壹、題目

Listen & Translate

貳、Team Name

隊伍名稱：

★☆無限期支持宿舍炊膳合法化☆★

Kaggle Name：

NTU_b03502125_★☆無限期支持宿舍炊膳合法化☆★

參、Members

組長：

倪嘉宏 b03502125

組員：

黃文鴻 b03701118

組員：

徐一真 b03901063

組員：

林威利 b03902047

肆、Work Divisions

倪嘉宏：行政、報告撰寫、第一種模型(初期)、實驗

黃文鴻：工作站維護、報告撰寫、開發環境維護、第一種模型(初期)

徐一真：第一種模型、報告撰寫

林威利：第一種模型(初期)、第二種模型、報告撰寫、實驗

伍、Preprocess & Featuring Engine

一、第一種 model:

1.Training Data

送入的資料包含已經處理好的音訊MFCC，以及對應的中文翻譯。這些Training Data讀入後，我們會先將對應的中文翻譯的部份每句前都加上'\t'、而每句後面都加上'\n' 之後將整個句子逐字編碼，編碼採用one-hot encoding。

2.Testing Data

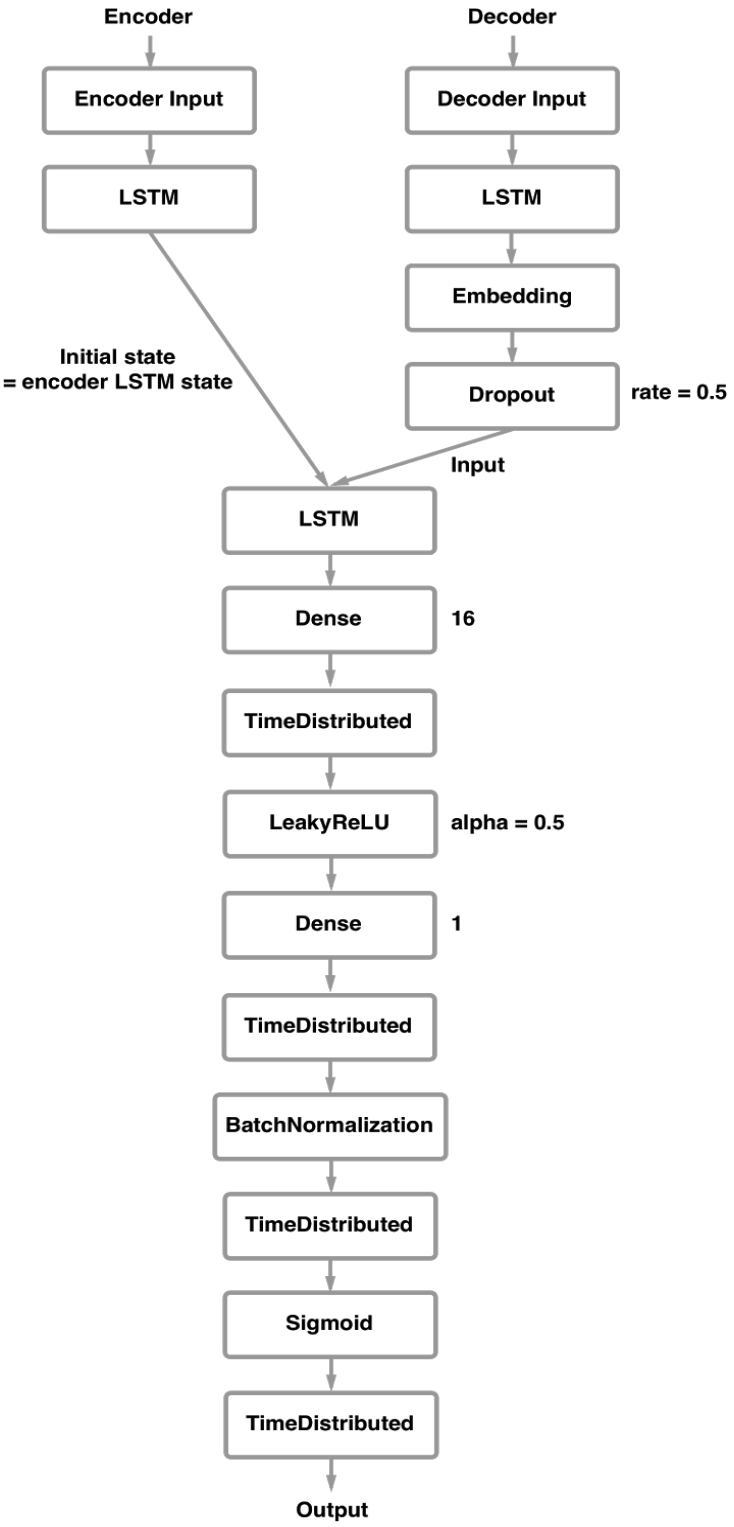
Testing Data 的部分，preprocessing的過程與 training data一樣。所有的答案(中文句子)都會通過與Training Data 相同的編碼。

