Deep Learning for Computer Vision

HW5 Report

Name: 張景程　Dep.: 電機三　Student ID: B04901138

[Problem1]

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

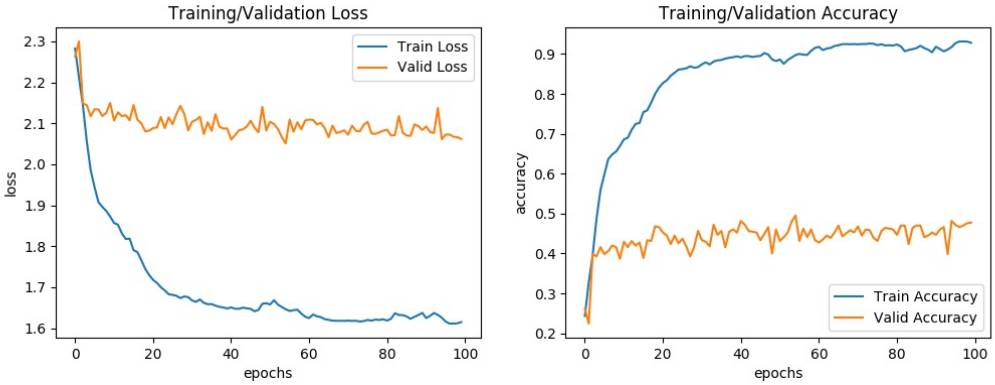
   在 Preprocess 的部分，首先利用助教提供的 code (fps=2) 將影片 downsample 成一系列的 frame，同時將這些 frame resize 成(3, 224, 224) 並 normalize 到 (0,1) 之間，以便進入後面的 pre-trained model。

   Pre-trained model 的部分，我選擇直接利用現成的 VGG-16，得到 size 為 (7, 512, 512) 的 output，經過 average pooling 後直接 flatten 並丟入後面的 fully-connected layer，架構如下：

```
CNN_Classifier(
  (main): Sequential(
    (0): Linear(in_features=25088, out_features=4096)
    (1): ReLU()
    (2): Linear(in_features=4096, out_features=1024)
    (3): ReLU()
    (4): Linear(in_features=1024, out_features=256)
    (5): ReLU()
    (6): Linear(in_features=256, out_features=11)
    (7): Softmax()
  )
)
```

   在實作上我並沒有 fine-tune 前面 VGG-16 的參數，單純只有 train classifier 的部份，optimizer 使用 Adam，並 train 了 100 個 epoch。

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



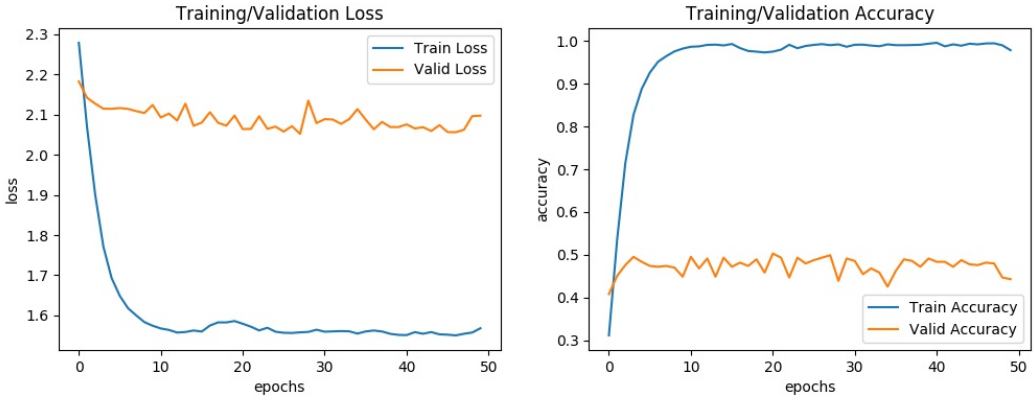
validation accuracy : 0.4951644100580271
從 learning curve 可看出，大約在 20~30 個 epoch 左右，loss 和 accuracy 就沒有太大變動，且有嚴重 overfit 的情形，推測是 training 和 validation 分布差異所致。

