r04325008

資料處理好之後，有3696個句子，最長的trainning句子有777個frame，所以我先把所有的句子padding到777，之後再使用mask讓cost function不計算padding的cost。Target先處理成39個類別之後才放進模型裡。

Model Description

1. RNN

我使用keras的sequential model。第一層先使用masking，讓之後的cost function可以忽略輸入全是0的frame。接著使一層的Bidirectional LSTM。Bidirectional LSTM後面接一層的LSTM。在這兩個RNN後都有接batch normalization，目的是為了要讓最佳化的更好。最後在RNN後面接上一些fully connected layers，並使用relu作為activation。最後一層的activation使用softmax，讓模型可以預測機率。模型建後好以後，使用categorical cross entropy計算cost，並使用adam解最佳化問題。

1. CNN

CNN model的前處理以RNN資料的前處理為基礎，再將每一個frame和其前後兩個frame concatenate在一起，形成一個維度為（3, 39）的frame。第一個frame和最後一個frame則在左右個補上39維的0。

RNN + CNN 的model與RNN相似。但在masking之前加入了ConvLSTM2D。這個方法可能沒有辦法很正確的做到masking，因為padding的input在經過convLSTM2D之後不會全為0。但keras的ConvLSTM2D不支援masking，所以沒有辦法將資料傳進model後馬上做masking。經過Convolution之後，將結果攤平，接著放入兩層LSTM，其後都接著batch normalization，最後接上fully connected做更多非線性轉換。一樣使用softmax轉換成機率，並用categorical cross entropy計算cost，使用adam解最佳化問題。

How to Imporve Performance

1. Bidirectional LSTM和batch normalization
2. 使用Bidirectional LSTM可以讓模型同時考量聲音資料兩個方向的資訊，讓模型可以從更多的資料中抽取更多的資訊。使用batch normalization是因為一般的update在計算時都是assuming其他variable constant的情況下計算出來的，但實際上在update時所有的variable都會變動，為了解決這個問題所以使用batch normalization，讓update時更可能讓cost下降。

Experimental results and settings

1. 我做出來的模型rnn是比較好的。這一方面可能是因為cnn的部分masking做的不是很正確（但應該還是可以學習到的，因為target是0的話，cost還是會是0，所以cost function的錯誤應該只是一個常數項的錯誤，update的方向應該還是正確的），另一方面可能是因為我只是用了前後兩個frame，或許frame數再多一點效果會更好。最後因為覺得cnn已經很複雜了，所以只加上了上兩個普通的LSTM，沒有使用bidirectional LSTM，這也可能造成　cnn的結果較差。

|  |  |
| --- | --- |
| model | accuracy |
| 2 lstm(200 200)+desne+softmax | 62.24% |
| 2 lstm(100 100) | 68% |
| 2 lstm(100 100)+relu+dropout 0.9 | 64.13 |
| 2 lstm(100 100)+relu | 66.16% |
| 3 lstm(100 100 100) + relu dropout 0.9 | 63.22% |
| 1 lstm(600) | 51% |