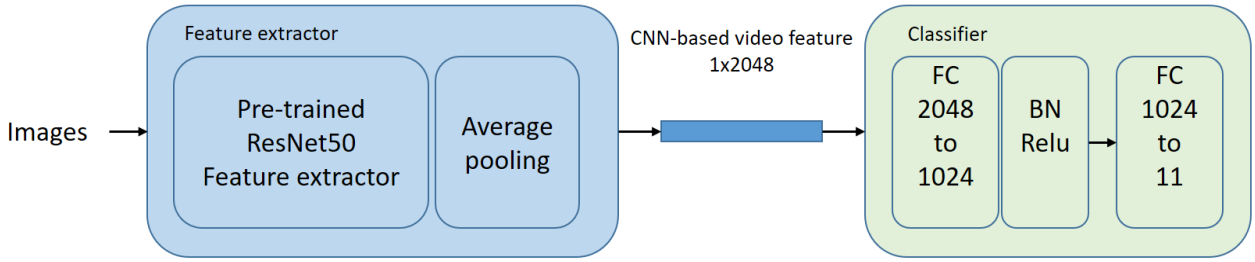
DLCV HW4

鄭承昀 機械碩一 R07522823

Problem 1 Data Preprocessing

1. Describe your strategies of extracting CNN-based video features, training the model and other implementation details (which pretrained model) and plot your learning curve (The loss curve of training set is needed, others are optional). (5%)

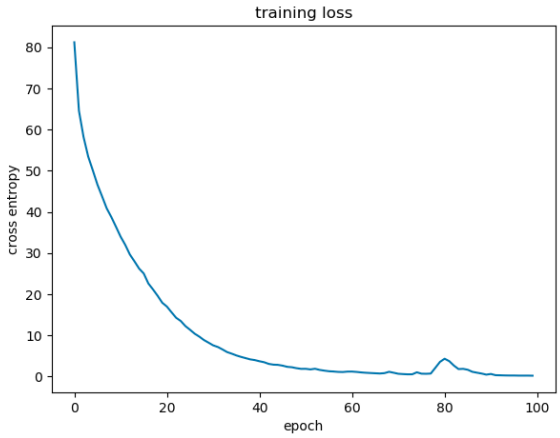


我是使用pre-trained的ResNet50作為我的feature extractor，我每一個frame的feature為1x2048，接著我將所有的feature做平均，成為最後的CNN-based video feature。其中，我down sample成4 fps。

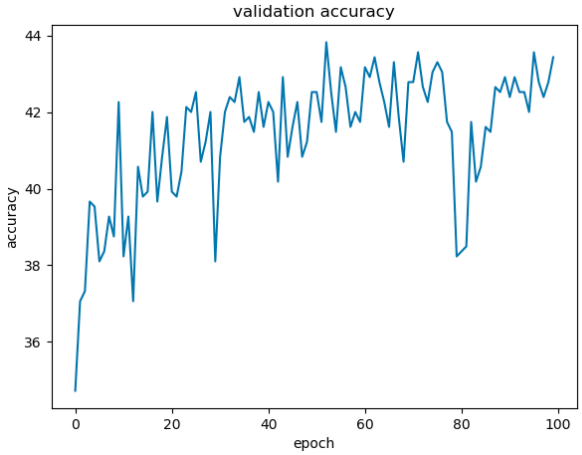
至於Classifier則為簡單的兩層fully-connected layer，輸出即為11類。