二、第二種 model:

基本上與第一種 model 相同，但 caption 部分的編碼採用單純 zero padding，沒有額外加上 '\t' 與 '\n'。

陸、Model Discription

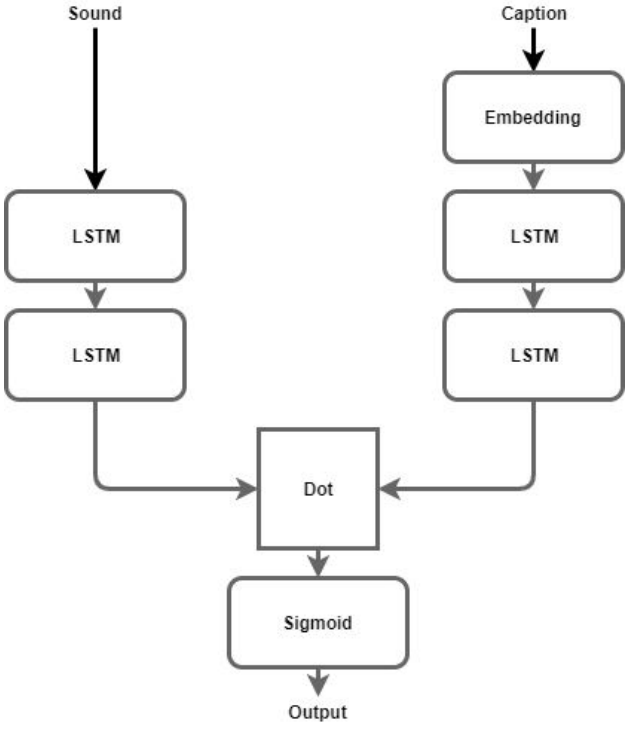
一、本組所使用的第一種 model 架構如圖所示：



MFCC資料當作encoder input，而將編碼後的中文翻譯當作decoder input，並設定答案為1。另外輸入的training data我們會再將MFCC部分搭配一組錯誤的中文翻譯，當作錯誤的範例並將答案設定為0。最後Model的輸入是encoder input 跟 decoder input，而輸出的答案與正確答案進行一個fit的動作。此Model所使用到的package包含：

- keras
- tensorflow

二、第二種 model 架構如圖所示：



我們的第二種model一樣先sample出正確與錯誤的範例，透過LSTM分別encode 聲音與文字的embedding之後做內積，讓正確的部分盡量接近 1，而錯誤的部分則是盡量接近 0。

