HW5

R06921062 蘇楷鈞

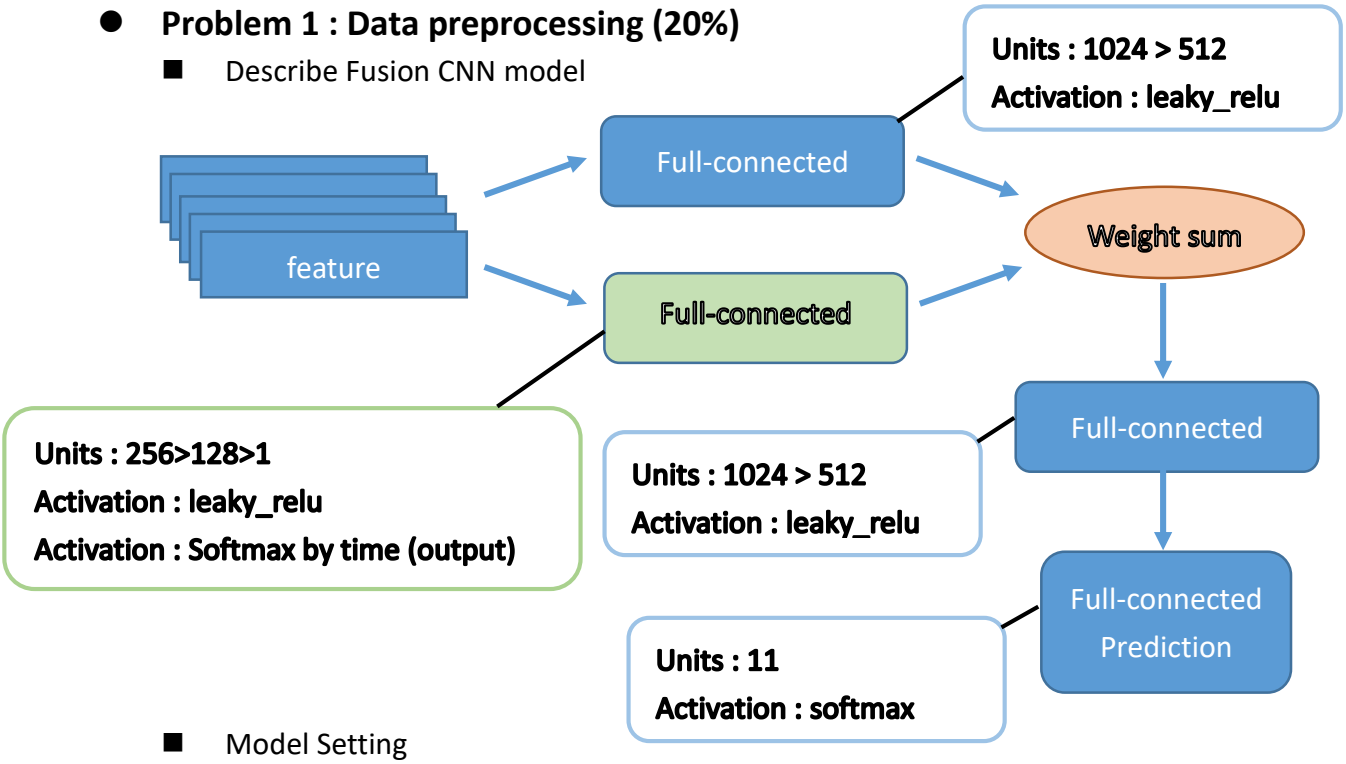
- **Default setting**
  - Feature extractor

    Resnet50 , training on Imagenet. Feature shape is [1,1,2048].

    ReadShortVideo(downsample_factor=12,rescale_factor=1)

    Slice Video until length is shorter than 50.

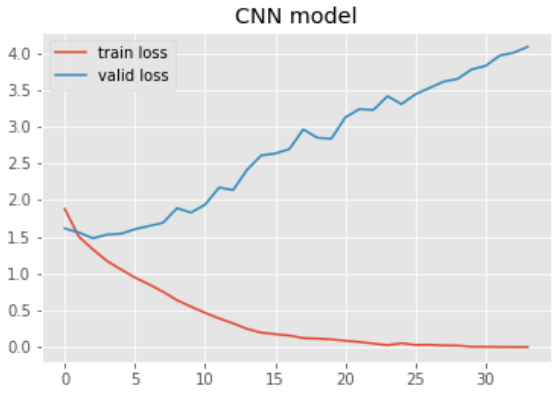
- **Problem 1 : Data preprocessing (20%)**
  - Describe Fusion CNN model

