Name: 何明倩　Dep.:電信碩一　Student ID:R06942039

[Problem1]

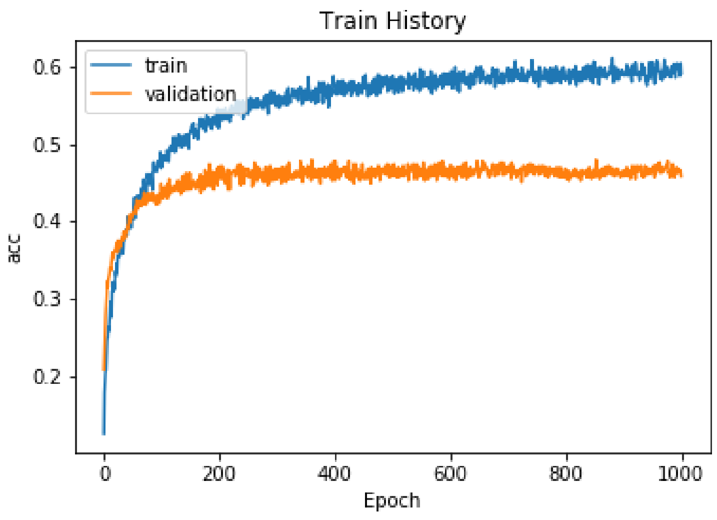
1. (5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.
   使用 keras 中內建的 ResNet50 來取出 CNN_based features (2048 維)，再將每個 video 取出來的數個 features 取平均。得到代表每個 video 的 feature 後，在設計一個 DNN 的 classify 以這些 features 當 input。
   下面是我設計的 DNN model：

