ADLxMLDS 2017 HW1 Report
B03b02014 張皓鈞

1. Model Description

- 前處理：

  首先將每一句話的frames都用全零的vector把frames的數量pad成最大frame數量，預測label也用sil的label作padding。我將訓練資料的矩陣形狀弄成(資料筆數，Frames數目，特徵維度)的形狀，而訓練label則是弄成(資料筆數，Frames數目, 48個phone class)的形狀。

  這次的問題可以看成將frames分成48種phone的分類問題，因此我用one-hot encoding的方式來表示label，訓練時讓模型輸出48個類別的機率，最小化分類交叉熵(categorical crossentropy)。

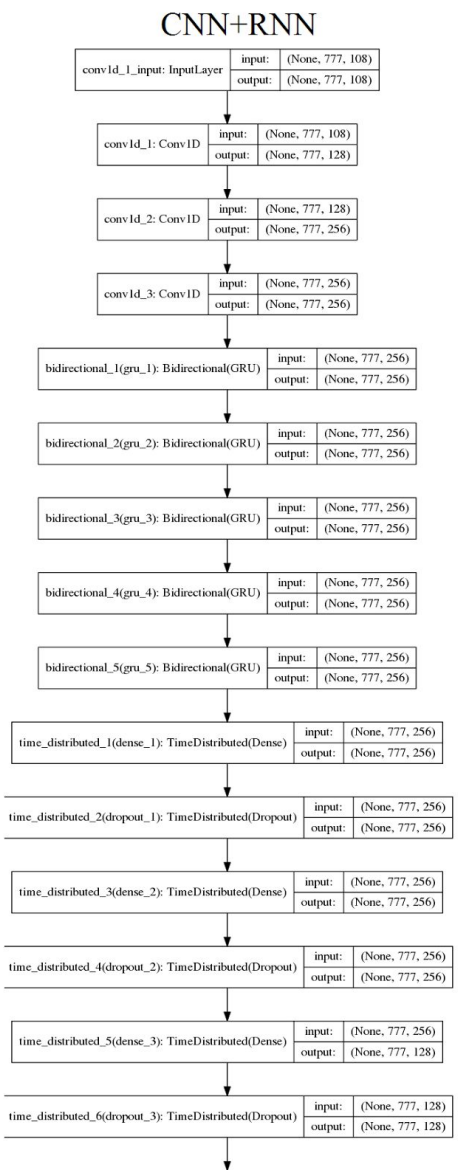
- RNN模型：

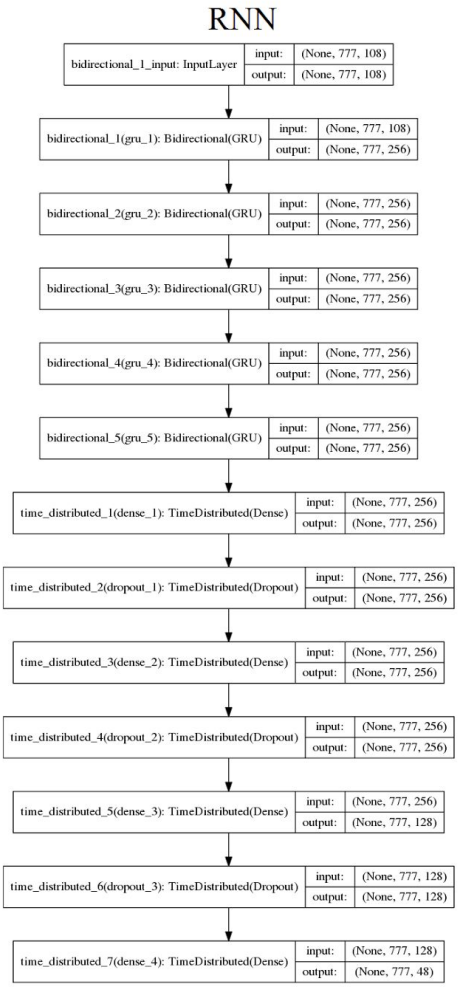
  我使用雙向RNN作為模型架構。　使用雙向的原因在於預測phone上，該frame的前後資訊皆和要預測的frames有關係，而且鄰近的frames在抽取時可能會抽取到部份重疊的聲音訊號，因此使用雙向的RNN可以保留到預測目標的前後資訊。 RNN選擇GRU的原因在於GRU因為比LSTM少了記憶單元(memory unit)的控制，在訓練速度上比LSTM還要快很多，因此選擇使用多層的GRU可以較有效率獲得訓練模型。

- CNN+(Bi)RNN：

  CNN在這裡的作用是作為特徵抽取器，利用多層且多個filter來抽取每一句話frames的特徵，再輸入進雙向RNN來進行訓練。在此CNN使用的是Convolutional 1D layer，Conv1D可以保有frame sequence的長度，並抽取所有frames的特徵。

