

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: 張承洋 Dep.: 電機三 Student ID: B04901056

[Problem1]

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

我利用 torchvision 的 models 套件下載 VGG16 的 pretrained model 來使用，首先用助教提供的 readShortVideo 函式將影片轉換成一個一個 frame 的圖片，再將圖片丟到 VGG16 model 生成 1000 維的向量，最後得到 影片長度 * 1000 的矩陣作為影片的 feature。針對每一部影片我使用 Sample & concatenate selected 的方法，取出影片 feature 中的第一個、中間的、最後一個 frame 並將它們 concatenate 起來，並使用一個簡單的 fully connected model 將影片從原本的 3000 維向量 (3*1000) 壓成 11 維的向量，最後經過一層 softmax，利用 Adam 作為 optimizer，learning rate = 0.0001，使用 torch.nn.CrossEntropyLoss() 進行訓練。

```
fcf(
  (dnn1): Linear(in_features=3000, out_features=1024, bias=True)
  (dnn2): Linear(in_features=1024, out_features=256, bias=True)
  (dnn3): Linear(in_features=256, out_features=128, bias=True)
  (dnn4): Linear(in_features=128, out_features=64, bias=True)
  (dnn5): Linear(in_features=64, out_features=11, bias=True)
  (softmax): Softmax()
  (dropout): Dropout(p=0.25)
)
```

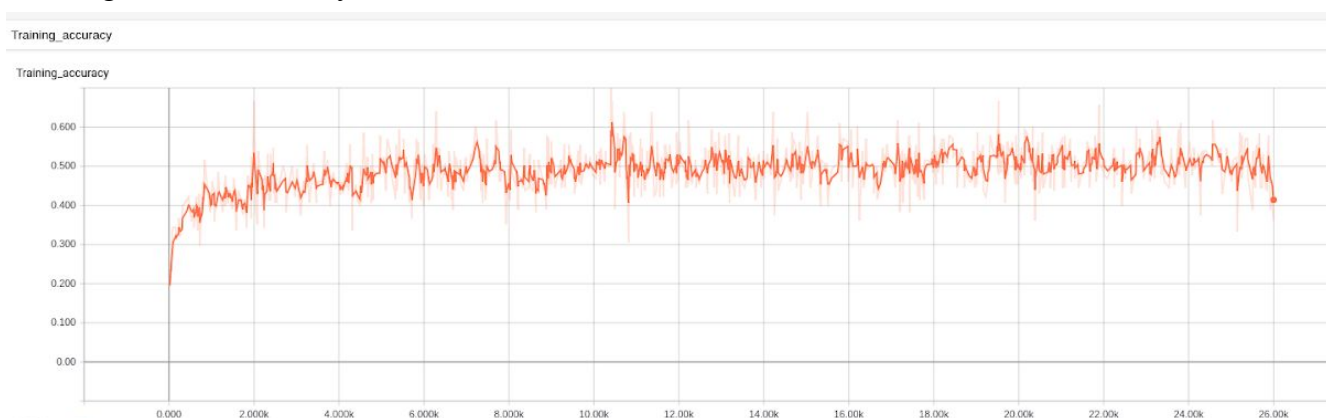
2. (15%) Report your video recognition performance using CNN-based video features and plot the learning curve of your model.

Performance:

accuracy = 0.427

(下面的 learning curve 是 training 時記錄下來的，與上傳的結果有落差)

Learning curve of accuracy :



Learning curve of loss :



[Problem2]