Training部份，optimizer使用Adam，learning rate為0.0001，loss function採用cross-entropy。

下圖為loss curve of training。

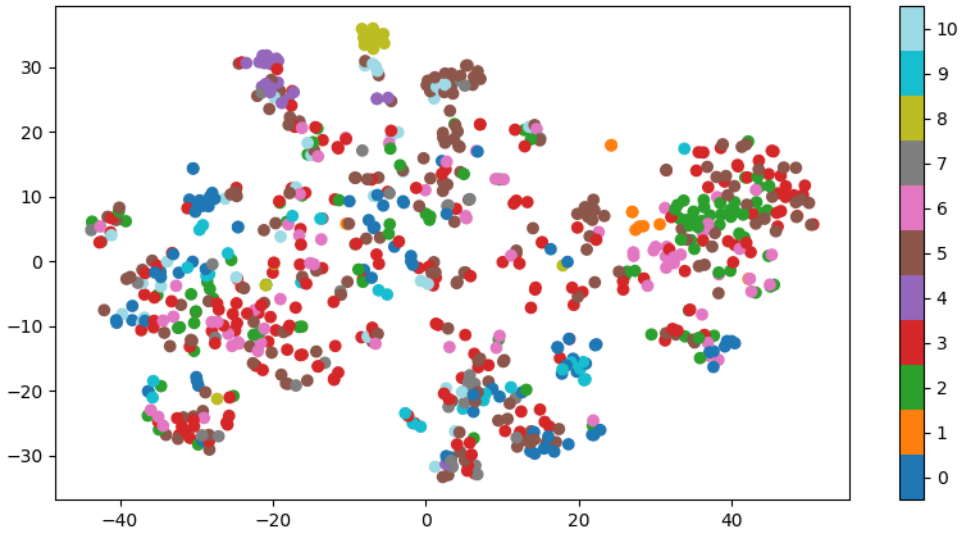


2. Report your video recognition performance (valid) using CNN-based video features. (5%)



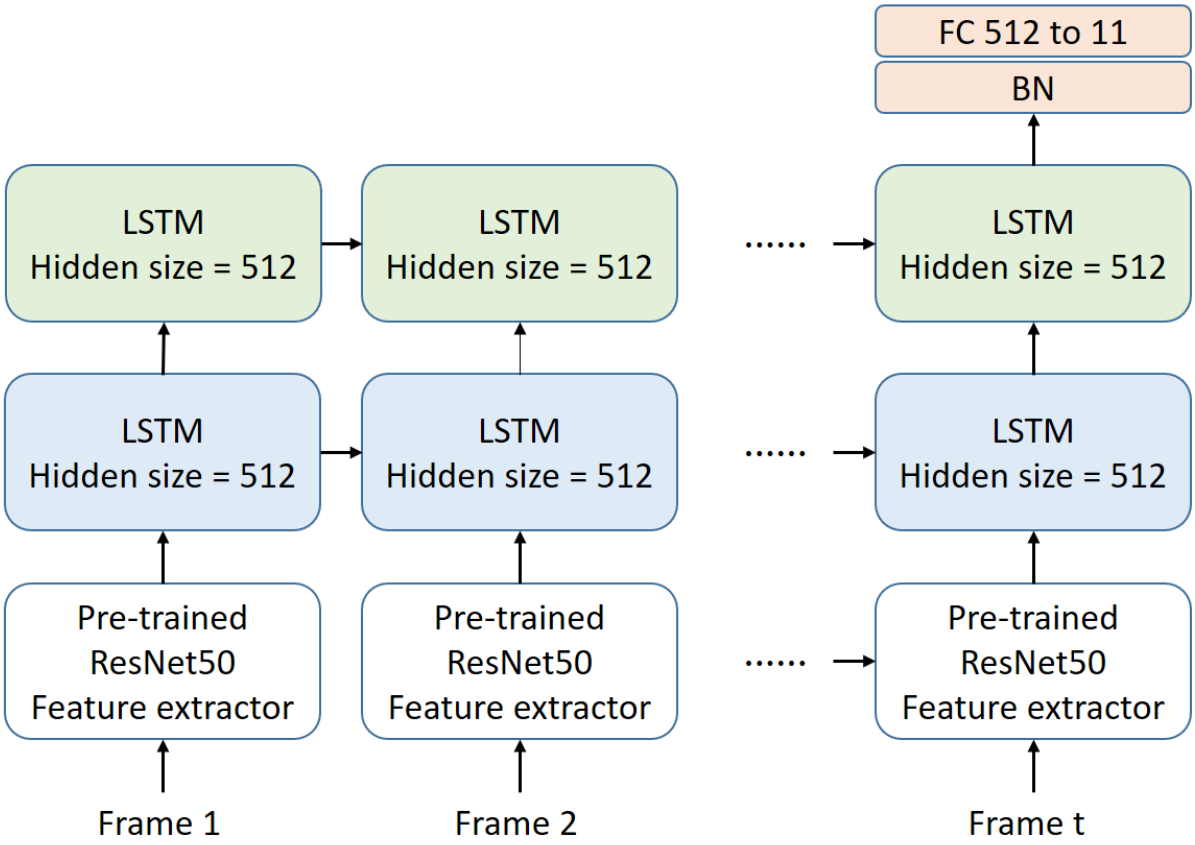
上圖為accuracy curve of validation。從圖上可以看出，在validation set上大概有42%左右的正確率。最後我採用epoch=100，其正確率為43.43%。