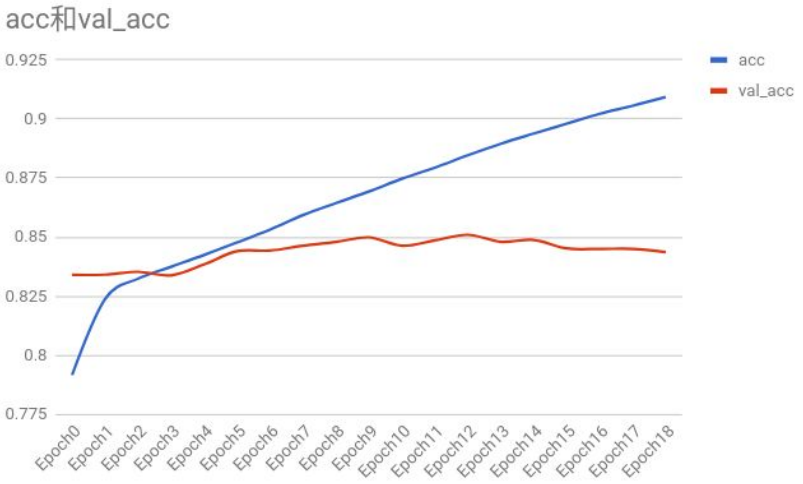
柒、Experiments & Discussions

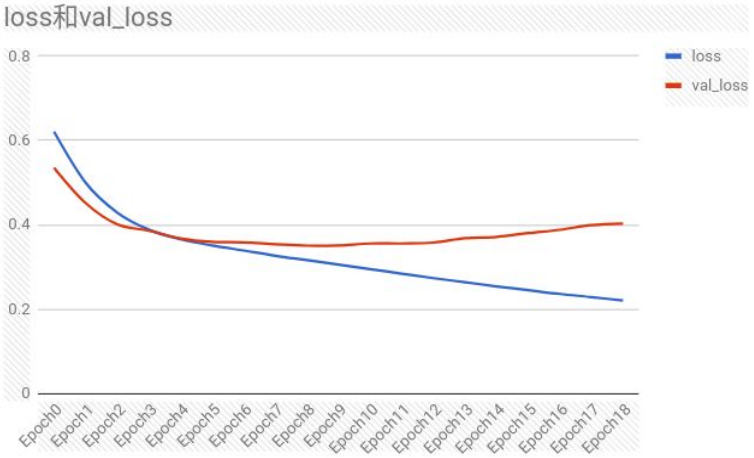
我們對兩種模型分別做了不同的實驗與討論，因此以下將各自描述實驗內容與結果。

一、第一種模型

1.模型訓練過程探討

下兩圖為分別為訓練的epoch數量對loss與對accuracy的做圖。需注意的是Accuracy該圖是用我們字己出的題目來計算的（也就是Model Discription一節所說的把正確答案標上1，再把MFCC資料對一個不正確的答案標上0。正確的答案與錯誤的答案分別佔一半。這種出法會使得隨機猜出的答案accuracy為0.5）。可以看出其實大約在第五個 epoch 就達到loss與 accuracy最佳的狀態，後面的epoch都只會使得訓練的結果overfit。





2.模型參數探討

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	(None, None)	0	
embedding_1 (Embedding)	(None, None, 256)	256	input_2[0][0]
input_1 (InputLayer)	(None, None, 39)	0	
dropout_1 (Dropout)	(None, None, 256)	0	embedding_1[0][0]
lstm_1 (LSTM)	[(None, 256), (None,	303104	input_1[0][0]
lstm_2 (LSTM)	[(None, None, 256),	525312	dropout_1[0][0] lstm_1[0][1] lstm_1[0][2]
time_distributed_1 (TimeDistrib	(None, None, 16)	4112	lstm_2[0][0]
time_distributed_2 (TimeDistrib	(None, None, 16)	0	time_distributed_1[0][0]
time_distributed_3 (TimeDistrib	(None, None, 1)	17	time_distributed_2[0][0]
time_distributed_4 (TimeDistrib	(None, None, 1)	4	time_distributed_3[0][0]
time_distributed_5 (TimeDistrib	(None, None, 1)	0	time_distributed_4[0][0]

Total params: 832,805
Trainable params: 832,803
Non-trainable params: 2

參數自變數以外的變數都為控制變數，下面探討的變數當中各項的預設參數分別為data_size = 5，leaky_alpha = 0.5，dense_width = 64