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

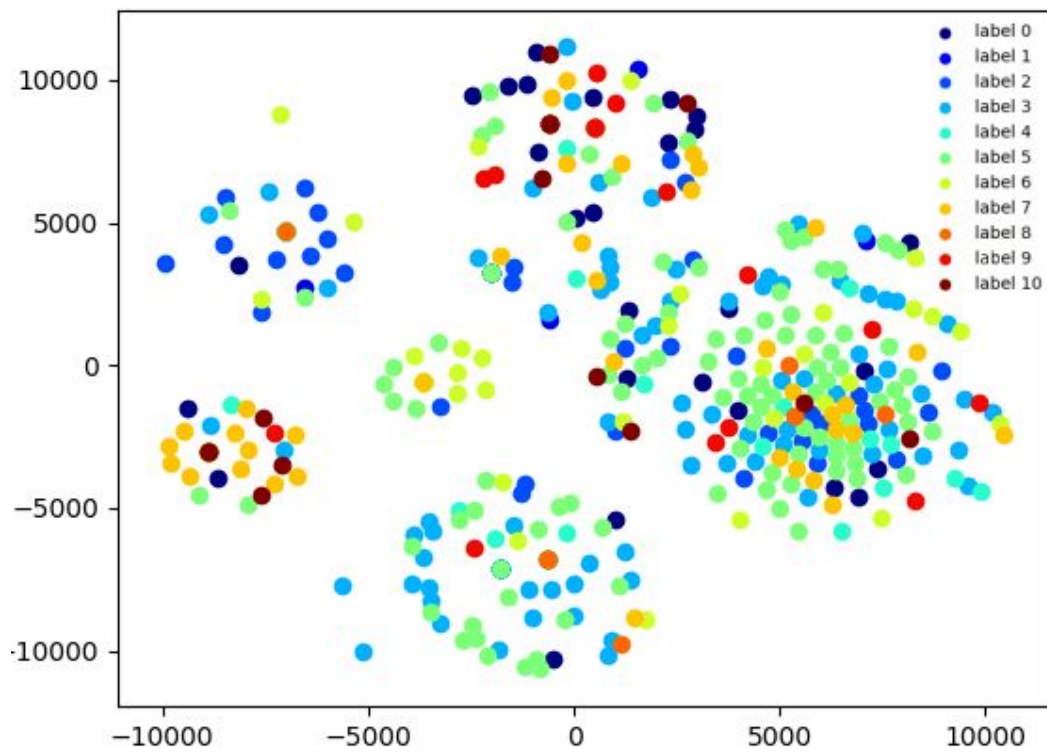
我使用 torchvision 的 models 套件的 VGG16 pretrained model，首先用助教提供的 readShortVideo 函式將影片轉換成一個一個 frame 的圖片，再將圖片丟到 VGG16 model 生成 1000 維的向量，最後得到 影片長度 * 1000 的矩陣作為影片的 feature。

針對每一個 batch 我使用 torch.nn.utils.rnn.pad_sequence，將每一部影片的 frame pad 到一樣長，再利用 torch.nn.utils.rnn.pack_padded_sequence 避免 model 將太多不必要的 padding frame 加入訓練過程，經過兩層 LSTM 之後將 512 維的 hidden layer 取出來經過一層 batch normalization 之後丟入 fully connected layer，將原本的 512 維向量壓成 11 維的向量，最後經過一層 softmax，利用 Adam 作為 optimizer，learning rate = 0.0001，使用 torch.nn.CrossEntropyLoss() 進行訓練。

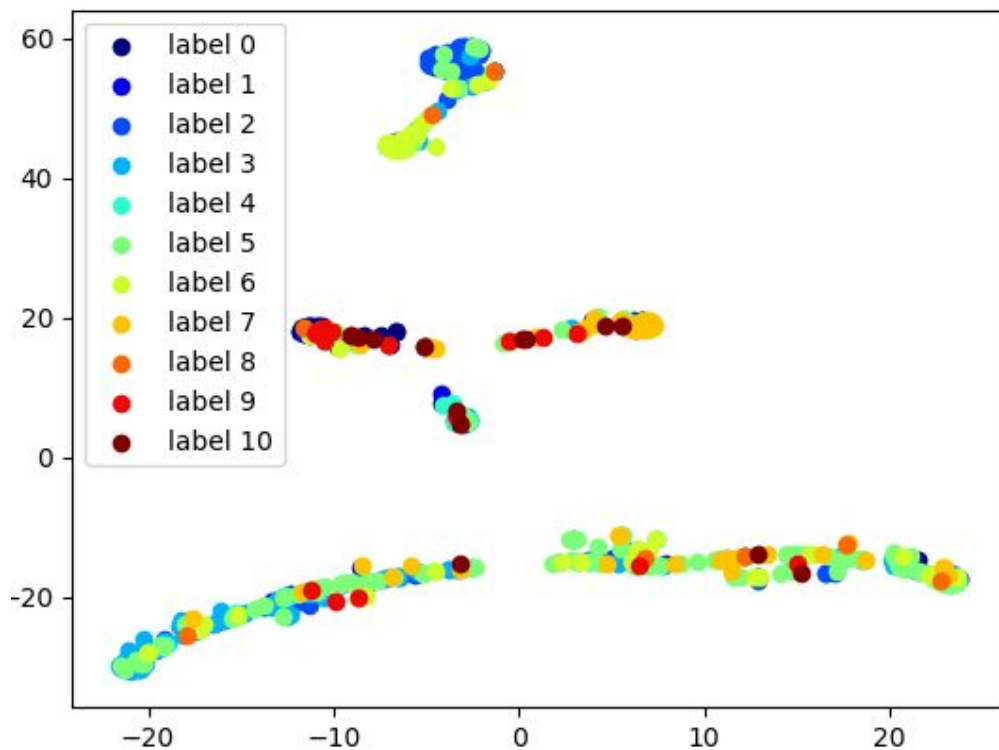
```
(lstm): LSTM(1000, 512, num_layers=2, dropout=0.5)
(bn_0): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
(fc_2): Linear(in_features=512, out_features=11, bias=True)
(softmax): Softmax()
```