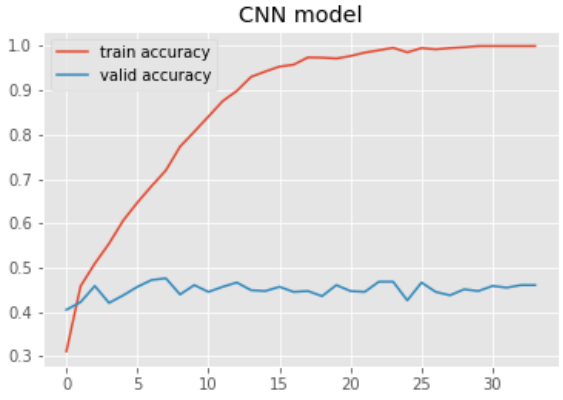


  - Model Setting

Objective function	Optimizer	Batch size	Earlystopping	Max time stamp
Categorical Cross-Entropy	Adam lr=1e-4	32	25	50

  - Learning curve



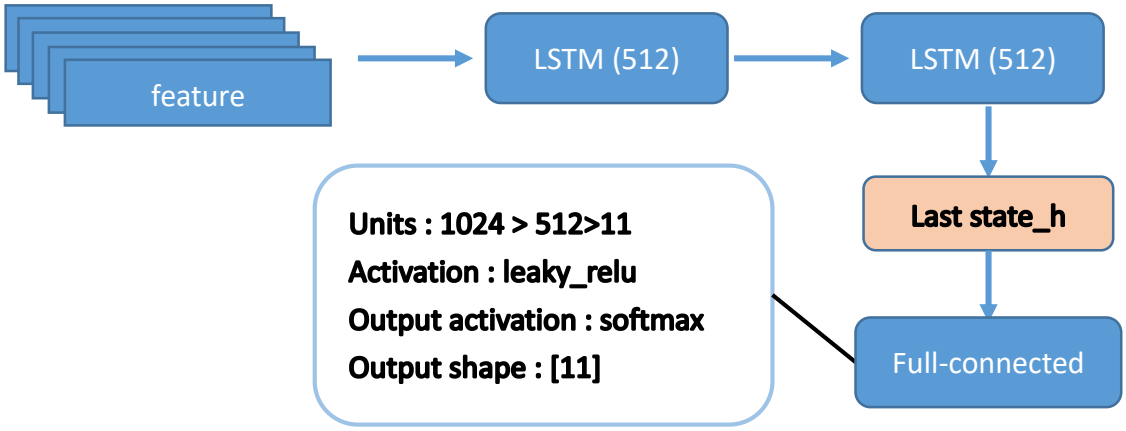


- Comment

  最好的準確率為 0.47388，在考量 CNN 缺少像 RNN 一樣讀取時間資訊的架構，因此想說已個動作的資訊融合來找出屬於這動作的 feature，因此設計兩個 fully-connected 的模型來分別輸出動作資訊以及其權重，最後做 weight-sum 取出資訊。而這樣的架構也比直接取平均來的好。

Problem 2 : Trimmed action recognition (55%)

