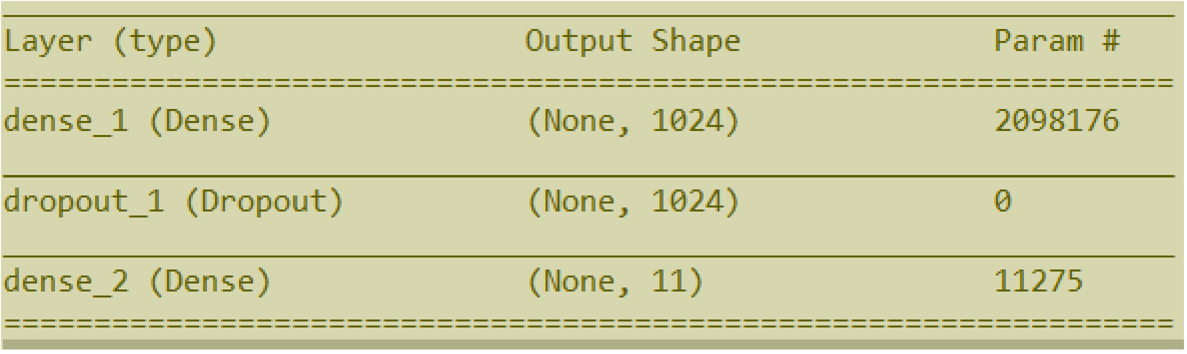
Please use this report template, and upload it in the PDF format. Reports in other format will result in ZERO point. Reports written in either Chinese or English is acceptable. The length of your report should NOT exceed 8 pages.

Name: 汪家銘　Dep.:電機四　Student ID:B03901111

[Problem1]

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



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 1024)              2098176
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 11)                11275
=================================================================
```

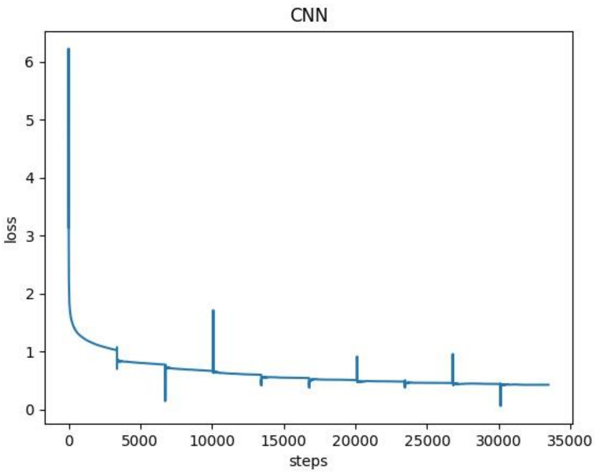
如上圖，我的CNN是將兩層的Fully-connected layer，換掉ResNet50的最後兩層，並freeze ResNet50的pretrain部分，並以Dropout避免過擬合。
之所以不考慮在後面繼續加FCN的原因是我們要輸出的是label而不是上次的圖，另一方面是要用CNN降到11維要很多層，且我這邊訓練效果不佳。

針對ResNet50要求輸入維度是(224, 224, 3)的部分，我在放入之前將每幀都做了手寫的random-crop(test的時候是centre-crop)，得到(224, 224, 3)的cropped-image，再通過ResNet50得到(2048)的CNN-feature，在接下來的所有問題中，我都是使用同一個feature。

Training detail：Optimizer = adadelta，Loss = categorical_crossentropy

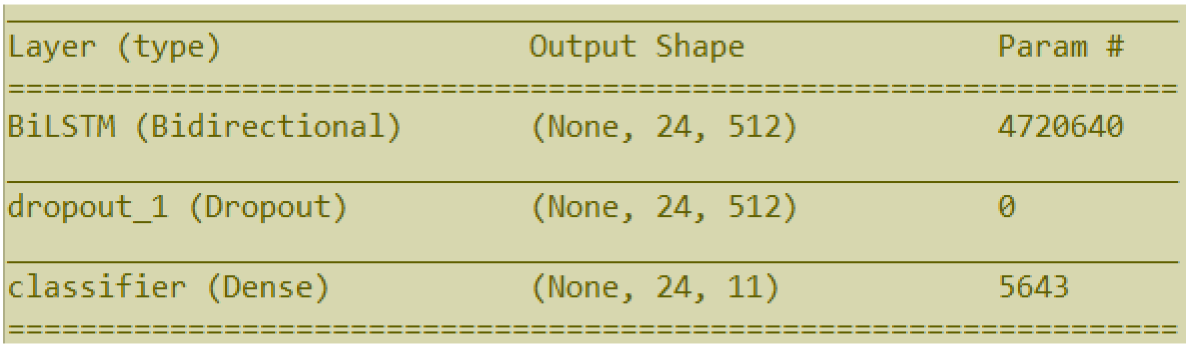
2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.

在本次作業中，我的 CNN 在 valid data 中逐幀的 ac.c 是 45%，而整部 trimmed video 的 acc.是 44.2%。learning curve 如下圖：(抖動是因為 ep 之間的斷層)



[Problem2]

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



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
BiLSTM (Bidirectional)       (None, 24, 512)           4720640
_________________________________________________________________
dropout_1 (Dropout)          (None, 24, 512)           0
_________________________________________________________________
classifier (Dense)           (None, 24, 11)            5643
=================================================================
```

