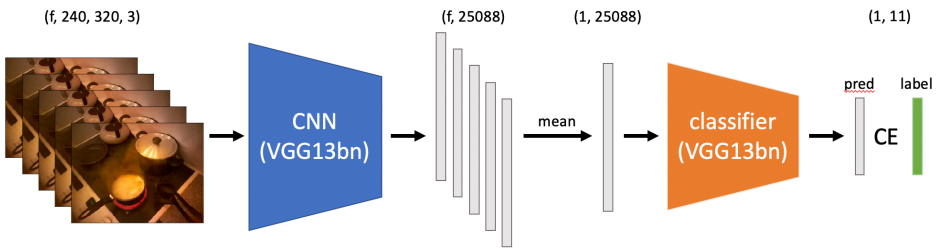
Homework #4

Deep Learning for Computer Vision

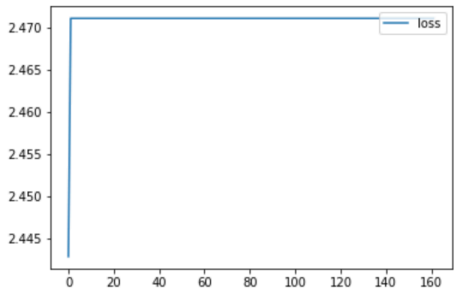
電機所碩二　林益璟　R06921076
Collaborator : No Collaborator

Problem 1 : Data preprocessing (20%)

- 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%)



Problem 1. 使用 VGG13_bn。實際的作法是將每個 frame 都過 CNN，再將全部 CNN 吐出的 features 做平均，再以該 mean feature 過 classifier 以此預測類別。訓練過程 CNN 部分是 fix 的，只訓練 classifier。



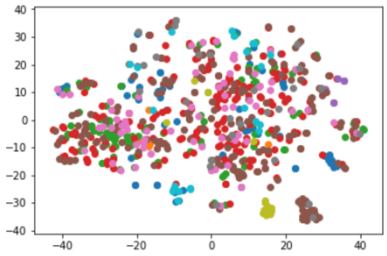
上圖是 Problem 1. 在訓練階段的 loss 與 accuracy，實際上在 epoch=160 內除了第一個 epoch 有改善 loss 但之後都不在進步，有嘗試改 classifier 或使用其他 backend model 但都沒效，猜想用 concat 或 weighted sum 來取代 mean 或許有機會訓練起來。

- Report your video recognition performance (valid) using CNN-based video features and make your code reproduce this result. (5%)

  8.84%（實際上因為訓練環境的顯卡僅有 8G mem.，部分資料會無法讀入，此情況佔約 5%，我直接預測為 1，故在有不同 mem 的環境下應會稍有浮動。）

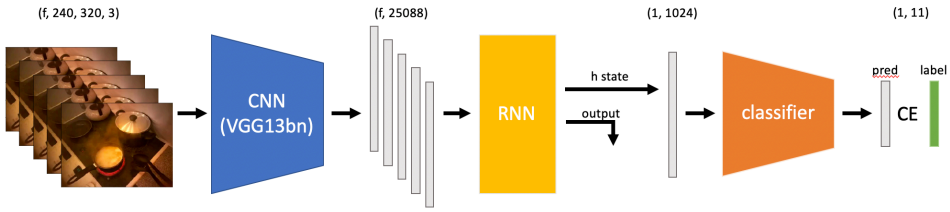
- Visualize CNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. (10%)

  我取 CNN 出來的 features 並 mean 後，分別依序丟入 PCA 及 TSNE (n_component 分別為(300, 2))得到右圖，因為我 Problem 1 沒有訓練起來，故 TSNE 的結果也沒有同 class 群聚的現象。



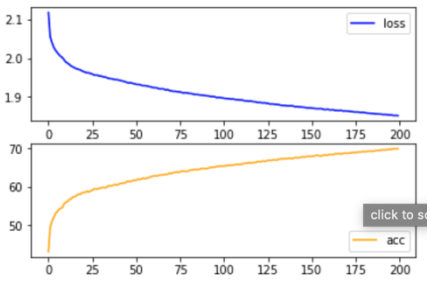
Problem 2 : Trimmed action recognition (40%)

- 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%)



在 Problem2 的 CNN part 也使用了 VGG13_bn，不同的是為了不要像前面一樣因大量空間被模型佔用造成，我在 init RNN 時 hidden 給 1024，雖然可能會使 capacity 下降但在這個 task 應沒有太多影響(穩定上升)，過了 RNN 後我是直接給將 output 捨棄並將 h state 往 classifier 送。

右圖是 Problem 2 的訓練結果，loss and acc 屬於 training set，從中可以看到與 Problem 1 相比，使用了 RNN 後能在訓練時得到穩定的提升。

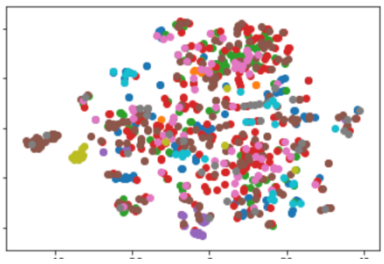


- Your model should pass the baseline (valid: 0.45 / test: 0.43) validation set (10%) / test set (15%, only TAs have the test set).

  在 valid 上得到的分數為 15.6%的正確率。

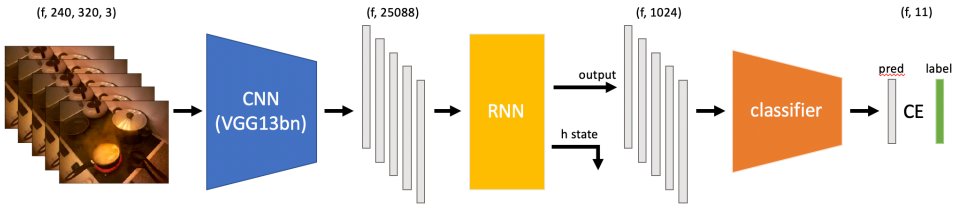
- Visualize RNN-based video features to 2D space (with tSNE) in your report. You need to color them with respect to different action labels. Do you see any improvement for action recognition compared to CNN-based video features ? Why? Please explain your observation (10%).

  從圖片中可能比較難看出來改進的地方，如果真要說，全域來看 class 仍然是發散的，但從局部區域來看，某些 class 有群聚的現象。



Problem 3 : Temporal action segmentation (40%)

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



在 Problem 3 我使用與前者相似的結構，但在 classifier input 的地方選擇使用 output (為了符合 frame 數)而非 h state。在訓練時也算穩定上升。值得一提的是，因為礙於運算資源，我在 batch 上無法取太大，需要自動將 h_state 取出再送回 next batch frame 中繼續運算接續的 frame，而不是在一次 model forward 下完成整個 video 的 RNN 運算。

- Report validation accuracy in your report and make your code reproduce this result. (20%)

  於 validation set 的正確率為 13.11%