HW1 106/10/28

報告製作：鄭皓謙

學號:r06922115

TIMIT dataset 分析過程：

作業可以選擇mfcc或是fbank資料集，我選擇了mfcc，配合找到的mfcc wav2features library方便我以後再深入了解speech recognition。

TIMIT為二維資料：

timestep dim	根據句子長度變化
features dim	根據mfcc算法定義，固定維度39，實數域

這次作業要的是phonemes的輸出，而不是frame wise的一對一輸出，原本我想嘗試接兩層RNN做auto-decoder-encoder，不過可能是模型太長而且中間我並沒有另外做masking，所以效果不好，所以最後採用的模型是 1 to 1, frame wise output，然後再另外做triming。

考慮給的資料輸入每個 time step frame有重疊的資訊，以及在features dim男女音可能在features dim上有shift offset的對應關係，先用CNN在兩個軸上做特徵擷取，經過flatten後接上RNN做一對一的輸出，最後做48-39 mapping跟triming。

資料被我padding到最長777長度，所以每個sample後面都會接一長串的padding 0，考慮到資料的平均長度，每筆資料通常有一半是padding的值，後面會提到如何處理的。

由於padding的關係，所以用keras內建的accuracy metric不能夠顯示有效資料長度內的準確率，所以我在loss.py裡面另外寫了loss_with_mask以及acc_with_mask。

模型實作過程：

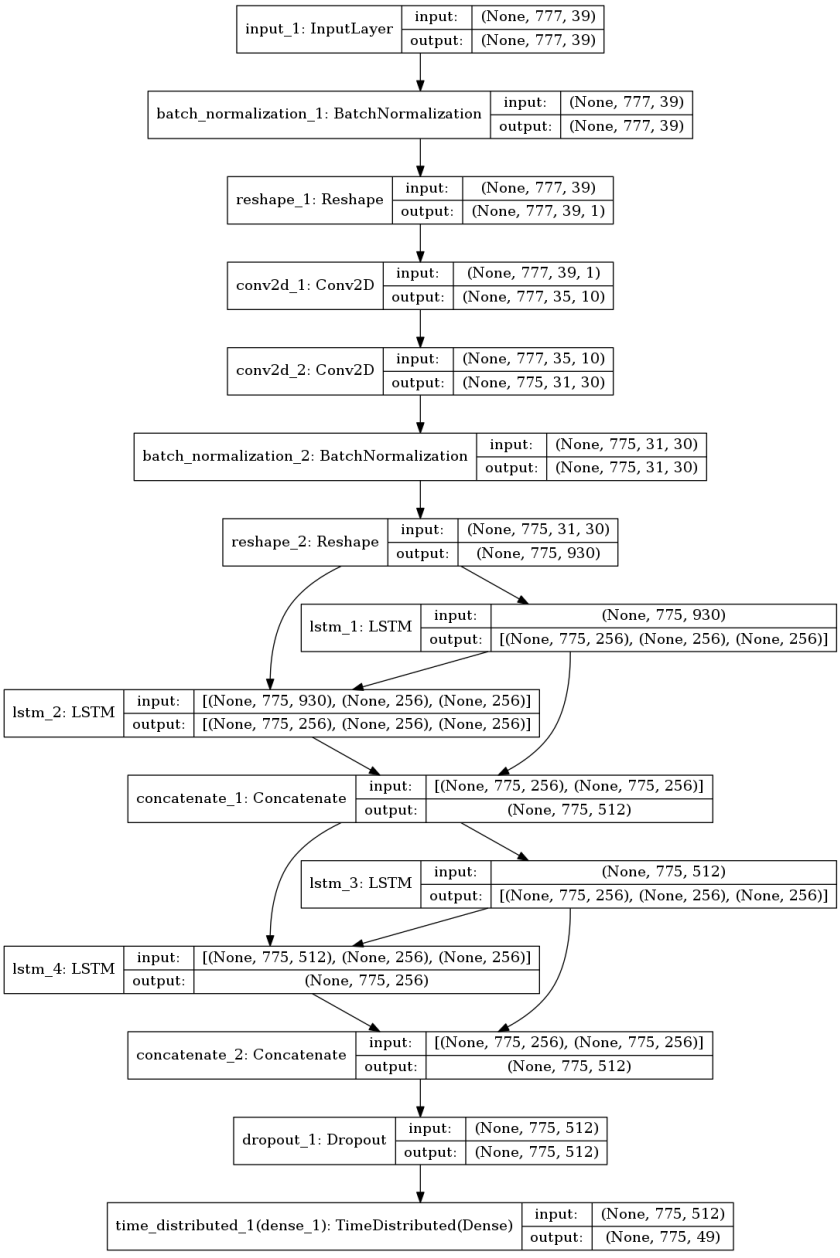
我之前做非固定長度輸出時，用categorical crossentropy，結果輸出都是sil以及padding symbol。

在一對一模型的途中，我發現因為padding資料的關係，大幅影響初期模型gradient decent的方向，這個可以由定義有做masking的loss function去避免，以及配合sample weight去調整。我將sil的權重調到0.2而padding value的權重則調整為0.01。

這次dataset並沒有觀察到明顯subsampling的性質，所以CNN後面並不會接pooling layer，而為了加強在time step上的關係，使用了bidirectional 的LSTM。

兩層CNN —> 兩層RNN —>softmax dense

CNN kernel選用的size是(1,5)跟(3,5)，希望由LSTM去處理time step上關係。



可以改進的地方

由於CNN有bias的關係，不能夠單純加一層Masking把0去掉，所以接到RNN的值裡面會有大部分是padding，調整sample weight的做法並不夠好，這是可以改進的地方，可以再寫一個層masking。試過將CNN的bias關掉，不過效果更差。

這個模型在跑training時容易跑到overfitting，由於我不夠清楚dropout該加在哪些地方，這也是我覺得可以改進的部分。

loss function我是寫了一個有masking的版本，不過因為是一對一模型，如果沒學到怎麼辨識padding的話 效果還是很差，所以我並沒有採用，只透過調整sample weight的方式讓模型盡量在初期不要踩到padding所造成的淺洞。如果是輸出不定長度的模型，我想配合CTC loss可以做得更好。

目前輸出做triming的部分，是將重複phoneme沒超過長度3的當作錯誤判斷去掉，我覺得可以另外練一個模型去學習更廣泛的分佈關係。3這個閾值是根據直覺設定的，因為我也沒有研究資料集裡面的統計特性。

RNN另外把hidden state接出來，原本打算做encoder decoder，不過失敗了，考慮之後加上masking的可能，並沒有用keras的bidirectional layer warper包。