Name: 郭冠軒　Dep.: 生機四　Student ID: B03611026

[Problem 1]

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

   我將每段影片降至 2 fps，再利用 ResNet50 pretrained model 將每個 frame 轉為 2048 維的 feature，因此每段影片的 feature 維度會變成 (number of frames, 2048)。接著，我將每段影片的 feature 取平均值，得到形狀為 (2048,) 的新 feature，並將新 feature 餵進數層 dense (fully-connected) layer 的 model 進行訓練。

   我使用的 model 是由三層 dense layer 所組成，並且在其中加上 dropout layer 避免訓練太快達到 overfitting。

   訓練 model 的 implementation details 如下：

   Epochs = 100, Batch size = 64, Optimizer: Adam optimizer (learning rate=10^{-4})

   Model architecture:

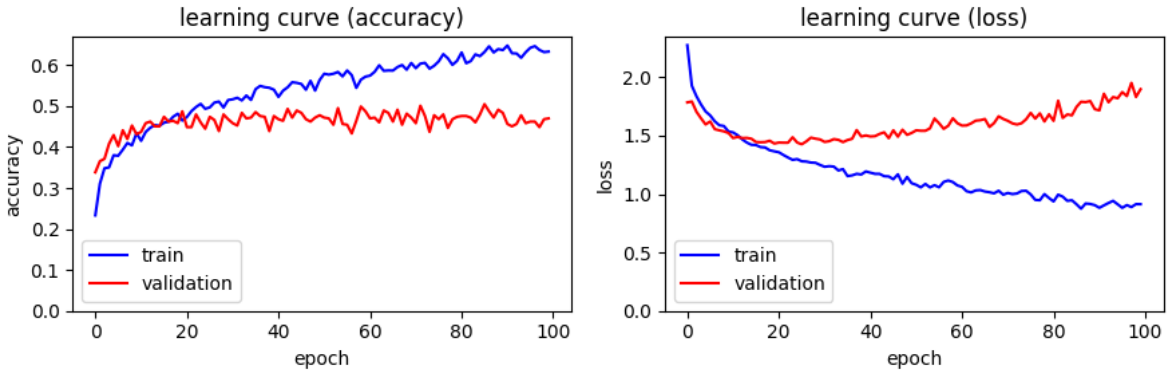
Layers	Details
Dense	Units = 512, Activation function = tanh
Dropout	Dropout rate = 0.5
Dense	Units = 128, Activation function = tanh
Dropout	Dropout rate = 0.5
Dense	Units = 11, Activation function = softmax

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

   Video recognition performance (at the 86th epoch):

Dataset	Training dataset	Validation dataset
accuracy	0.6323	0.5048

   Learning curves:



[Problem 2]

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

我使用數層 LSTM layer 建構 RNN model，並將第一題中取得的 CNN feature 利用 padding 統一 sequence 長度後，餵進 RNN model 中，並且將最後一層 LSTM layer 所吐出的 output 餵進 dense (fully-connected) layer，最後產生預測結果。

訓練 model 的 implementation details 如下：

Epochs = 100, Batch size = 64, Optimizer: Adam optimizer (learning rate=10^{-4})

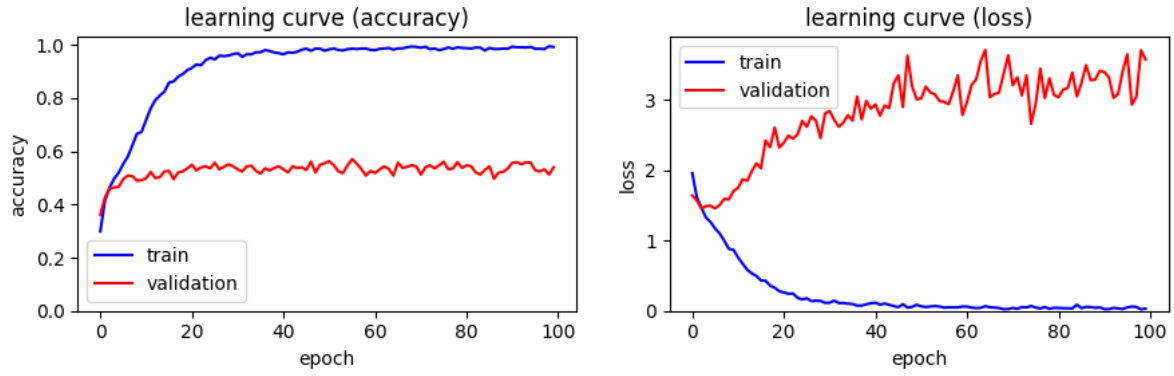
Model architecture:

Layers	Details
LSTM	Cell = 256, Dropout rate = 0.2, Recurrent dropout rate = 0.2, Activation function = tanh
LSTM	Cell = 256, Dropout rate = 0.2, Recurrent dropout rate = 0.2, Activation function = tanh
LSTM	Cell = 256, Dropout rate = 0.2, Recurrent dropout rate = 0.2, Activation function = tanh
Dense	Units = 512, Activation function = tanh
Dropout	Dropout rate = 0.4
Dense	Units = 128, Activation function = tanh
Dropout	Dropout rate = 0.4
Dense	Units = 11, Activation function = softmax

Video recognition performance (at the 56th epoch):

Dataset	Training dataset	Validation dataset
accuracy	0.9858	0.5706

Learning curves:

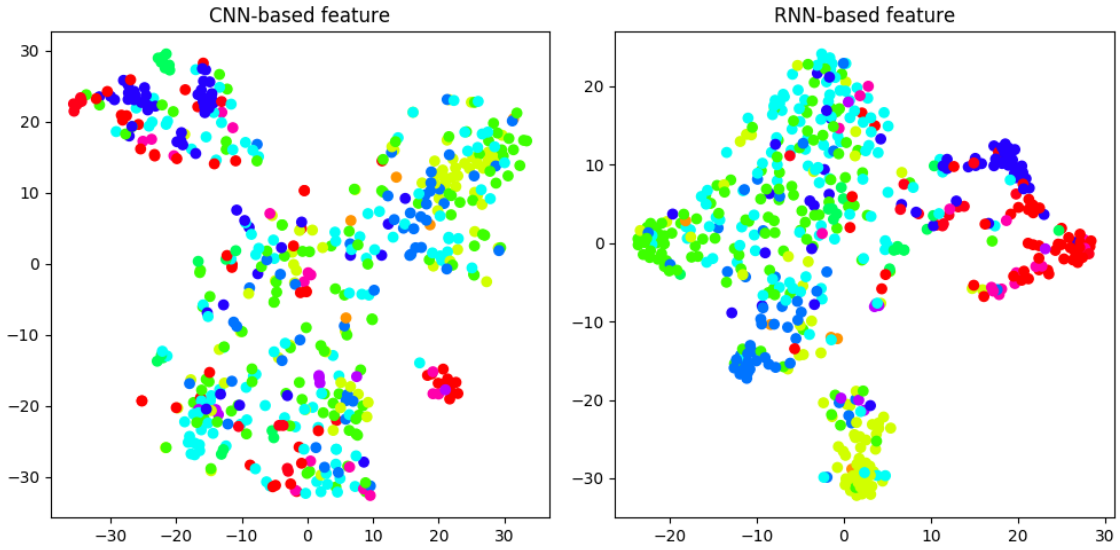


2. (15%) Visualize CNN-based video features and RNN-based video features to 2D space (with t-SNE). You need to generate two separate graphs and color them with respect to different action labels. Do you see any improvement for action recognition? Please explain your observation.

   下圖是我將 validation dataset 中 517 段影片的 CNN-based 及 RNN-based features，經由 t-SNE 降至二維之後，按照 action label 所畫出來的分布圖。

   從圖中可以觀察到，與左側相比，右側 RNN-based feature 的 label 被分得比較好，同一種的 action label 在空間分布中較為靠近。

   就實際訓練結果而言，使用 CNN-based feature 的預測正確率為 0.5048，而使用 RNN-based feature 的預測正確率為 0.5706，RNN 的表現也的確比 CNN 好。



[Problem 3]

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

   由於訓練影片的長度長短不一，我先將每部影片切成數段長度為 250 個 frame 的短影片，以便餵進 model 做訓練。而最終在預測 validation 影片的 label 時，也是將影片切成數段 250 個 frame 的短影片，並將 model 預測的結果進行適當的剪裁與拼接，得到最後的 output label。

   承襲 Problem 2 的 model 架構，我將最後幾層 dense (fully-connected) layer 以 time distributed wrapper 包起來，使每一個 time step 的 LSTM output 都可以各自通過 dense layer，進而達到預測每一個 input frame 的 label 之目的。

   訓練 model 的 implementation details 如下：

   Epochs = 100, Batch size = 16, Optimizer: Adam optimizer (learning rate=10^{-4})