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 :



從上面兩張圖可以看出，CNN-based feature 傾向於將資料集中成一團一團，RNN-based feature 則是將資料集中成條狀分佈，而且資料分群程度較為優秀。

我的猜測是 CNN-based feature 看的是一張一張的圖片，RNN-based feature 則會考慮時間前後的圖片，對於同樣一張圖片如果前後情況不同就有可能造成 ground truth label 不同，所以 RNN-based feature 在這點表現得比 CNN-based feature 還要優秀，可以想像說 RNN-based feature 是在 CNN-based feature 上再納入時間這個因素，對於原本模糊不清的分類進行進一步的細分，進而提高分類表現，也提高了影片辨識的準確率。

[Problem3]

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

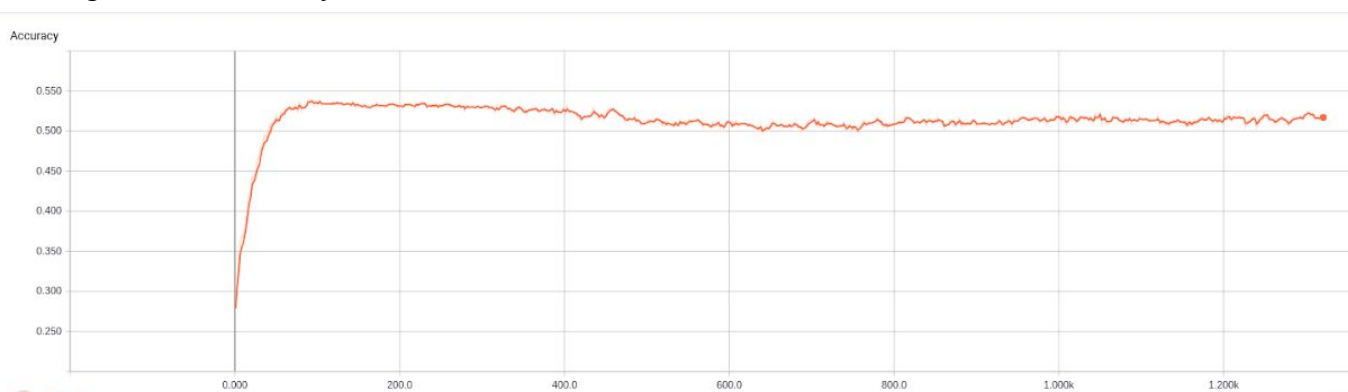
基本上這一題的 RNN model 跟前一題差不多，主要的差別有：

- a. 在這一題當中，我使用了 max_step 作為一部影片的 frame 上限，選擇的方式是 random.sample，而沒被選擇到的 frame 就捨棄掉。
- b. 我將 LSTM 的 output (max_step * batch_size * 512) 直接丟到 FCN 得到 (max_step * batch_size * 11) 的 tensor 再經過一層 softmax，利用 Adam 作為 optimizer，learning rate = 0.0001，使用 torch.nn.CrossEntropyLoss() 進行訓練。
- c. 在更新參數的部份，針對每一個 step 算出 CrossEntropyLoss 並進行相加得到 total_loss，利用 total_loss.backward() 進行參數的更新。

