ADL x MLDS assignment report

學號：B03901030 系級： 電機四 姓名：蕭晨豪

1. Model Description
   - RNN
     我使用的 RNN 為 Bidirectional 的 LSTM cell，hidden dimension 為 512，layer 為 2，layer 中的 dropout 為 0.4，使用的 feature 是 mfcc。我的 model 直接一次性的將所有 frame 吃進來並且輸出一個 predicted sequence。
   - CNN+RNN
     我使用的 CNN+RNN model 分成兩部分，CNN 的部分使用 fbank，先將所有的 frame 用一個寬度為 5 的 window 刷過，每一個 window 使用 kernel 為 5x5 的 10 個 CNN filter 抽出 10 個 feature，將這十個 feature 與 mfcc concatenate 起來之後送入 unidirectional LSTM cell，LSTM 的 hidden dimension 為 256，layer 為 2，layer 中的 dropout 為 0.5，最後輸出一個 predicted sequence。
   - Details
     我使用的套件是 pytorch，其中 output 和 criterion 的組合是 Log Softmax + NLLLoss(negative log likelihood)，使用的 optimizer 是 lr 為 0.001 的 Adam，我切了 10%的資料當 validation data，並選擇 validation edit distance 最好的 model 上傳。

2. How to improve your performance
   - Methods
     - 在 trimming 的部分先將<sil>去除，之後連續三個以上一樣的 phone 才被我加到 final output 中
     - 將 feature 做 normalization，減去 mean 再除以 std
   - Why
     - 更改 trimming 應該是讓我過 baseline 的最大功臣，主要是觀察到 training data 的 label 中每個 label 至少都重複三次以上(因為人說話的 phone 不可能變化太快而且 frame 有重疊)，所以這樣可以非常有效的清除不小心認錯的 phone
     - 主要是我一開始 train 時 loss 不會穩定的收斂，我認為加 normalization 可以讓 model 更輕鬆的學習(scale 比較小)，我較好的兩個結果也都有經過 normalization

3. Experimental Results

- Compare RNN and CNN

  我比較了 RNN 和 CNN 做在不同 feature 上的 performance，當兩者皆使用 mfcc 時純用 RNN 效果(13.87)會比 CNN+RNN(15.16)好，而使用 fbank 時 CNN+RNN(14.67)的效果會比純 RNN(15.16)好。(紅字為最好的 validation edit distance)與同學討論和查詢資料後，原因似乎是因為 mfcc 的產生過程中 discrete cosine transform 的係數會有一些被捨棄掉，以致於喪失掉某些 locality 的資訊，不適合 CNN 做辨認。

  根據這個實驗我最後的 CNN+RNN model 先嘗試從 fbank 中抽取 feature 再加入 mfcc，不過就結果來看並沒有比直接用 bidirectional LSTM 做 mfcc 好，或許可以將 CNN 上的 RNN 改成 bidirectional 試試看。

- Compare models
  - 原本使用的 RNN model 是一層的 unidirectional LSTM，在 training 時很快就 overfit 了，validation loss 只能下降到 0.98 左右，edit distance 也還在 20 出頭。後來認為參數量不太夠便嘗試了兩層的 LSTM，情況的確有所改善，validation edit distance 降到了 16 左右。
  - 另一個嘗試的方向是 bidirectional，因為這次的 task 並沒有需要依時序進行，因此可以使用 bidirectional 的模型。Bidirectional 從兩側獲取資訊可以解決一部分的 vanishing gradient 問題，在預測較長的 sequence 時也應有較好的成果。最後最好的結果的確是由 bidirectional 兩層的 LSTM 達到的。