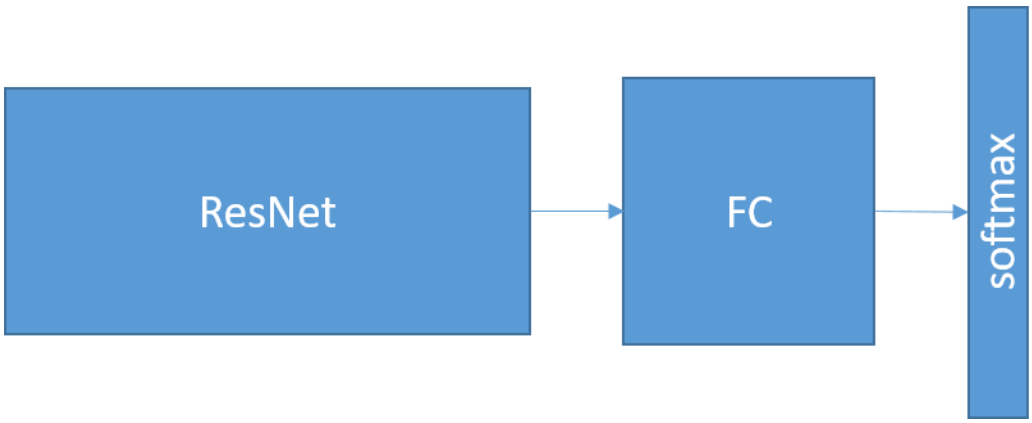
Name: 籃聖皓　Dep.:電信碩一　Student ID:R06942143

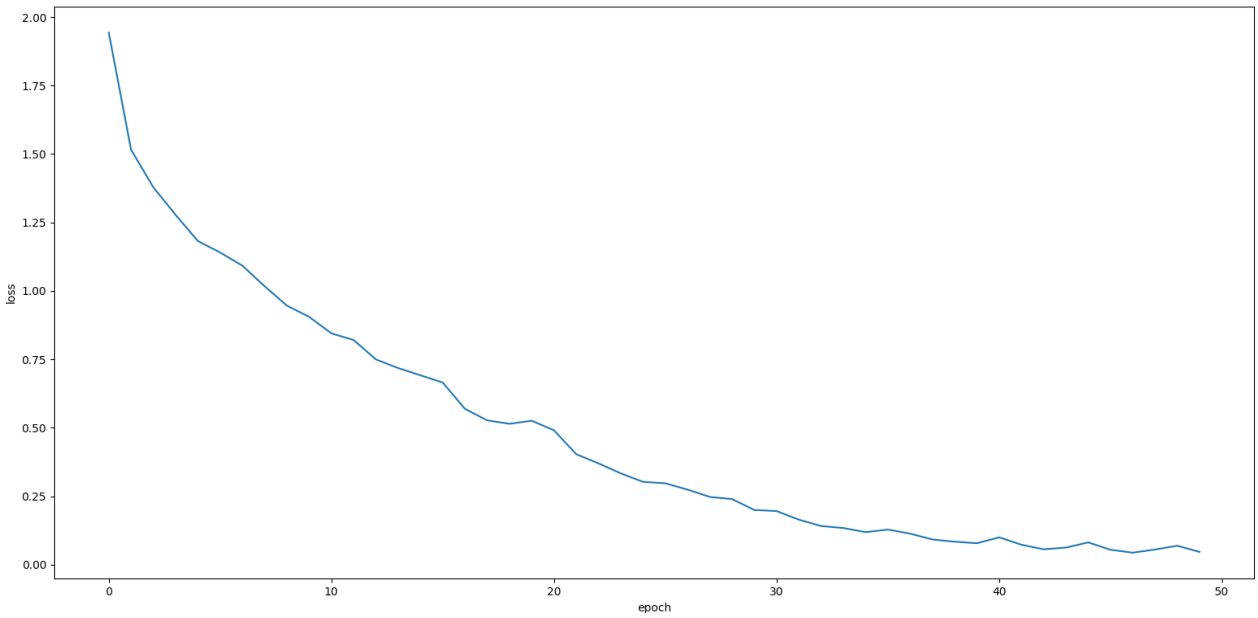
[Problem1]