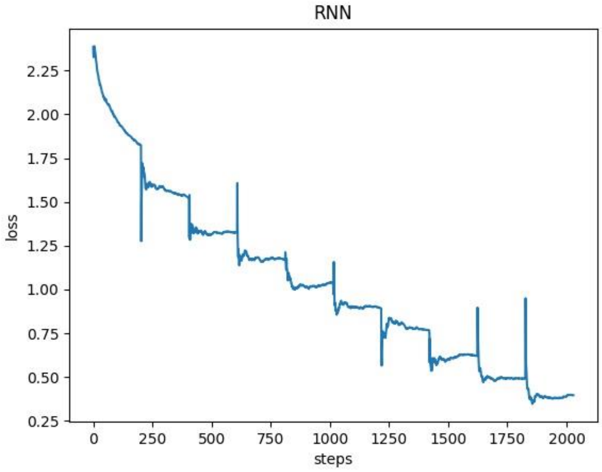
如上圖, 我的 Trimmed RNN 架構是 Bidirectional-LSTM (24 nodes 的 hidden layer), 最後再外接一層 Fully-connected layer,產生 one-hot labels.

我的前處理是首先通過 ResNet50 取 2048 維度的 feature。再設定一個 maxTime(=16)，將長度超過 maxTime 1.5 倍(24)的影片遞迴對半至 16 以下，最後 zero-padding 產生(517, 24, 2048)的 data。與此同時，對 label 也做同樣處理，shape 為(517, 24, 11)

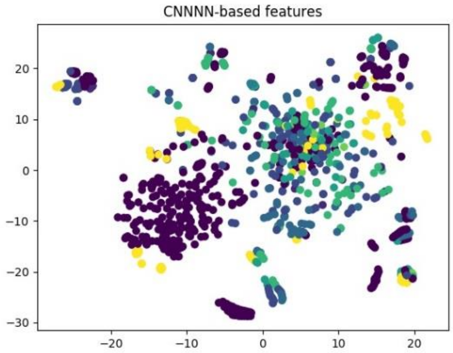
通過 RNN 後每個影片產出的是(24, 11)維的 predict，和 label 做 cross-entrophy 得到 loss，Optimizer 為 adadelta。

在 predict 的時候，將 predict 去掉 padding 部分(之前有存)，再取各幀平均取 argmax 得到 predict-label，最終的 trimmed-acc 為 0.51，有超過 baseline 0.45。

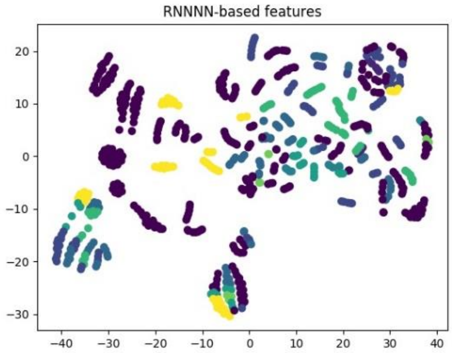
附上 learning curve 如下：



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投影

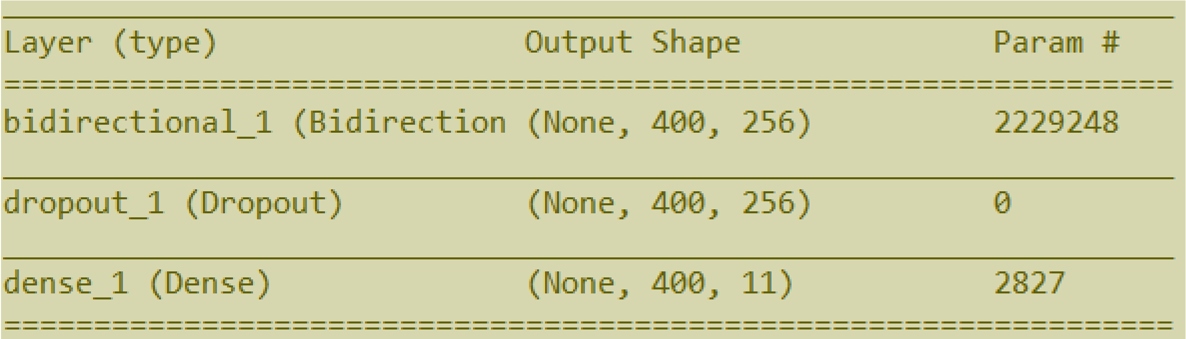


RNN投影

由上圖可看出，RNN由於是train-by-film，輸出的feature也呈現線狀，由這種線狀連接把同一個class的data弄到了一起，從而增加了acc.。反觀CNN就比較趨近多點圓分佈，各label之間就分不開。