3. Visualize CNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels.(10%)



Problem 2 Trimmed Action Recognition

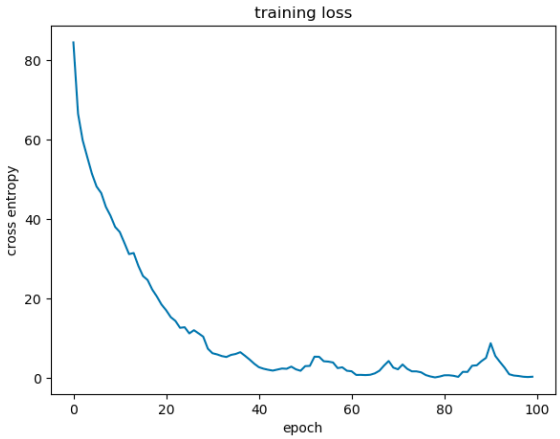
1. Describe your RNN models and implementation details for action recognition and plot the learning curve of your model (The loss curve of training set is needed, others are optional). (5%)



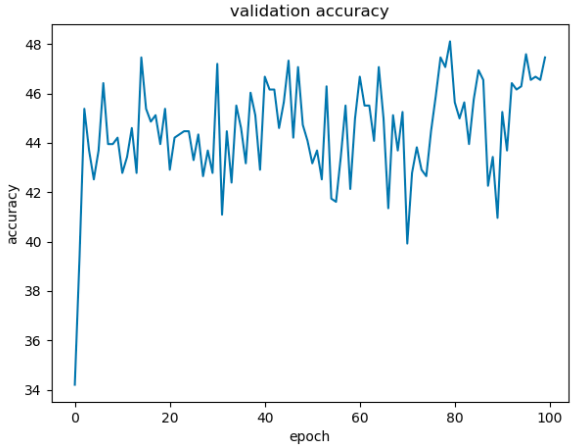
我採用的是兩層的LSTM作為我的RNN model。其中，兩個LSTM的 hidden size為512。最後面接上一個fully-connected layer進行分類。這邊我一樣down sample成4 fps。

Training部份，optimizer使用Adam，learning rate為0.0001，loss function採用cross-entropy。

下圖為loss curve of training。

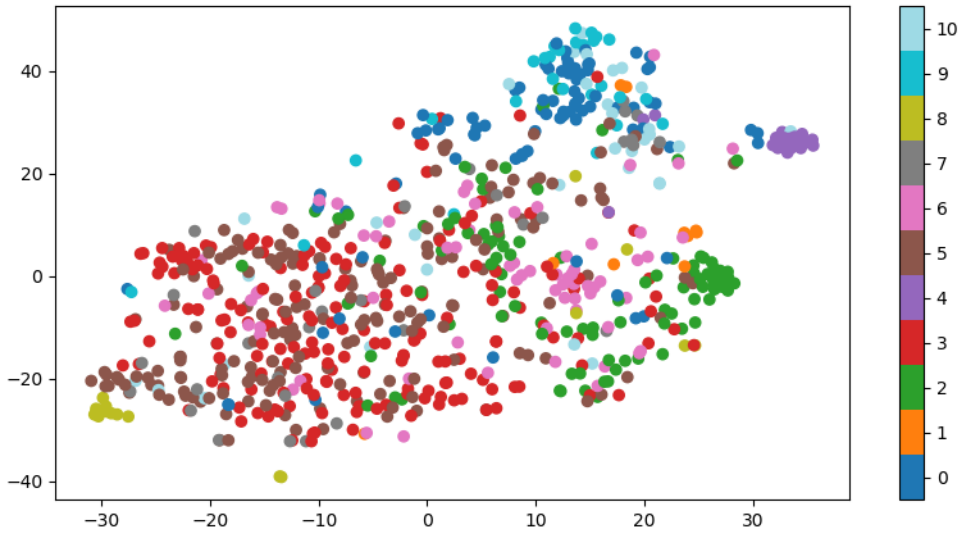


2. Report your video recognition performance (valid) using RNN model.



上圖為accuracy curve of validation。從圖上可以發現，在validation set上的正確率在45%左右。最後我採用epoch=100，其正確率為47.46%。