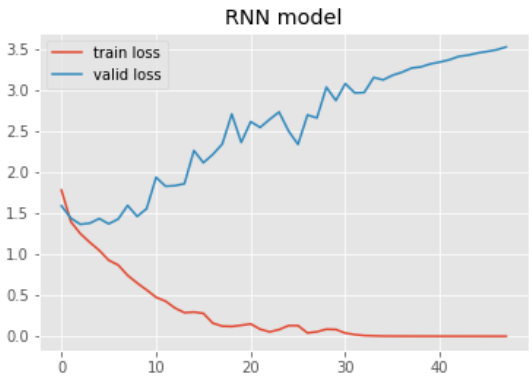
- Describe RNN models

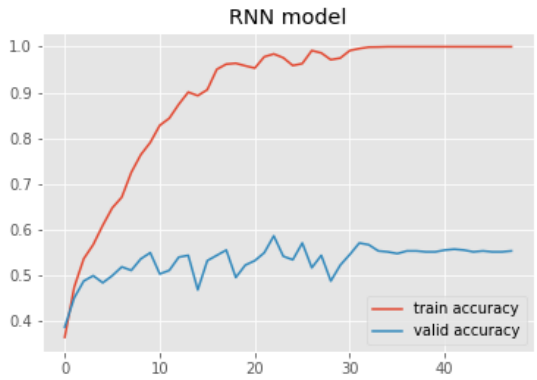


- Model setting

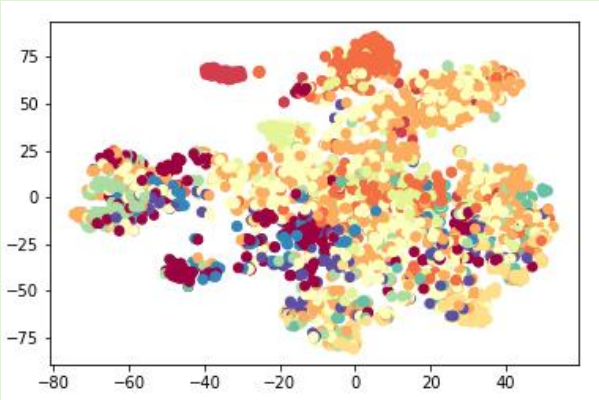
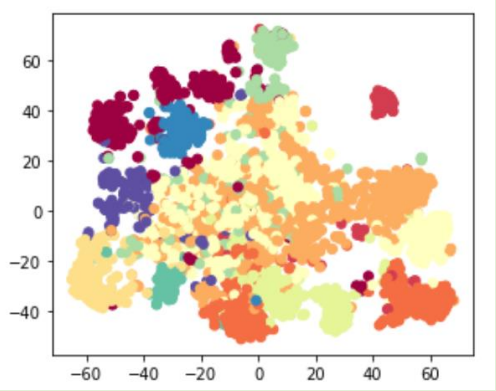
Objective function	Optimizer	Batch size	Earlystopping	Max time stamp
Categorical Cross-Entropy	Adam lr=1e-4	32	25	50

- Learning curve





■ Visualize

Fusion CNN	RNN
	

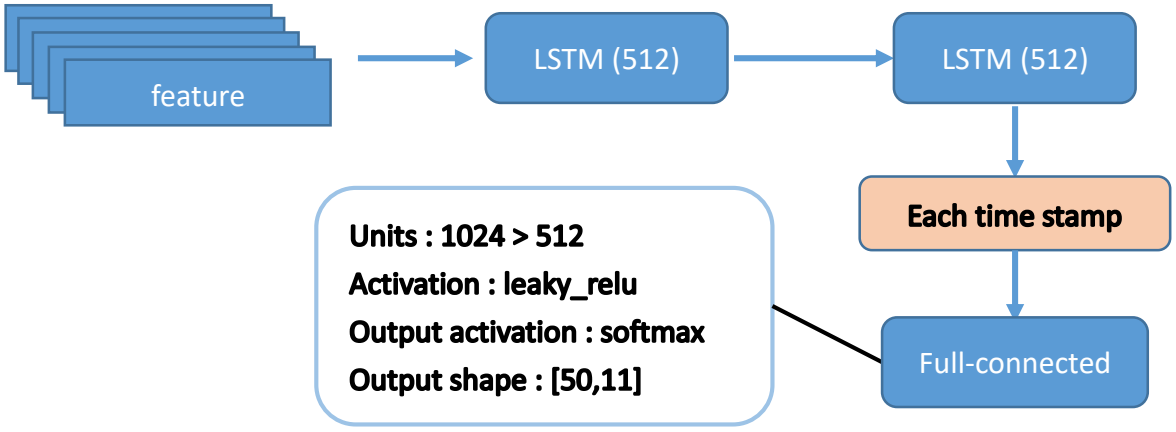
■ Comment

RNN 模型和 CNN 的模型，除了很明顯地在 accuracy 上前者勝過後者，從模型取出的 feature(RNN 為 last state_h，CNN 為 weight-sum)來看，前者較能把各類別在空間中區隔開來，後者類別常混在一塊。以模型設計的角度來看，CNN 不包含時間資訊，而 RNN 則有時間上順序的資訊，以 take 和 put 這兩動作來說，就必須有時間上順序才知道是 take 還是 put。

	CNN	RNN
Accuracy	0.4738878143133462	0.586073500967118

● Problem 3 : Temporal action segmentation (25%)