[Problem3]

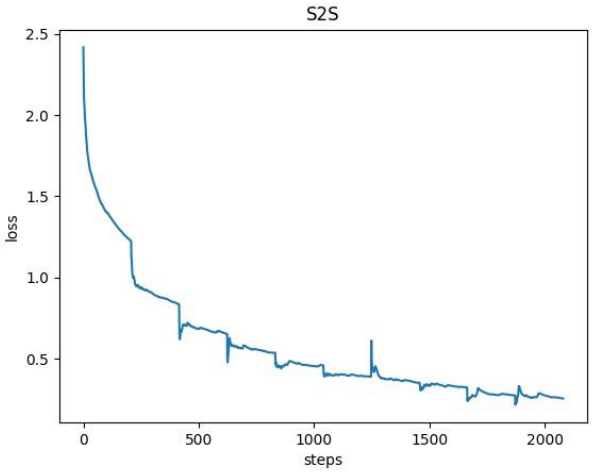
1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
bidirectional_1 (Bidirection (None, 400, 256)          2229248
_________________________________________________________________
dropout_1 (Dropout)          (None, 400, 256)          0
_________________________________________________________________
dense_1 (Dense)              (None, 400, 11)           2827
=================================================================
```

如上圖,我的模型是一層的 Bidirectional-LSTM (400 nodes 的 hidden layer),最後再外接一層 Fully-connected layer,產生 one-hot labels

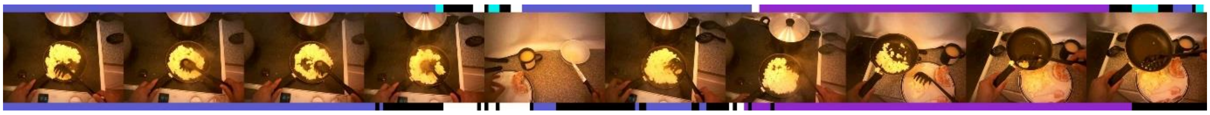
2. (10%) Report validation accuracy and plot the learning curve.



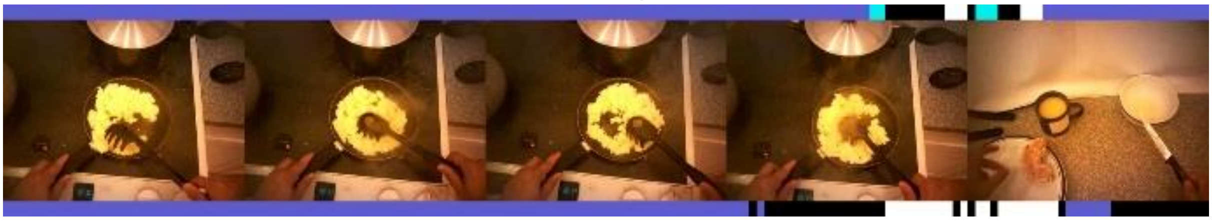
對五個 full-video，acc 分別是：0.60, 0.68, 0.61, 0.56, 0.51，平均正確率是 0.61

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).

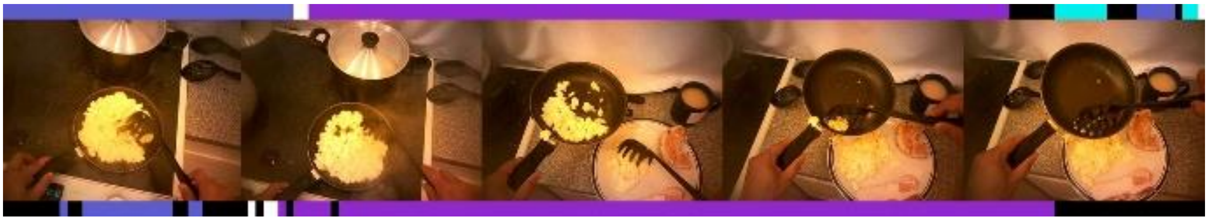
   我截取了影片中間的 320 幀，片段正確率為 65%的片段如下：



full-image



half-image(1)



half-image(2)

我截取這部分的原因是，這段 label 的多樣性比較豐富。有些片段最高 acc.可以達到 70%以上，但那些片段幾乎都是單一 label 的片段。

從我截取的部分可以看出，RNN 對長時間的片段預測準確性較高，但如果動作頻繁切換，RNN 預測正確性就會迅速下降（看中間淺藍和白色部分）

[BONUS]

NONE