Model architecture:

Layers	Details
LSTM	Cell = 256, Dropout rate = 0.5, Recurrent dropout rate = 0.5, Activation function = tanh
LSTM	Cell = 256, Dropout rate = 0.5, Recurrent dropout rate = 0.5, Activation function = tanh
LSTM	Cell = 256, Dropout rate = 0.5, Recurrent dropout rate = 0.5, Activation function = tanh
Time Distributed (Dense)	Units = 512, Activation function = tanh
Dropout	Dropout rate = 0.5
Time Distributed (Dense)	Units = 128, Activation function = tanh
Dropout	Dropout rate = 0.5
Time Distributed (Dense)	Units = 11, Activation function = softmax

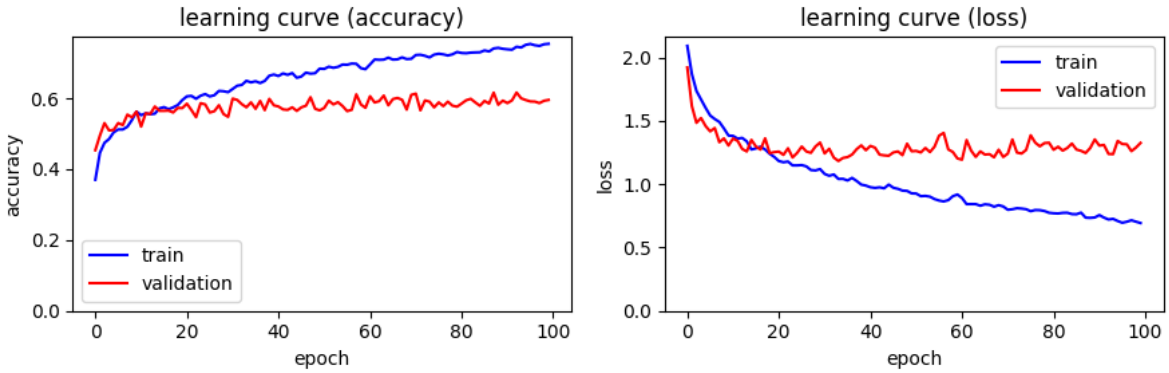
2. (10%) Report validation accuracy and plot the learning curve.

Video recognition performance (at the 93th epoch):

Dataset	Training dataset	*Validation dataset
Accuracy	0.7457	0.6164 (**0.6252**)

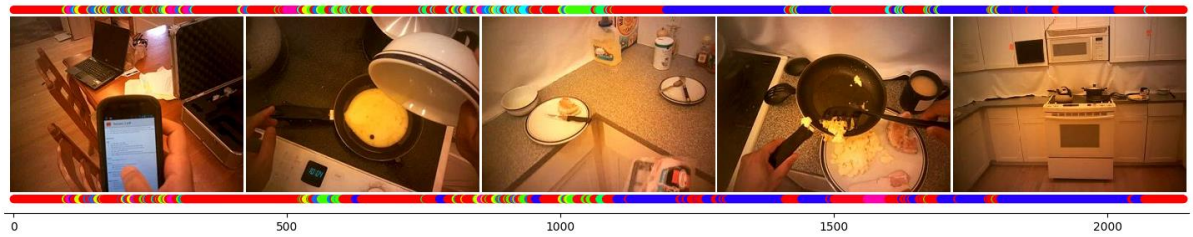
*由於訓練中得到的 validation accuracy 是由已經切成長度為 250 個 frames 的 validation video 去計算，相較於原始的 validation video，經過處理過後的 validation video 有些 frame 會重複出現不同的影片段落，因此原始影片和處理過的影片所計算出來的 validation accuracy 會有所差異。0.6164 為處理過後影片的 validation accuracy，0.6252 為原始影片的 validation accuracy。

Learning curves:



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).

   我將 validation video 中 ”OP01-R03-BaconAndEggs” 的預測 label 結果呈現於下圖，這段影片共有 2140 個 frame，上排為 ground truth labels，下排為 predicted labels。



   從圖中可以觀察到，對於一些 action labels 變化複雜的影片段落，predicted labels 的準確程度相對沒有那麼好，另外，影片中有許多紅色與深藍色標記的較長段落的 label 錯誤預測成彼此。

   我認為將影片統一切成固定長度進行訓練時，如果影片開頭的 action labels 非紅、深藍色，而且是變化複雜的段落，會因為 LSTM 還沒發揮功用，導致影片開頭得到較為不佳的預測結果。