1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.
   使用 ResNet，來自於 keras 的 pre-train model，一開始將影片切出來的 frame 都放入 resnet 提取特徵，提取出 2048 維度的特徵值，參照投影片上的方式，接一個 fully connected 的 DNN，最後接上一個 softmax 分成 11 個類別，而其中是採取 average pooling 的方式。

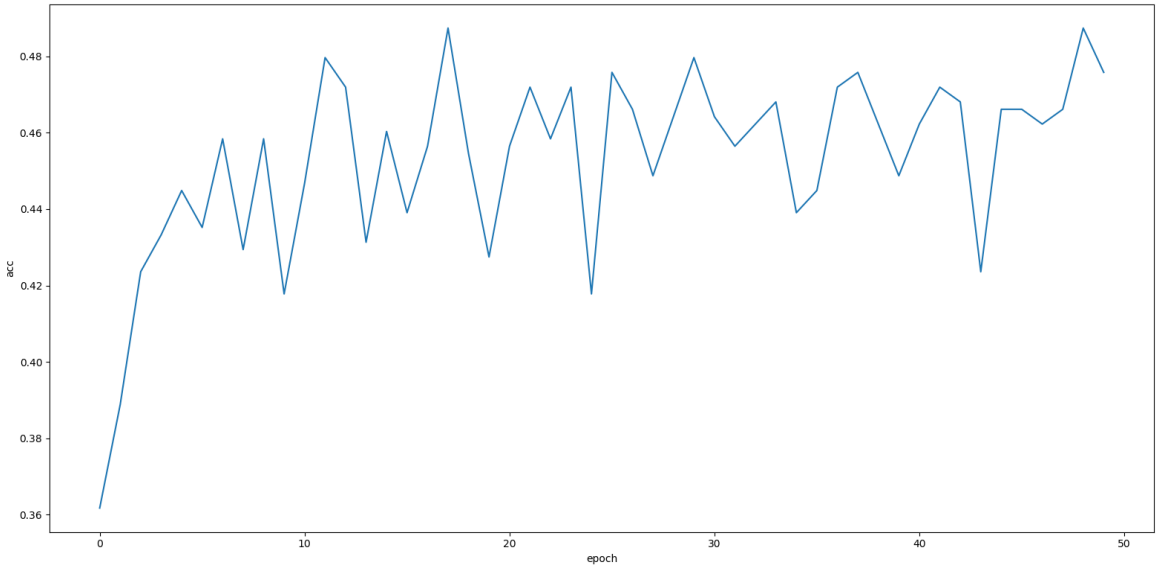


   Optimizer: adam
   FC: Dense(512)+Dense(11,Softmax)

2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.
   最後的準確度 validation acc 為 0.4738878143133462
   到了後面不論後面 DNN 怎麼調整準確度的結果都差不多，如果要再提高準確度可能要改 pretrain model。



