ADL 2017 HW1 Report

B04902004 王佑安

1. Model Description

   RNN:

   Input ->
   BiLTSM(250)->
   BiLTSM(250)->
   BiLTSM(250)->
   BiLTSM(250)->
   Linear->
   Output

   Loss function: Cross Entropy
   Optimizer: RMSProp lr=1e-3 alpha=0.87

   RNN+CNN:

   Input ->
   Conv1d(64, 7) ->
   Conv1d(64, 7) ->
   BiGRU(64) ->
   BiGRU(64) ->
   Linear ->
   Output

   Loss function: Cross Entropy
   Optimizer: Adam lr=1e-2

2. How to improve your performance

(1) 將 speaker 的性別當作 feature

data 是由真實的聲音轉為向量，因此男女的生頻率差異應該也會影響 training，所以我把 speaker id 表示性別的第一個字母也當作一個 feature 處理。

(2) Bidirectional RNN：

根據常理判斷，一個 phone 除了跟前面的 phones 會有關聯，後面接著的也應該要有，因此使用 bidirectional 讓 predict 時也能參考到後面的 phones。比較後 bidirectional 表現確實比只把參數增加兩倍來得好。

(3) 移除連續長度短的 phones：

觀察 training data 後可以發現 phones 幾乎都是很多個相同的連續，而 predict 出來的常有一兩個單獨夾雜在一整串相同的 phones 中間，可以判斷這可能是 model predict 時候產生的雜訊而不是正確答案，而且依樣可以跟據常理判斷講話是確實很少有單個 phone 短暫的夾在其他 phones 中間。嘗試把連續長度小於 n 的 phones 直接從 output 中移除，發現在 n=3 時表現提升最多。

3. Experimental results and settings

(1) RNN

Layers	2	2	4	4
Hidden Size	64	250	64	250
Accuracy	0.752	0.757	0.763	0.782
Edit Distance	10.21	9.81	9.66	8.77

嘗試了不同層數跟 hidden size，層數跟 hidden size 都是越高表現越好，但超過 4 層或 hidden size 再提高都會讓 training 的時間慢很多，或是出現無法 fit 的情況。

另外也嘗試了 LSTM 跟 GRU 的比較，兩者表現其實差不多沒列入表格，但 LSTM 的 hidden sate 比以 GRU 多，所以 training 花的時間比較久。

(2) RNN+CNN

Model	CNN*4+RNN*2	CNN2+RNN*2
Accuracy	0.772	0.784
Edit Distance	9.26	8.55

加上 CNN 的表現其實跟只有 RNN 差不多，沒有明顯的進步。而我在 train 的時候 RNN 層數只要超過 2 就會沒辦法 fit，因此沒有嘗試加上更多 RNN，而 CNN 層數增加似乎也沒有提昇表現。

實驗後我認為兩種 model 表現差不多，我認為可能是因為我使用了 Bidirectional RNN 的關係。CNN 可以將周圍的 phones 卷積起來，而原本的 RNN 只能看到前面的 phones，有了 Bidirectional 讓 RNN 也能同時看到前後的 phones，達到類似 CNN 的效果，因此表現差不多。