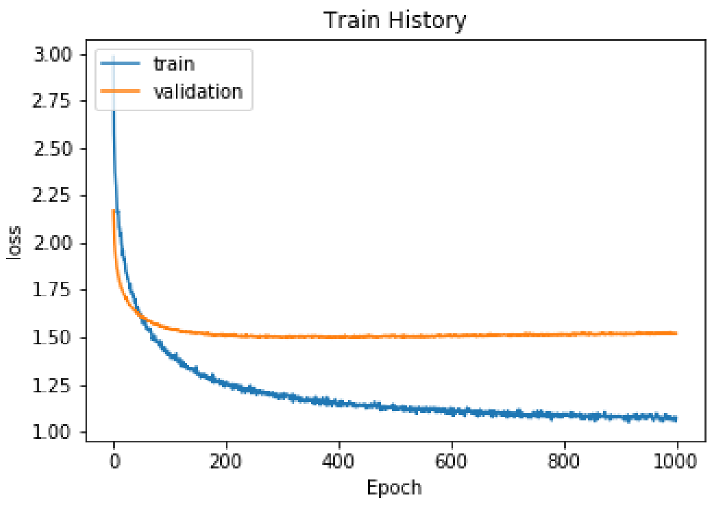
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
DNN_input (InputLayer)       (None, 2048)              0         
_________________________________________________________________
dropout_3 (Dropout)          (None, 2048)              0         
_________________________________________________________________
DNN_output (Dense)           (None, 11)                22539     
=================================================================
Total params: 22,539
Trainable params: 22,539
Non-trainable params: 0
_________________________________________________________________
```

   這裡的 dropout 為 0.5，選擇的 optimizer 為 Adam，loss 選擇 categorical_crossentropy。

2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.
   P.S. accuracy 是取到數點下第二位四捨五入，我選擇第 361 epoch

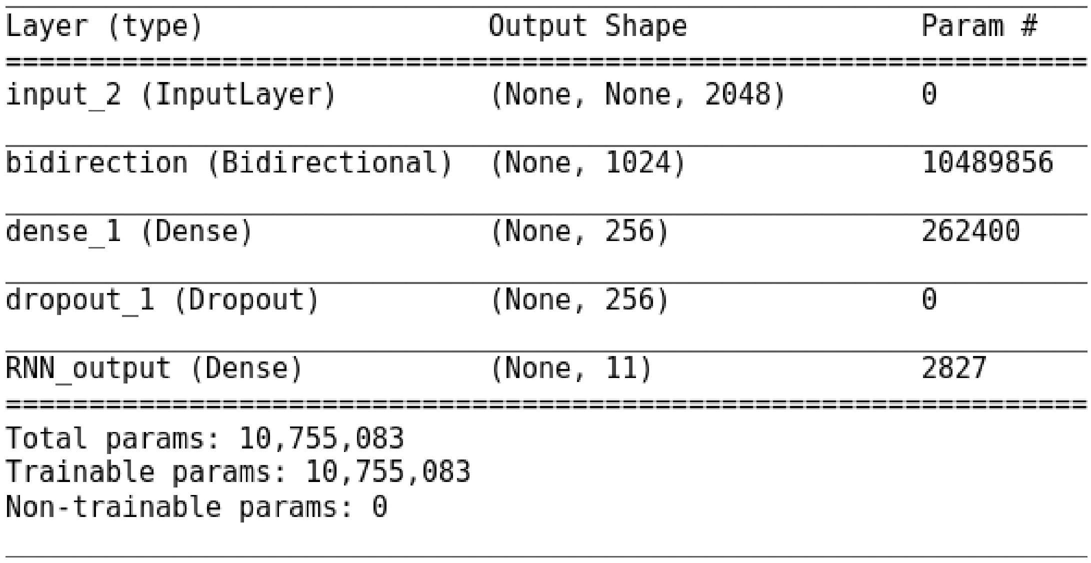
	Training set	Validation set
accuracy	0.64	0.48





[Problem2]

1. (5%) Describe your RNN models and implementation details for action recognition.
   利用 keras 中內建的 ResNet50 來取出 CNN_based features (2048 維)，但是每個 video 取出來的 features 數量不同，所以我就利用 keras 內建的 sequence padding 來做 zero padding，將每個 video 的 features 數量都 padding 成跟最多 features 的 video 的數量一樣(time_steps=234)。底下我的 RNN model 架構：

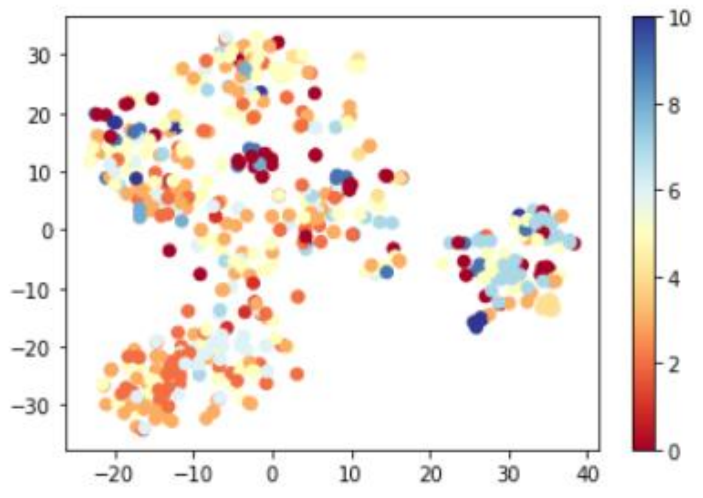
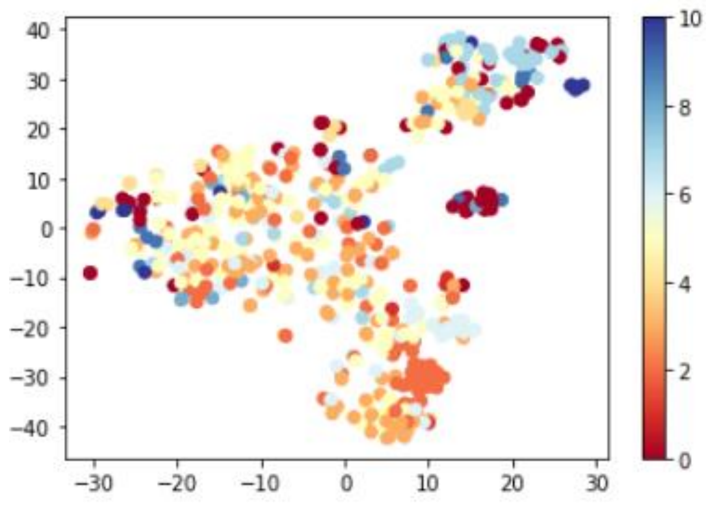


```
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         (None, None, 2048)        0