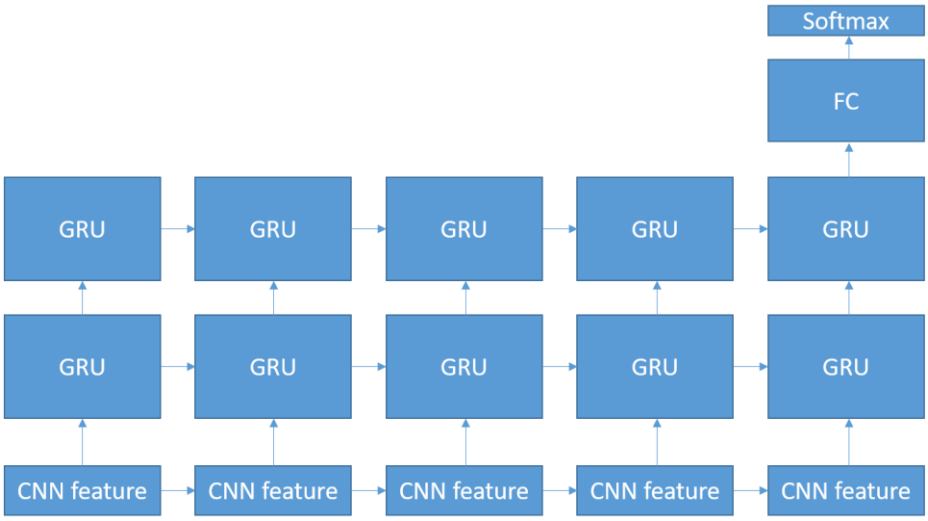
Train loss



Val_acc

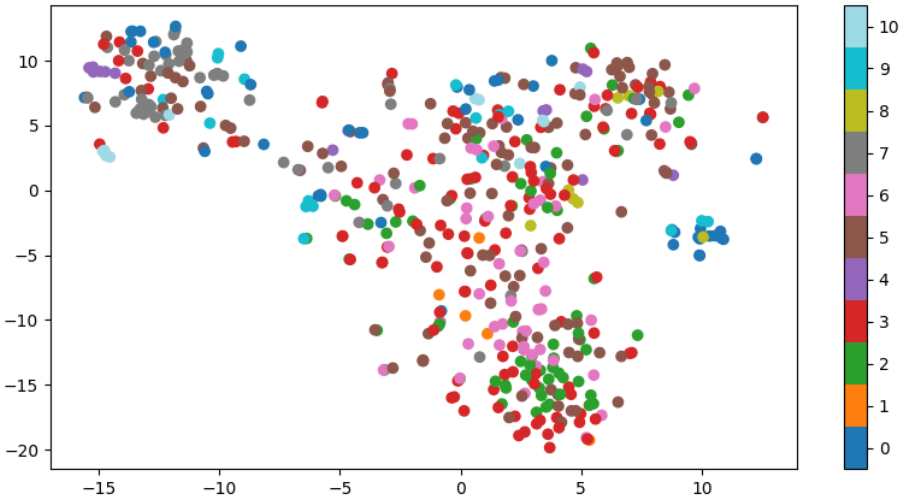
[Problem2]

1. (5%) Describe your RNN models and implementation details for action recognition.
這邊把原本用來提取 CNN 的 resnet 提取過來，而對於影片長度不一樣的問題，事都採取 zero padding，把長度都調整到最長的長度，接著把提取出來的 CNN feature 接入 RNN 之中，然後最後判斷結果。如下圖：