- 基本模型架構：





2. Model Improvement

- Feature Concatenation：

  將fbank和mfcc的特徵向量接在一起，形成108維的向量，接受兩種方法取得的聲音特徵，有明顯讓模型的預測效果變好。

- Batch Normalization：

  把Convolutional 1D or 2D layer的輸出經果Batch正規化能讓模型的訓練更快收斂，因為特徵值之間的數值範圍被縮小到固定的範圍裡。

- Blending：

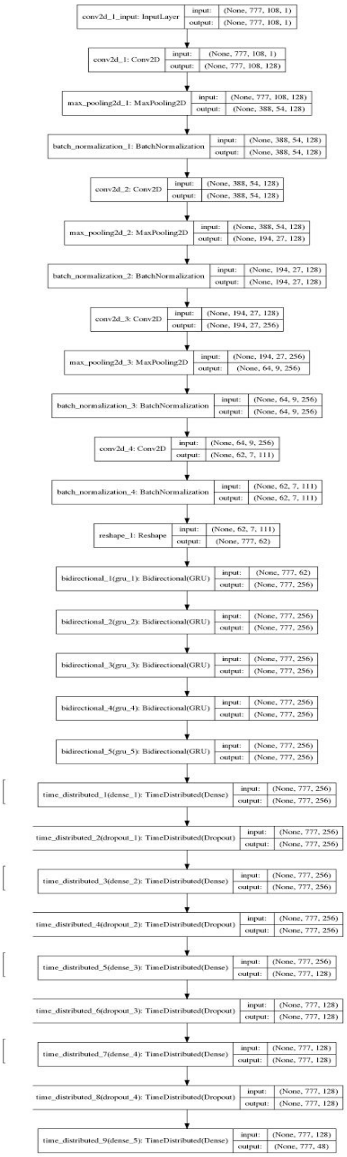
  將數個模型輸出的48個類別的機率分別按模型本身的預測能力進行加權平均，綜合三個模型的預測結果，增加互補的可能性。

3. Experimental results and settings

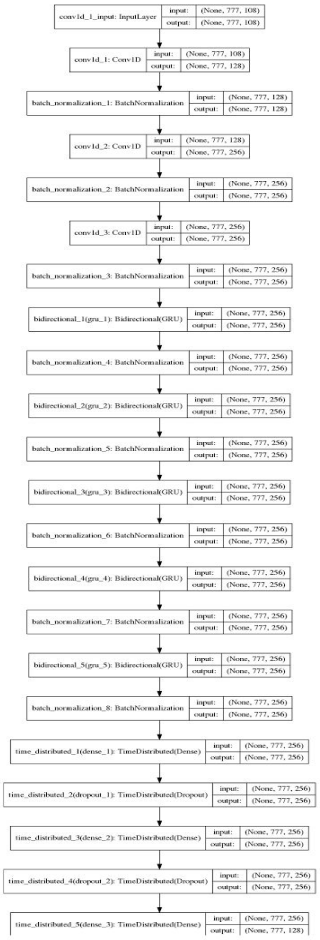
- 嘗試過的模型：
  - Bi-RNN：此模型架構為上述基本的雙向RNN
  - Conv1D+Bi-RNN：此模型架構為上述基本的CNN+雙向RNN
  - Conv1D(w/ BN)+Bi-RNN：此為CNN+雙向RNN，每層Conv1D的輸出皆經過Batch Normalization
  - Conv1D(w/ BN)+Bi-RNN(w/ BN)：此為上述的CNN+RNN，每層的Conv1D和Bidirectional GRU的輸出皆經過Batch Normalization
  - Conv2D(w/ BN)+Bi-RNN：我們另外嘗試將每句話的frames看成是一張(777, 108, 1)的圖片，並用2D的convolutional layer來抽取特徵。

  以上模型訓練資料皆經過前處理且將兩種特徵接在一起，皆訓練至training / validation loss < 0.4以及training / validation Phone準確率 > 0.9才停止 。
- 詳細模型架構：

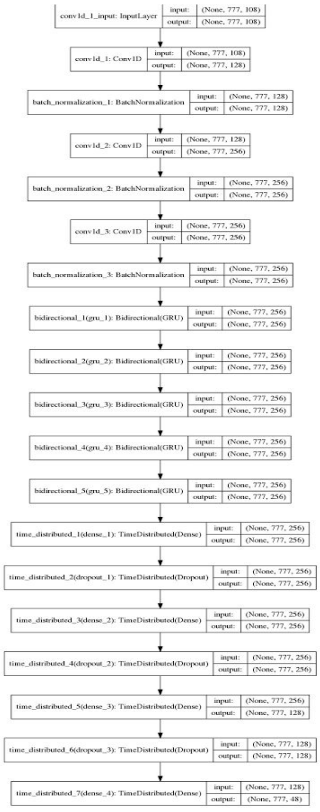
Conv2D



Conv1D (all BN)



Conv1D (conv BN)



- 訓練結果：

(Accuracy為phone預測準確率，edit distance則是此模型在public testing data上預測句子的phone sequence的平均edit distance)

Model	Bi-RNN	Conv1D + Bi-RNN	Conv1D (w/ BN) + Bi-RNN	Conv2D (w/ BN) + Bi-RNN	Conv1D(w/ BN) + Bi-RNN(w/ BN)
Total # of parameters	1,535,408	2,238,000	2,240,560	2,264,155	2,245,680
Edit distance	11.99435	13.71186	10.34463	19.73446	10.87570

- 比較分析：

1. 從最終的預測結果來說，我們可以看到雙向RNN的表現比加上Conv1D的雙向RNN還要好，然而Conv1D經過Batch正規化後卻可以超越單純的雙向RNN。我的推測是因為BN可以將通過Conv1D的特徵壓縮到範圍相近的數值，在RNN進行gradient descent可以有比較快的收斂速度。
2. 關於另外嘗試的Conv2D，和Conv1D的差異在於Conv2D有使用到MaxPooling將feature map抽取成較低維度的特徵，且並非一次性地從整個frame序列抽取特徵，而是將frames x 特徵看成一張圖片，分區域抽取特徵。 在訓練速度上，Conv2D的收斂速度明顯緩慢，Loss和準確率的下降也比Conv1D來得緩慢，而且訓練資料準確率在0.89時，Validation的準確率還在0.84左右，顯示有overfitting的現象。 在預測能力方面，Conv2D的表現明顯較差，原因推測是因為MaxPooling的處理會喪失一些序列上連續性的特徵。
3. 在Conv1D都有BN的前提下，將雙向RNN的輸出也經過BN並沒有明顯的進步。