bidirection (Bidirectional)  (None, 1024)              10489856

dense_1 (Dense)              (None, 256)               262400

dropout_1 (Dropout)          (None, 256)               0

RNN_output (Dense)           (None, 11)                2827
=================================================================
Total params: 10,755,083
Trainable params: 10,755,083
Non-trainable params: 0
```

   這裡的 dropout 為 0.2，選擇的 optimizer 為 Adam，loss 選擇 categorical_crossentropy。

	Validation set
accuracy	0.55

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-based video features	RNN-based video features
	

根據這兩張圖片我們可以發現 CNN、RNN_based features 的差異並沒有很大，但是在 validation 上 accuracy 可以發現 RNN_based features 的 accuracy 高出 0.07 (RNN model 在 validation 上的 accuracy 是 0.55)，可能在 accuracy 上沒有差異太多，所以 RNN_based features 才沒有比 CNN_based features 更有群聚的現象。

[Problem3]

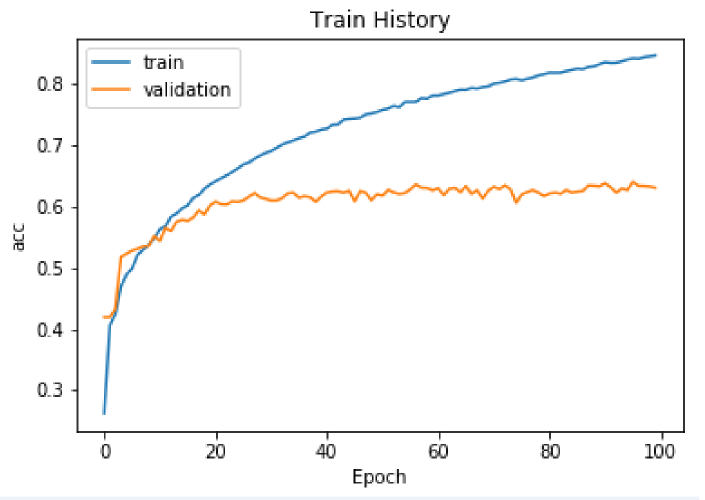
1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.
   利用 keras 中內建的 ResNet50 來取出 CNN_based features (2048 維)，我抉擇的 time_step 為 250，每個 video 每隔 50 frames.就會取一次 250 個 features，這樣做會比直接每 250 個 features 切成一個 sequence 還能讓 model 學到 frame 和 frame 之間的關係，因為直接每 250 個 features 切成一個 sequence 的話，就會讓前後的 sequence 失去相關性，這樣也來也能增加許多的 training dats。
   底下是我設計的 RNN model：

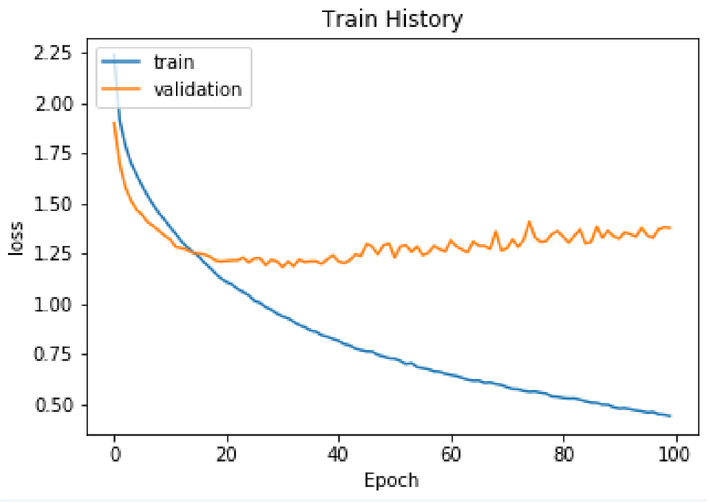
```
Layer (type)                 Output Shape              Param #
=================================================================
RNN_input (InputLayer)       (None, None, 2048)        0
_________________________________________________________________
bidirection (Bidirectional)  (None, None, 1024)        10489856
_________________________________________________________________
dense_1 (Dense)              (None, None, 256)         262400
_________________________________________________________________
dense_2 (Dense)              (None, None, 32)          8224
_________________________________________________________________
dropout_4 (Dropout)          (None, None, 32)          0
_________________________________________________________________
RNN_output (Dense)           (None, None, 11)          363
=================================================================
Total params: 10,760,843
Trainable params: 10,760,843
Non-trainable params: 0
_________________________________________________________________
```

   這裡的 dropout 為 0.2，選擇的 optimizer 為 Adam，loss 選擇 categorical_crossentropy。
2. (10%) Report validation accuracy and plot the learning curve.

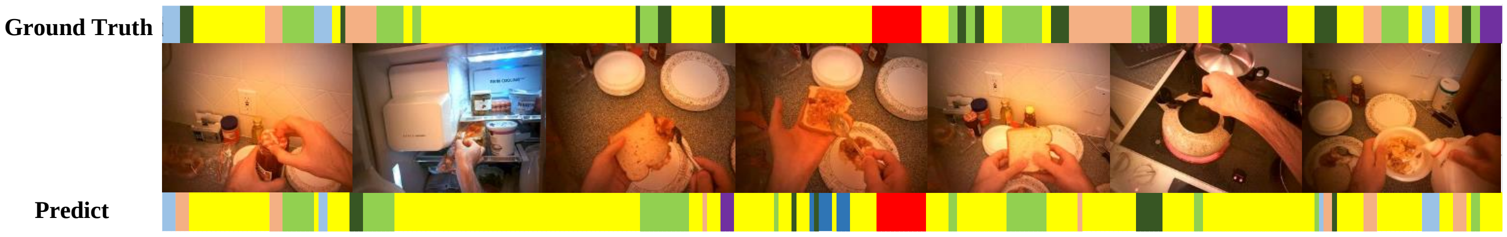
   P.S. accuracy 是取到數點下第二位四捨五入

	Validation set
accuracy	0.66





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).



label	0	1	2	3	4	5	6	7	8	9	10
color	🟨	🟥	🟧	🟩	🟩	🟩	🟦	🟦	🟦	🟪	🟪

從 visulization 的結果可以發現有很多 frame 很容易被誤判成 label 0 那是因為 label 0 的 frame 有很多且 label 0(other) 有很多不同類型的 frame 導致 model 容易將 frame 誤判成 label 0。像是 label 9 (pour)就全部都被誤判成 label 0 (other)。label 3(Take)和 label 5(Put)也容易誤判，可能是動作很類似，容易造成誤判。

[BONUS]