■ Describe Seq2Seq model



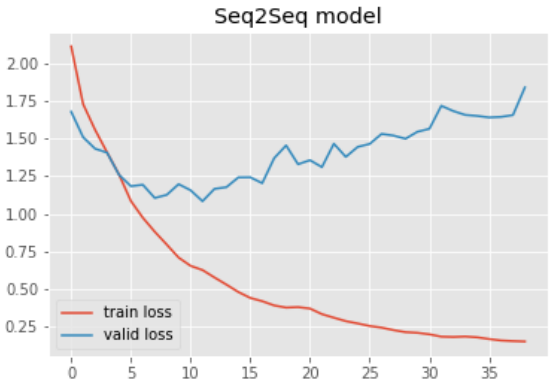
■ Model setting

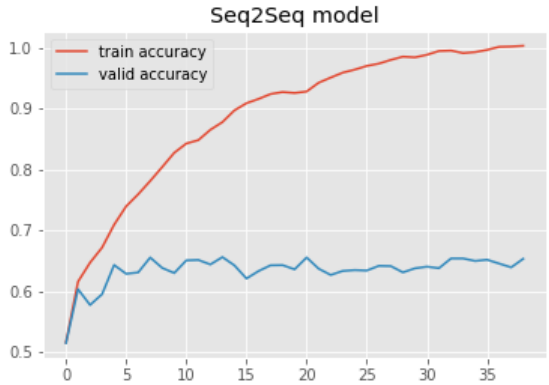
Objective function	Optimizer	Batch size	Earlystopping	Max time stamp
Categorical Cross-Entropy	Adam lr=1e-4	32	25	192

- Preprocessing

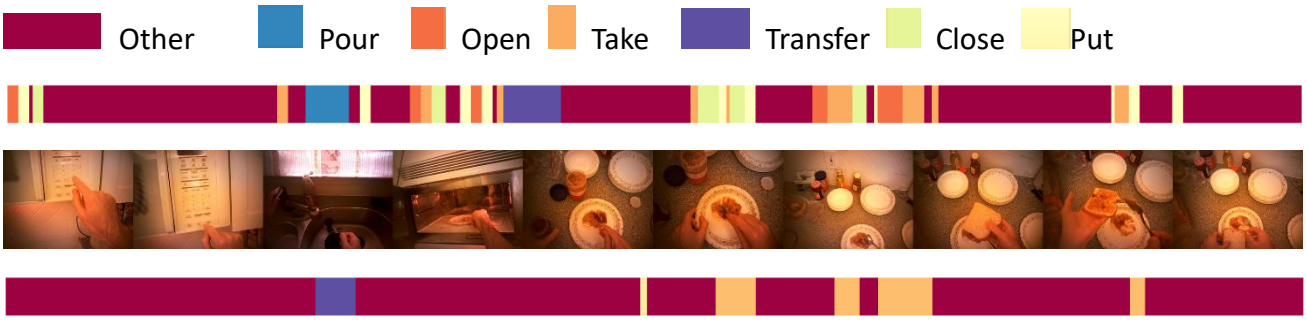
  把影片切成最小長度 160，對大 192，視影片總長而定，另外訓練資料集再亂數挑取區間，訓練資料共有 317 筆。