(1)關於 time_distributed_3 當中的 LeakyReLU

time_distributed_3為一 Leaky Rectified Linear 層。調整其 leak 的斜率對於最後的正確率有顯著的影響。簡單列表如下：

Leaky alpha	Accuracy	Loss
0.3	0.8464	0.3512
0.5	0.844	0.3617
1.0	0.8458	0.3582

(2)關於 time_distributed_1當中的 dense_width

time_distributed_1為一dense層。此參數為dense的node數量

dense_width	Accuracy	Loss
16	0.844	0.3617
32	0.8484	0.346
64	0.851	0.3408

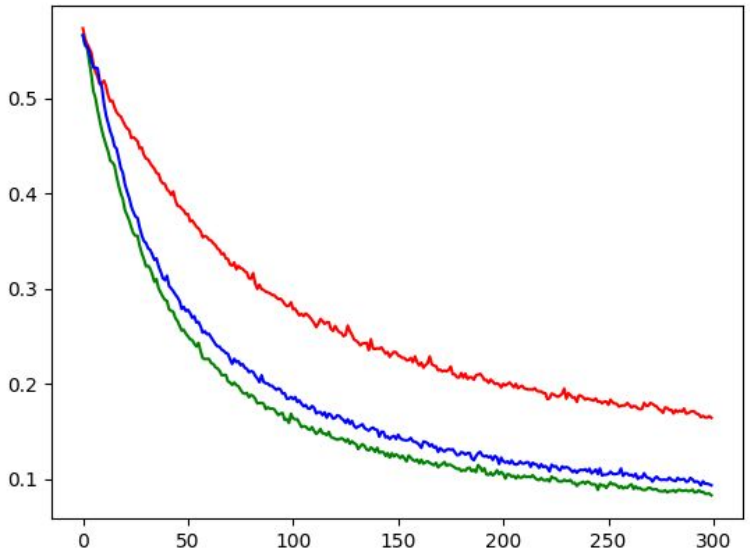
(3)關於 data_size

我們的訓練過程需要給電腦一份資料，資料當中包含許多組nfcc資料與中文翻譯的配對，其中有些配對是正確有些配對是錯誤的。而產生錯誤配對的方法是由給定的NFCC配上其原本對應中文翻譯的第下n筆翻譯。n的大小分別是1至data_size。舉例來說，data_size是5的情況下，產生的總資料數量會是題目給定的資料數量的6倍多，而其中第1份資料答案被標注為1，第2至5份資料答案皆被標注為0，而內容則分別是NFCC資料對應其原本中文翻譯的後1至4筆翻譯。

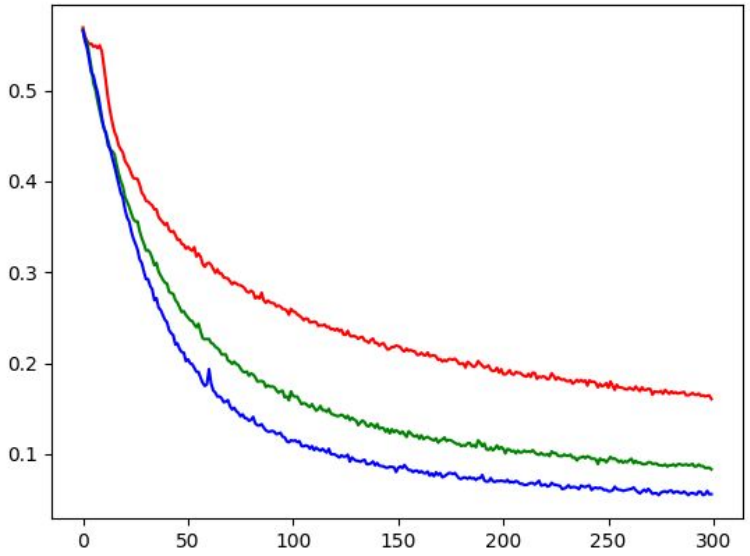
data_size	Accuracy	Loss
1	0.6646	0.6012
2	0.7088	0.5386
3	0.7872	0.4416
4	0.8252	0.3857
5	0.844	0.3617