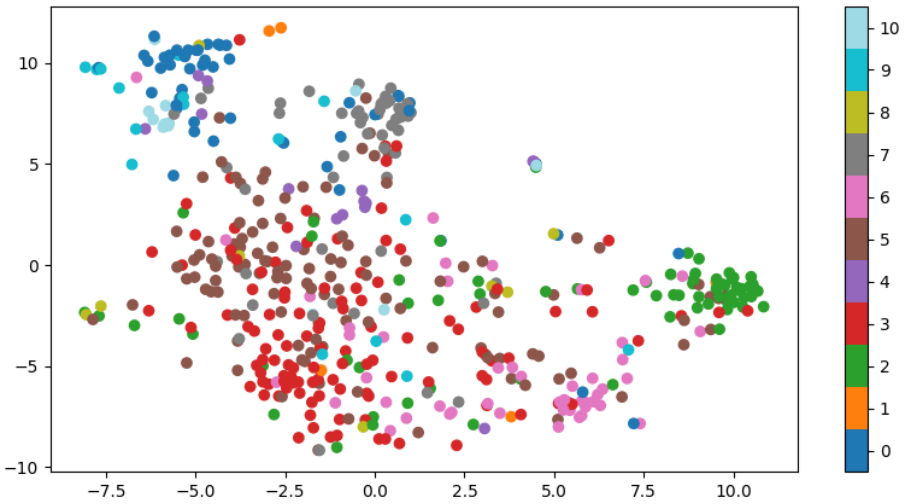


是使用 bidirectional 的 GRU 的架構，GRU 的 cell 都用了 512。

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 average pooling feature



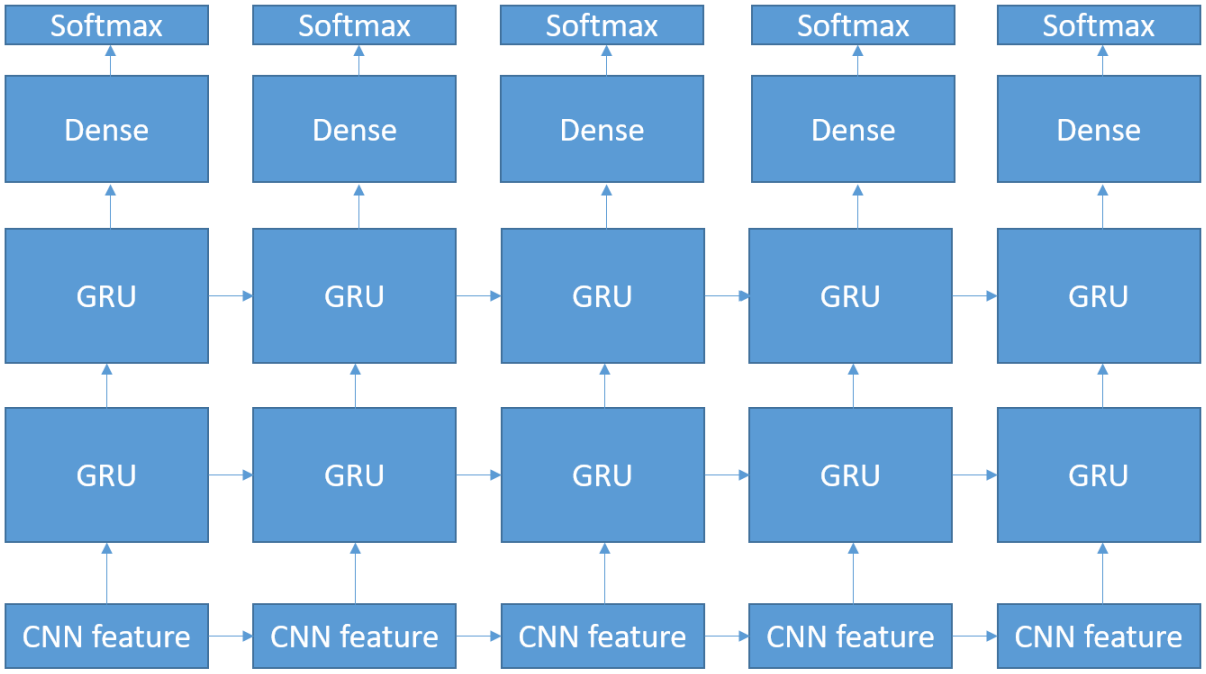
RNN feature

比較這兩張圖以上面 CNN 分出來的結果來看紅色，其實是分布在很多地方的，這樣並不好去區別，而在 RNN 之中可以看到紅色比較集中在左下，在其他的 label 之中也可以觀察到這些，綜合起來 CNN 分鐘來的特徵是比較分散的，雖然可能可以看出有一些 label 有些集中，不過還是整體糊在一起的。

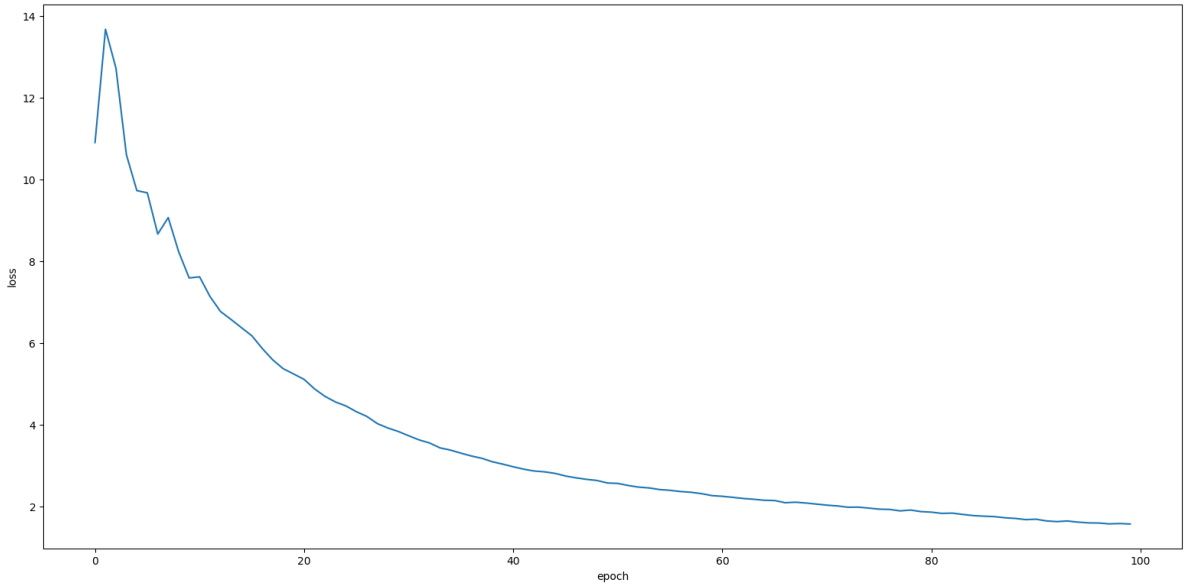
而再看下面這張 RNN 的圖，可以很明顯的看出來，有些顏色像是綠色就很明顯的集中，灰色也是，可以看出來 RNN 分出來的樣本是比較集中的，是有提升結果的。可以從這兩張 feature 的圖形比對出來，RNN 這個 model 分類出來的結果是好許多的。