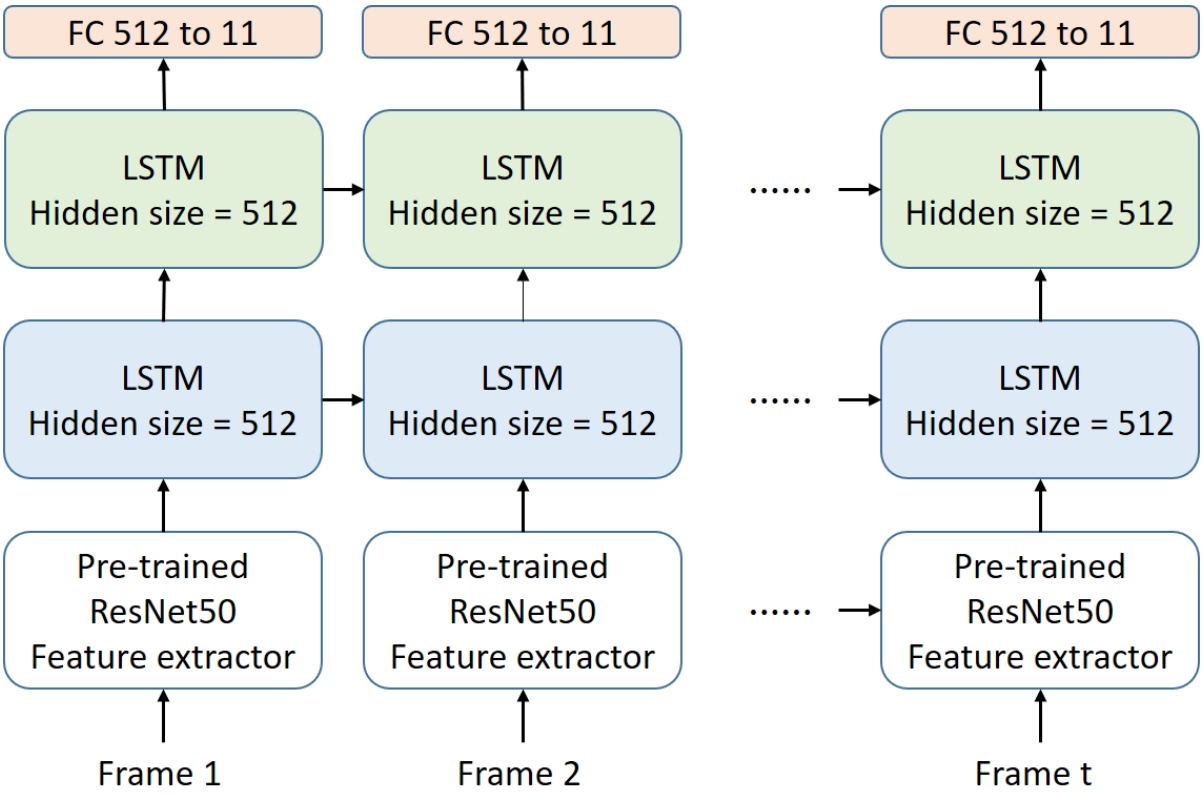
3. Visualize RNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. Do you see any improvement for action recognition compared to CNN-based video features ? Why? Please explain your observation (10%).



從上一題的圖和這張圖我們可以發現其實RNN feature和CNN-based feature在透過tSNE表示在2D平面上時，RNN feature的分類和CNN-based feature的分類差不了多少。這其實從分類的正確率來看也能看出來，RNN feature就比CNN-based feature好一點而已。然後仔細觀察後，我發現這兩種feature比較聚集跟比較分散的類別也差不多。

