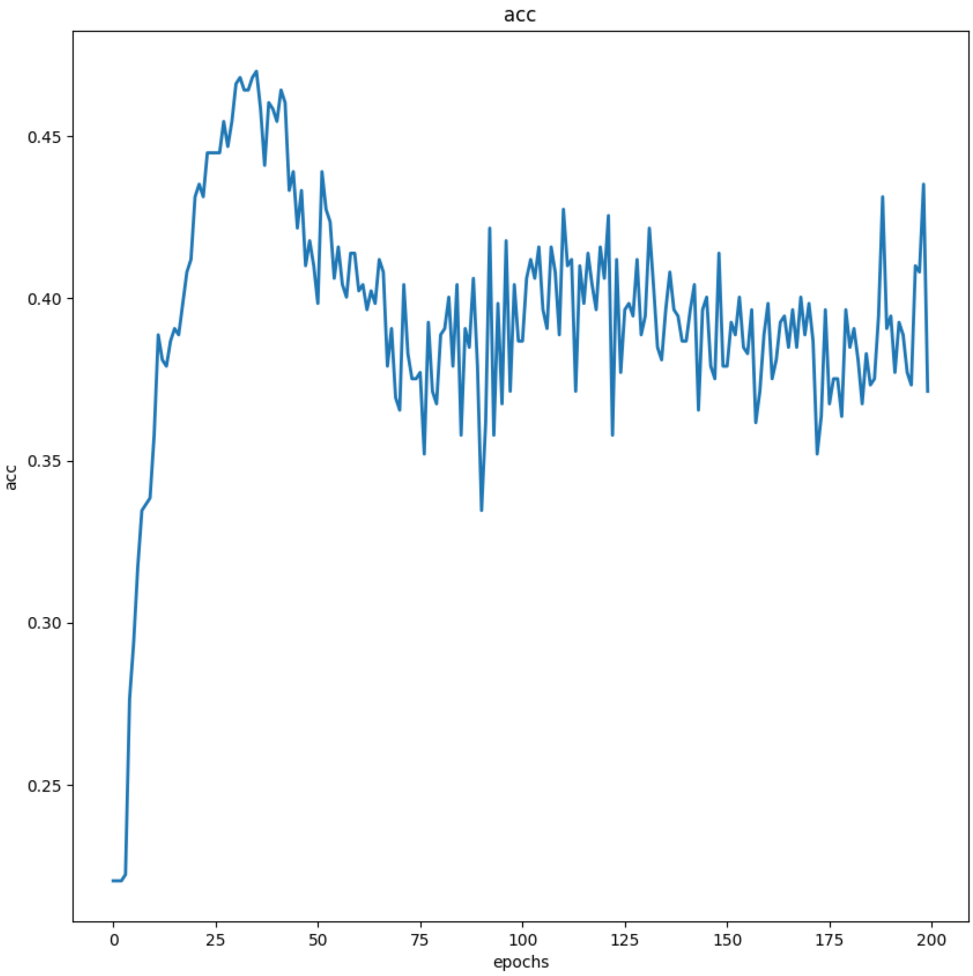




[Problem2]

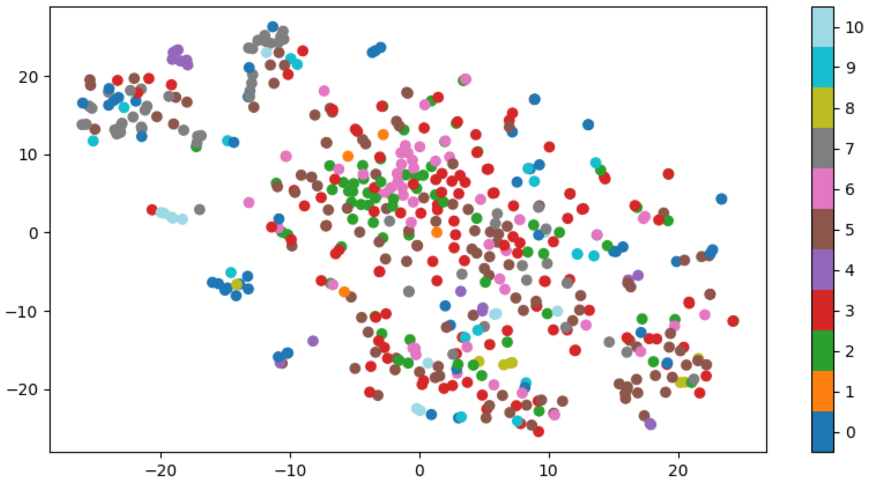
1. (5%) Describe your RNN models and implementation details for action recognition.
首先取每部影片改成 16 個 frame，本來只有取 4 個或 8 個的時候結果很差，改成 16 之後一次就過 baseline 了。接著把取好的 frame 給前面的 resnet50 產生 feature 之後就是 RNN 每一步的 input，LSTM 的部份是用單層的雙向結構把原本 4096 維的 feature 輸出成 512*2 維，然後藉由三個全連接層分別再降成 256、64 以及 11 維變成 predict label。optimizer 是 Adam(lr=0.0001)，batch size 則是 8。最好的準確度到達 0.47，相較第一題的模型有顯著的進步。


loss
loss
epochs
0.25
0.20
0.15
0.10
0.05
0.00
0
25
50
75
100
125
150
175
200

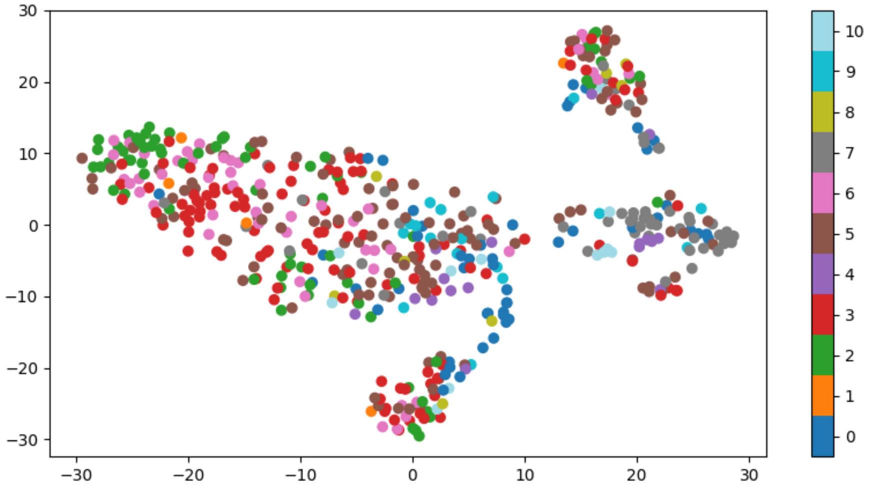

acc
acc
0.25
0.30
0.35
0.40
0.45
0
25
50
75
100
125
150
175
200
epochs

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 tSNE



RNN tSNE



[0:Other 1:Inspect/Read 2: Open 3: Take 4: Cut 5: Put
6: Close 7:Move Around 8: Divide/Pull 9: Pour 10: Transfer]
CNN 的 feature 看上去很亂，好像只有 close 和 open 比較聚集，但是兩者的分別又不大，可能是由於從高維投影下來的關係，導致本來可能有些區分最後卻還是混在一起了；至於 RNN 的部份就有比較明顯的分群，不過其實預測的準確度本來就也不到 50%，所以大致上看起來同類的都在一起，每一類卻還是有不少與大群分離的點。

[Problem3]

1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.
   我的想法是把第二題的模型做一點簡單的修改，改成每一個 step 都輸出一個結果去做 label 的預測。

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

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]