[Problem3]

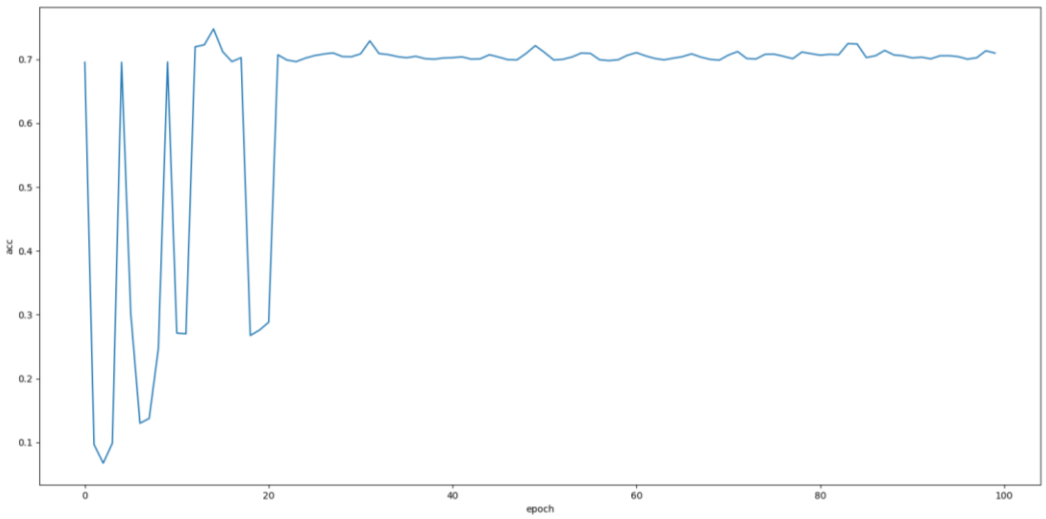
1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.
   以原本的第二題的 RNN 的架構，把其中的 sequence 都提取出來，並將每個 sequence 接上一個 label，這邊有在 GRU 之中多加入了 dropout，有提升 0.01 左右的效果不過沒有很顯著，架構如下圖。
   在訓練的時候，是每隔 2 個 epoch 就重新取一次連續的 500frame。
   是使用 bidirectional 的 GRU 的架構，GRU 的 cell 都用了 512。



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



Train_loss



Valid_acc

(這邊會跟下面的 val_acc 結果不大一樣，是因為這邊的 validaiton 的都經過 zero padding，因此這邊的結果會比較高一點，是因為後面 zero 的部分都可以完整地預測)

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).

下面是分別每個 val_acc 對應的結果，

OP01-R03-BaconAndEggs.txt	0.6341121495327103
OP02-R04-ContinentalBreakfast.txt	0.5874200426439232
OP03-R02-TurkeySandwich.txt	0.46557759626604434
OP05-R07-Pizza.txt	0.41656365883807167
OP06-R05-Cheeseburger.txt	0.5595588235294118
Average acc:	0.5326464541620323

以下是 label 與顏色的對應表

0 other 1 read 2 open 3 take 4 cut 5 put 6 close 7move 8 divide9 pour 10transfer

是挑選 OP01-R03-BaconAndEggs 之中的 16801.jpg~20401.jpg 的這 300 個 frame 的 label 結果，上面的是 groundtruth、下面的結果是 predict 的結果

可以看得出來這次訓練的模型，比較能處理的是連續的動作，像是幾個 frame 的很突然的動作，就沒辦法被抓出來，可能他打開東西之後移動，那很可能因為移動的 frame 比較多，就都判斷成移動了，但對於連續的動作有較好的結果。