[Problem2]

1. (5%) Describe your RNN models and implementation details for action recognition.
Preprocess 的部分，和 Problem 1 的方式幾乎相同，唯一差別只有通過 VGG-16 之後，由於後面是要丟入 RNN，因此必須將 average pooling 移除，後面的 RNN 架構如下：

```
RNN_Classifier(
  (lstm): LSTM(25088, 512, num_layers=2, dropout=0.5)
  (bn1): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True)
  (fc1): Linear(in_features=512, out_features=11)
)
```

將 CNN 抽取出的 feature 丟入簡單的 LSTM 後，把最後一個 timestep 得到的 hidden state 丟入 fully-conneted layer 來進行 classification，optimizer 使用 Adam，batch size 設為 32，並 train 了 50 個 epoch。
Training Process 如下：

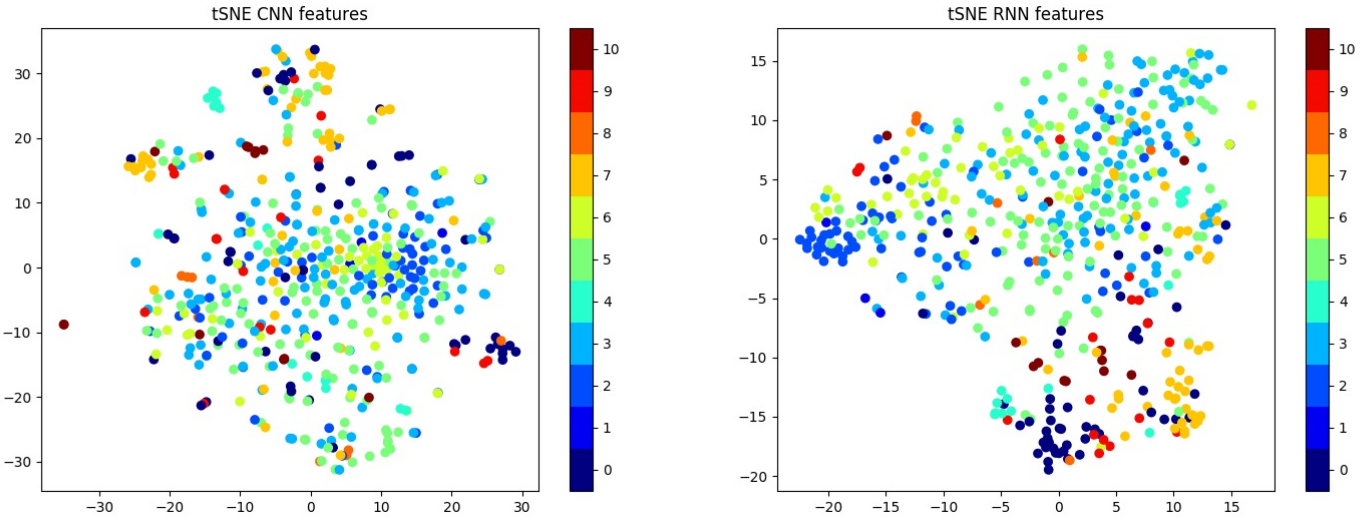


validation accuracy : 0.5183752417794971
(azure 上 0.5087040618955513)

從圖中可以看出，其實訓練過程收斂的速度很快，大約十個 epoch 左右 loss / accuracy 就已漸趨穩定，同樣也有 overfit 的情形，但平均而言，準確率較第一小題的 CNN-based 要高。

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.

