Homework 1 - End-to-end Speech Recognition

學號：R08942085 系級：電信丙 姓名：陳芃彣

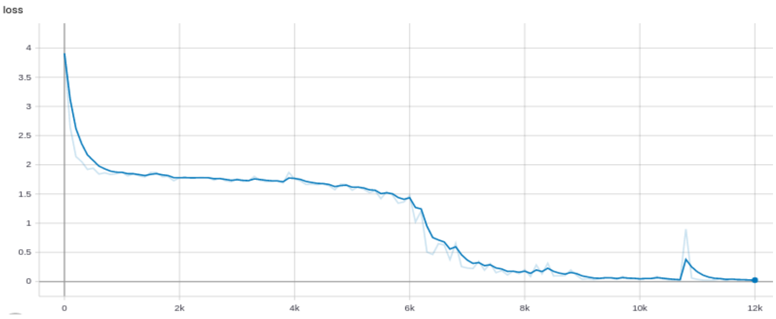
學號：R08942086 系級：電信丙 姓名：趙達軒

學號：R07942091 系級：電信丙 姓名：許博閎

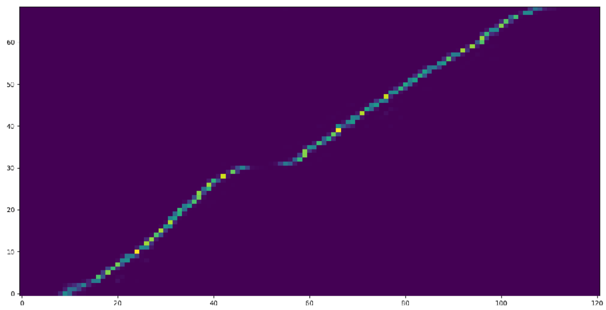
學號: R08945002 系級：生醫電資 姓名：陳玠妏

1. **(2%) Train a seq2seq attention-based ASR model. Paste the learning curve and alignment plot from tensorboard. Report the CER/WER of dev set and kaggle score of testing set.**

Learning Curve:



Alignment plot:



CER/WER of dev set:

```
===========  Result of result/decode_example_dev_output.csv ===========
-----------------------------------------------------------------------
| Statics             |  Truth        |  Prediction    | Abs. Diff.    |
-----------------------------------------------------------------------
| Avg. # of chars     |  66.99        |  66.93         |  0.48         |
| Avg. # of words     |  17.14        |  17.11         |  0.03         |
-----------------------------------------------------------------------
---------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
---------------------------------------------------------------
| Character     | 2.7174        | 2.56          | 0.00/18.18    |
| Word          | 8.9494        | 7.69          | 0.00/50.00    |
---------------------------------------------------------------
```

Score of testing set:

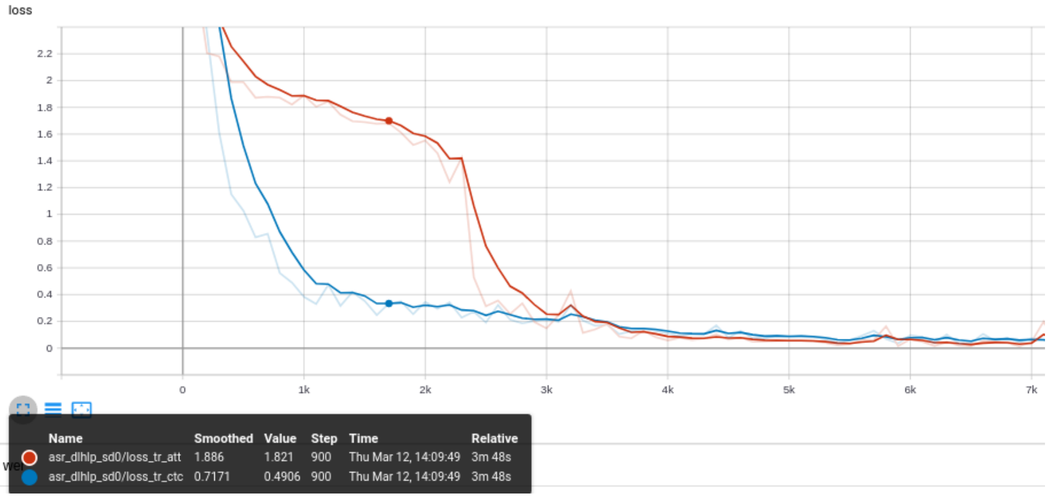
Name	Submitted	Wait time	Execution time	Score
asr_v1.csv	just now	0 seconds	0 seconds	1.77600

