

DLCV HW5

Name: 王冠驊 Dep.:電信碩二 Student ID:R05942102

[Problem1]

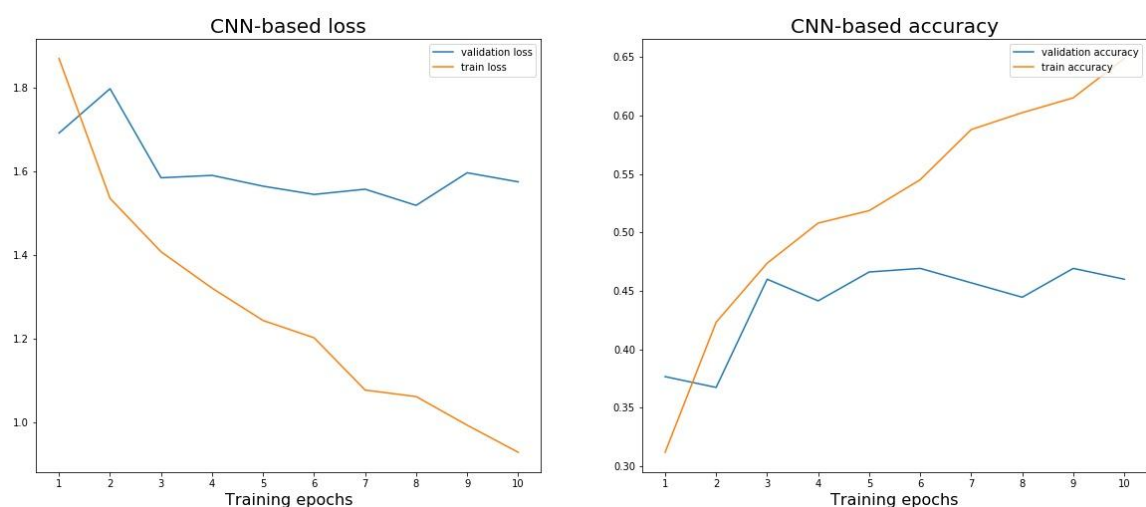
(5%) Describe your strategies of extracting CNN-based video features, training the model and other implementation details.

我們使用 pre-trained 的 ResNet50 對每一部 video 中的 frame 抽取 feature，最後我們將同一部 video 的所有 frame feature 取平均當作此 video 的 feature，這個方法雖然簡單，但在產生 video feature 的同時也損失不少原本在 frame feature 資訊。我們使用 3 層的 dense layer 針對將每一個 video feature 做分類。在訓練的過程中，我們不會去 fine-tune ResNet50。

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

最終，我們在助教提供的 valid videos 上測試；accuracy 為 0.47195。

下圖為 model training 的 learning curve。(圖中的 validation data 是從助教提供的 train videos 中切出來的，與上面測試所使用的 valid videos 不同)



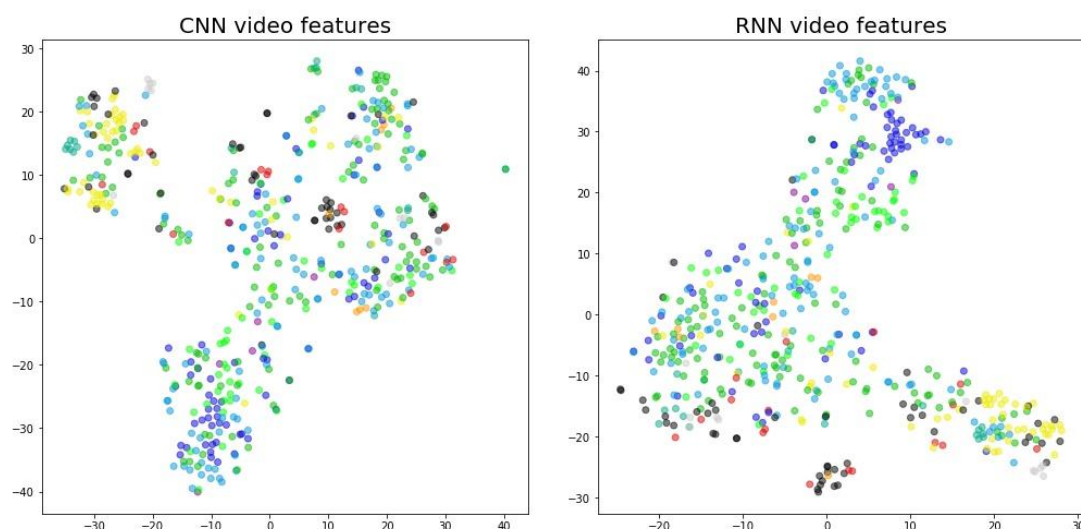
[Problem2]