Problem 3 Temporal Action Segmentation

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

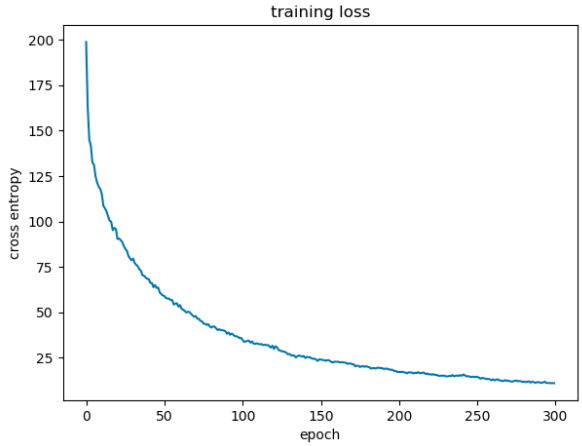


基本上這邊我採用和上一題一模一樣的model，兩層的LSTM，其hidden size皆為512。而在每一個step的輸出都會經過一層fully-connected layer進行分類。

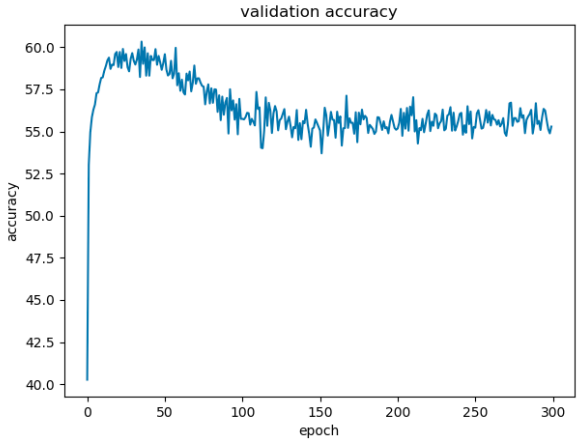
另外，在training data上，我將所有的影片以350frame為一單位將training data切開，其中，兩兩相鄰的部份重疊30個frame。

Training部份，optimizer使用Adam，learning rate為0.0001，loss function採用cross-entropy。

下圖為loss curve of training。

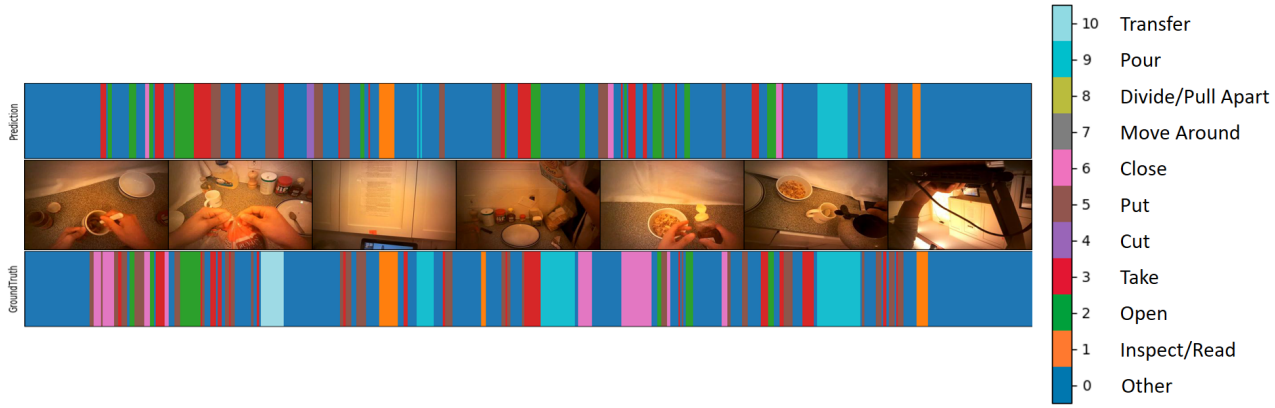


2. Report validation accuracy in your report and make your code reproduce this result.



上圖為accuracy curve of validation。從圖上可以發現，在validation set上的正確率在56%左右。最後我採用epoch=100，其正確率為55.75%。

3. Choose one video from the 7 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 500 continuous frames).



上圖為OP03-R04-ContinentalBreakfast這部影片的visualization（共取539個frame）。從圖上可以看出在Other, Inspect/Read, Open, Take這幾類分辨的正確率特別高。其可能原因為這幾類的training data比較多。而Close, Transfer的辨識正確率則是最低的。

Reference

1. https://github.com/thtang/DLCV2018SPRING/tree/master/hw5