- Learning curve





- Temporal action segmentation

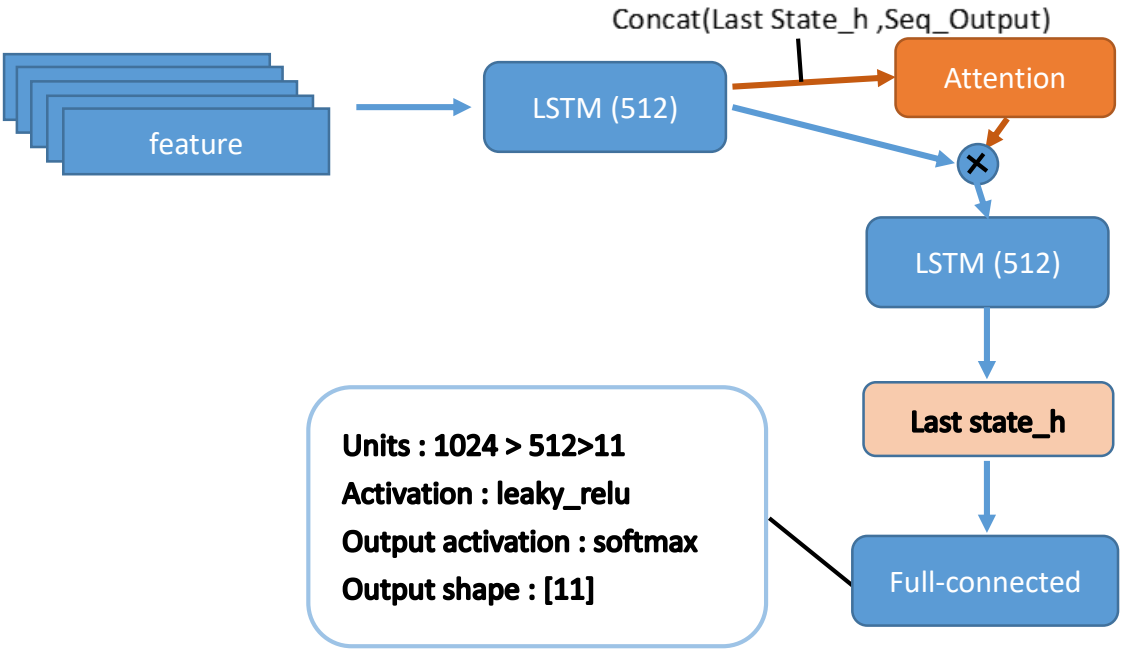


影片擷取自 ContinentalBreakfast 第 199 張到第 459 張的圖片，上方是 ground truth，中間是以間格 35 從第十張開始抽取，下方是預測的標籤，可以看出對於 other 都很好辨認，take 也相較其他標籤，辨認出來的機會較高，可能是因為在資料集中，other 的 label 佔了多數。

Bonus

- Trimmed action recognition

模型和第二題大致相同，但在過完第一層 LSTM 後，將最後的 state_h 和第一層 LSTM 的輸出做合併，經過 attention network 後產生 50x1 的向量，而不實作對每個 time stamp 的原因是，對每筆影片都只對應一個類別，因此以 attention 的架構去找出該注意的時間點。

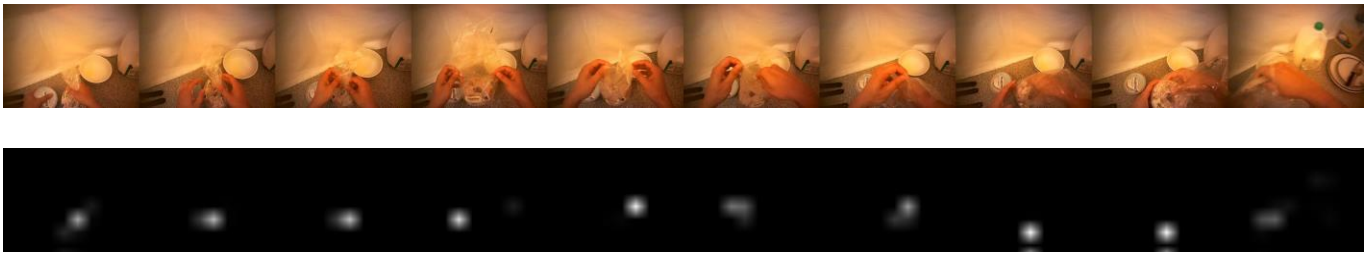


下圖是隨機取五筆 valid data 做實驗，發現大部分都給第一層 LSTM 最後的輸出最大的權重，可能是一層 LSTM 就足以產生出足夠的資訊，因此 attention network 幾乎都給後面的 time stamp 較大的權重。其 label 分別為[5:Put , 7:Move Around , 3:Take , 3:Take , 10:Transfer]。

■ Temporal action segmentation

將 Resnet50 最後一層 Conv2D 取出 feature map(8x10x2048)當 Seq2Seq 模型的 Input，將第一層 LSTM 每個時間點的 state_h(512)複製成 80x512，和 flatten feature map(80x2048)合併，經過 attention network(fully-connected 512>128>1)後，產生出 8x10x1 的 score 和 feature map 相乘(註：score 總合為 1)，最後經過 global average 產生 2048 維的 feature。雖然準確率只有 0.54025，並沒有比第三題好，可能是加輸入的 feature 經過 attention 後分布可能更廣，在直接挪弄第二題的模型可能無法完整學習。

下圖代表著連續從 valid dataset 取出的 10 張相片所計算出的 score(每次影片的長度為 1 60 至 192 不等)，例如 open 就關注在開口附近，還有當要準備 take 時，會擴大注意範圍到手的動作。



Put Put Put Open Open Open Take Take Take Take