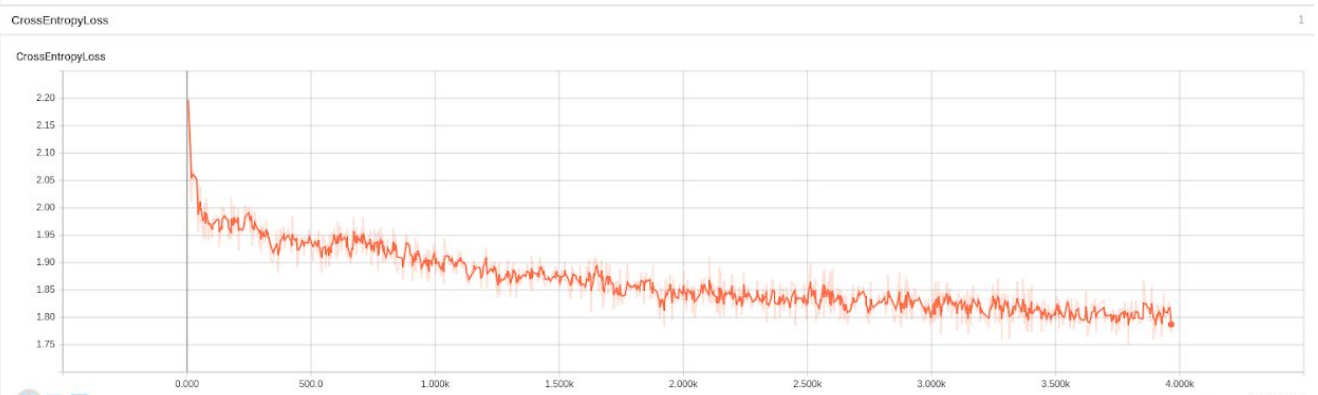
2. (10%) Report validation accuracy and plot the learning curve.

validation accuracy : 0.5617

learning curve of accuracy :

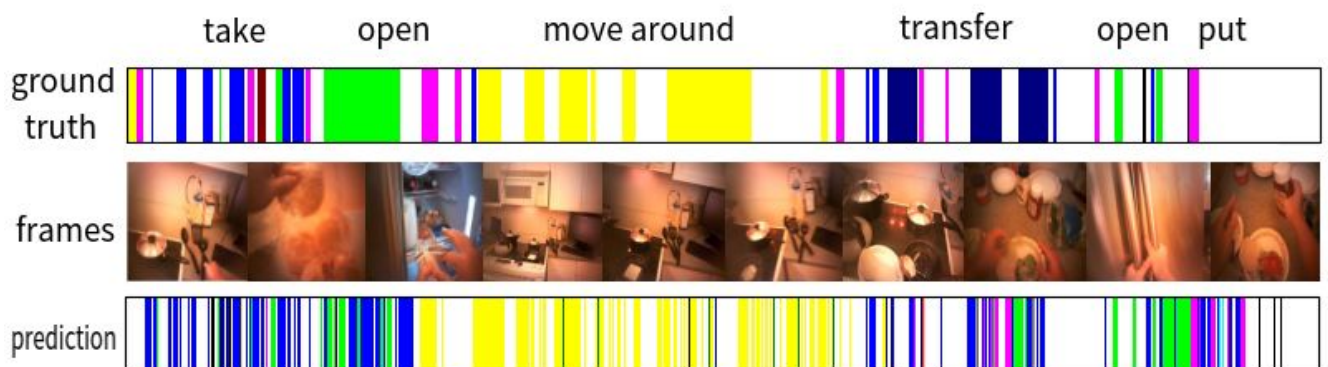


learning curve of loss :








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).

selected video : OP06-R05-Cheeseburger , frame 690 (08281.jpg) ~ end



Label	Action	Color
0	other	
1	inspect/read	
2	open	
3	take	
4	cut	

5	put	
6	close	
7	move around	
8	divide/pull apart	
9	pour	
10	transfer	