二、第二種模型

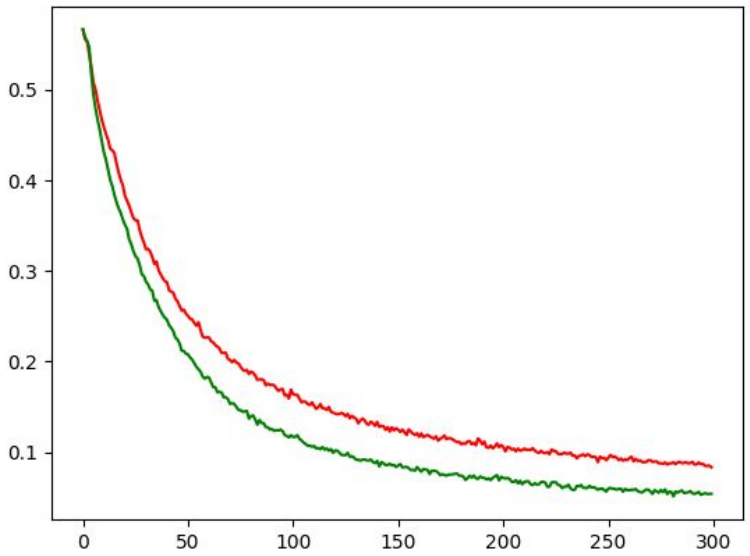
1.LSTM層數

X軸: epoch Y軸: training loss 紅色: 單層LSTM 綠色: 雙層LSTM 藍色: 三層LSTM	

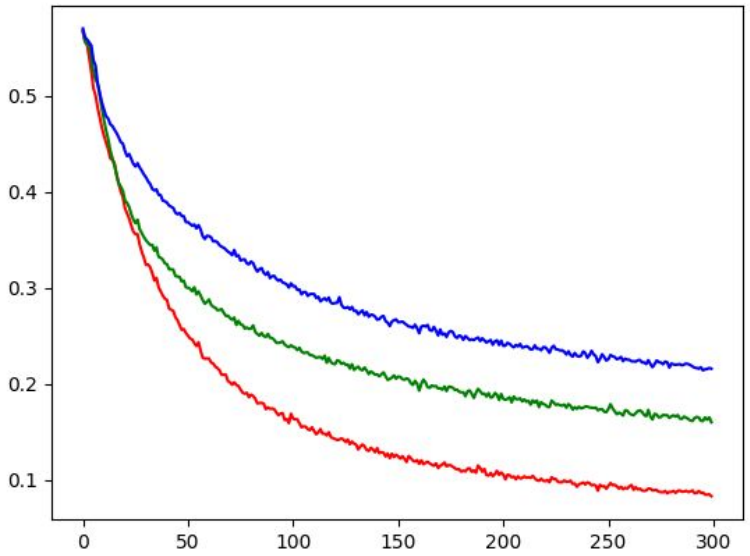
2.LSTM hidden dimension

X軸: epoch Y軸: training loss 紅色: hidden = 50 綠色: hidden = 100 藍色: hidden = 200	

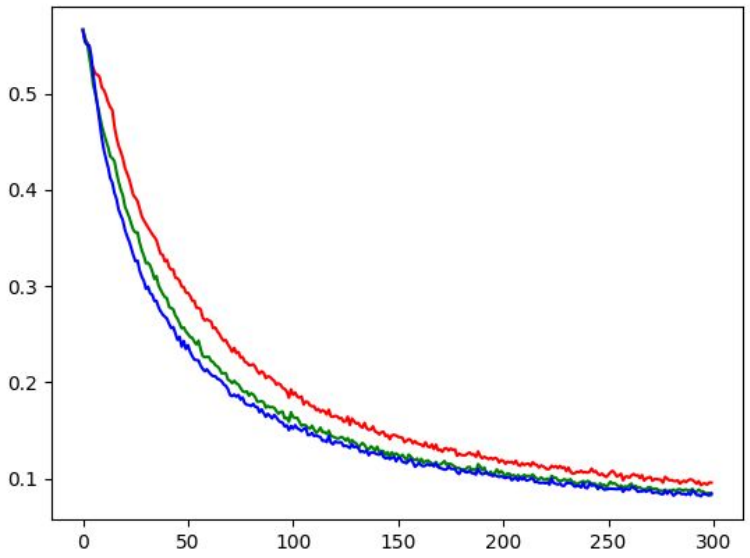
3.bidirectional LSTM

X軸: epoch Y軸: training loss 紅色: LSTM 綠色: bidirectional LSTM	

4.Dropout

X軸: epoch Y軸: training loss 紅色: dropout=0.0 綠色: dropout=0.25 藍色: dropout=0.5	

5.word embedding dimension

X軸: epoch Y軸: training loss 紅色: hidden = 50 綠色: hidden = 100 藍色: hidden = 200	

捌、Reference

[1]Krupakar, H., Rajvel, K., Bharathi, B., Deborah, S. A., & Krishnamurthy, V. (2016, February). A survey of voice translation methodologies—Acoustic dialect decoder. In *Information Communication and Embedded Systems (ICICES), 2016 International Conference on* (pp. 1-9). IEEE.

[2]Wöllmer, M., Zhang, Z., Weninger, F., Schuller, B., & Rigoll, G. (2013, May). Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 6822-6826). IEEE.

[3]Bérard, A., Pietquin, O., Servan, C., & Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.