(a) CNN-based **(b) RNN-based**



從兩張圖可以比較出，CNN-based 的 feature 投影到二維以後全都混雜在一起，很難分辨出每個 class 之間的區別，相反地，RNN-based 的 feature 有些 class 很明顯的比較聚集在一起，分群的效果較好，從 validation accuracy 上來看也是 RNN 的準確率較高。

[Problem3]

1. (5%) Describe any extension of your RNN models, training tricks, and post-processing techniques you used for temporal action segmentation.
Preprocess 的部份和第二題相同，但由於這裡的影片長度過長，如果把影片全長的 frame 直接丟進 RNN 會 train 不起來，且有過多無謂資訊，因此我使用 random sample 的方式，把每部影片通過 CNN 所得到的 feature 隨機取出 512 個 frame feature 以及其相對應的 label，並以第二小題 model 的 weight 來 initialize，通過 LSTM 後再將每個 time-step 的 output 接入 fully-connected layer 來做 classification。

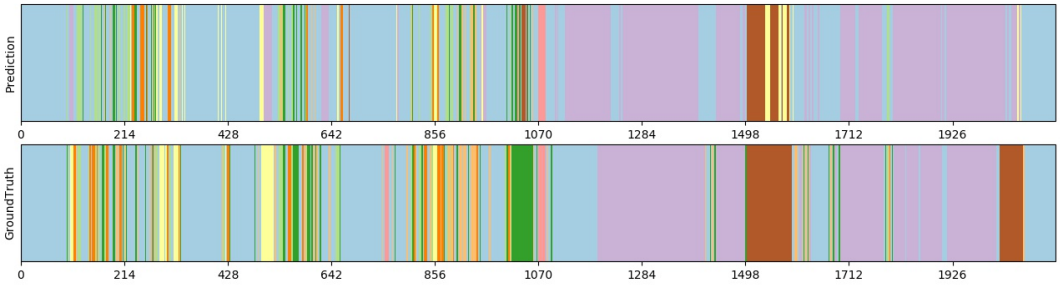
2. (10%) Report validation accuracy and plot the learning curve.

Video name	Accuracy
OP01-R03-BaconAndEggs	0.652803738317757
OP02-R04-ContinentalBreakfast	0.5991471215351812
OP03-R02-TurkeySandwich	0.5460910151691949
OP05-R07-Pizza	0.5290482076637825
OP06-R05-Cheeseburger	0.6154411764705883

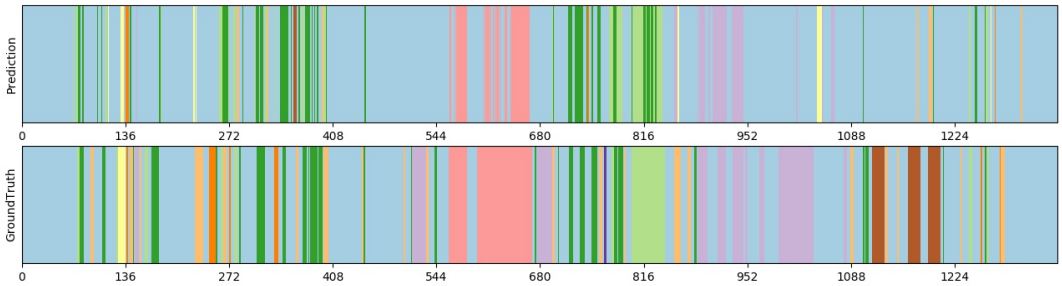
average validation accuracy： 0.5885062518313008

可以看出 training loss 持續下降，但 validation 的部分，無論是 loss 還是 accuracy，大約在 30~40 個 epoch 左右就收斂，accuracy 大約在 0.57~0.58 左右，訓練效果還算可以。

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).
以下選擇兩部影片來做 visualization 的呈現，圖中不同的顏色代表不同的 label，每張圖中，上排為 predict 的結果、下排為 GroundTruth，結果如下：
OP01-R03-BaconAndEggs :



OP06-R05-Cheeseburger :



可以觀察到其實 prediction 有很大一部分是淺灰色，也就是 label 為 0 (Other)，可能是因為 training data 裡 0 就佔了很大的比例，再加上訓練過程是採用

random sample 的方式，可能會遺漏某些重要資訊，因此預測時也較容易傾向於產生 0 的結果。即使如此，我們仍依舊能從圖中觀察到，部分有顏色的 label 預測的還不錯，像是第一張圖粉色及綠色的地方，整體而言，predict 的結果和 ground truth 的分佈情形仍稱的算是一致。