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

我們使用與在 CNN-based model 相同的方法抽取 frame feature。不同的是，RNN model 可以在不同的 time step 吃入不同的 frame feature。如此一來，RNN model 就可以完整的獲得所有包含在 frame feature 內的資訊。在這裡，我們使用 bidirectional 的 LSTM，並取出最後一個 frame 輸入後的 hidden state 視為整個 video 的 feature。與 CNN-based 相同，我們使用 3 層的 dense layer 針對將每一個 video feature 做分類。在訓練的過程中，我們不會去 fine-tune ResNet50。

(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.

下面為 CNN-based video features 與 RNN-based video features 的視覺化結果。



從結果看來，分群的效果並不是太好，有比較明顯的分群類別只有圖中的黃色、藍色以及黑色。並且 CNN-based video features 與 RNN-based video features 在分群效果上差異並不大。RNN model 在 valid videos 上測試；accuracy 為 0.4893，與上一題的 0.47195 差異也不大。推測原因可能為在這邊的每一部 video 都為同一個 action，故在裡面的每一個 frame 應該都帶有相似的 feature，若我們對所有的 frame feature 取平均並不會損失太多的資訊。

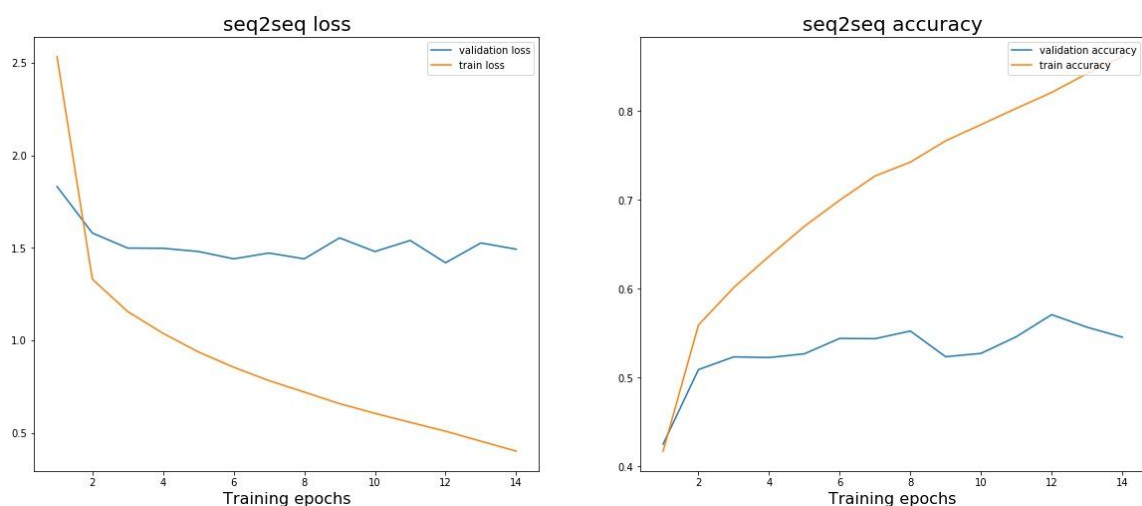
[Problem3]

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




在 Problem3 所使用到的 RNN model 架構與 Problem3 的相同。然而，在 Problem3 中，我們會讓 RNN model 在每一個 time step 都吐出它的 hidden state。在取 training data 的時候，我們會以 200 個 frame 為一個間隔，取出一筆長度為 400 (maximum time step) 個 frame 的資料當作一筆 training data。最終我們會得到 512 筆 training data。我們也嘗試使用更小的間隔來增加最後 training data 的數量。但經過實驗後，我們發現這樣的設定會得到較低的 validation accuracy。

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

我們在 valid videos 上測試的 accuracy 為 0.5819。



(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).

Ground Truth	
Frames	
Prediction	

從結果看來，有比較明顯大部份 **match** 到的部分的為圖中顯示紫色的地方，其 **label** 為 “Other”。除此之外，其他 **ground truth labels** 的分部則大多較為零碎，**model** 雖然沒辦法準確的抓到不同 **label** 變換時的邊界，但是在大致的位置都有預測出該 **label**。