Complete

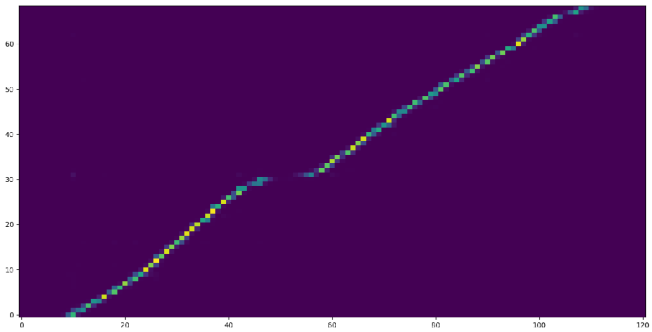
Jump to your position on the leaderboard

2. (2%) Repeat 1. by training a joint CTC-attention ASR model (decoding with seq2seq decoder). Which model converges faster? Explain why.

Learning Curve: blut: ctc, red: att



Alignment plot:



CER/WER of dev set:

```
============  Result of 3_dev.csv ============
-----------------------------------------------------------------------
| Statics              |  Truth        |  Prediction     | Abs. Diff.   |
-----------------------------------------------------------------------
| Avg. # of chars      |  66.91        |  66.99          |  0.42        |
| Avg. # of words      |  17.11        |  17.14          |  0.03        |
-----------------------------------------------------------------------
-----------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
-----------------------------------------------------------------
| Character     | 2.2535        | 2.42          | 0.00/20.00    |
| Word          | 7.6211        | 7.60          | 0.00/75.00    |
-----------------------------------------------------------------
```

Score of testing set:

alex_beam_process.csv 1.09600

2 hours ago by r08942085

在語音識別中，我們的數據集是音頻文件和其對應的文本，但通常音頻文件和文本很難再單詞的單位上對齊。因此需要在預處理操作時進行對齊，如果不使用對齊而直接訓練模型時，由於人的語速的不同，或者字符間距離的不同，會導致模型很難收斂。

CTC 引入了 blank，每個預測的分類對應的一整段語音中的一個 spike，其他不是 spike 的位置認為是 blank。對於一段語音，CTC 最後的輸出是 spike 的序列，並不關心每一個音素持續

了多長時間。

ASR 的過程是一幀 MFCC39 維向量進去，然後出一個 label。假設，“你好” 這個音訊共有 200 個 MFCC 特徵幀。這 200 個特徵幀對應著 200 個輸出結果，就結果空間而言，共有音素數目^200 種可能。

而 CTC 認為，計算目標函式的時候，上例中的 200 個 MFCC 特徵，得到的 200 個模型的結果，每個小結果都對應著所有音素上的一個概率分佈。然後計算所有能對映成“ㄋㄧˇㄏㄠˇ“結果的音素路徑的概率值，讓這個值越大越好就行了。

但是這樣一來，計算量就非常的大，指數級的計算量。CTC 就使用了類似 HMM 中的向前向後算法來計算。發現進行反向傳播的時候，每一幀 MFCC 對應的結果的導數，都可以利用前一時刻的兩個狀態的結果直接求到。這樣一來，整體計算量就急劇萎縮成了 7*T*音素個數，因此更容易收斂。

3. (2%) Use the model in 2. to decode only in CTC (ctc_weight=1.0). Report the CER/WER of dev set and kaggle score of testing set. Which model performs better in 1. 2. 3.? Explain why.

CER/WER of dev set:

```
============  Result of 2_dev.csv ============
 ------------------------------------------------------------------------
| Statics                | Truth          | Prediction     | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars        | 66.73          | 66.99          | 0.70          |
| Avg. # of words        | 17.07          | 17.14          | 0.07          |
 ------------------------------------------------------------------------
 ----------------------------------------------------------------
| Error Rate (%)| Mean           | Std.           | Min./Max.      |
 ----------------------------------------------------------------
| Character     | 2.9846         | 2.72           | 0.00/25.00     |
| Word          | 10.5576        | 9.04           | 0.00/75.00     |
 ----------------------------------------------------------------
```

Score of testing set:

2.csv 1.62800

a few seconds ago by r08942085

add submission details

jointly trained encoder + seqtoseq decoder 成績比較好（第 2 題）

可能原因為純 CTC 解碼通過預測每個幀的輸出來識別語音，算法的實現基於假設每幀的解碼保持彼此獨立，因而缺乏解碼過程中前後語音特徵之間的聯繫，比較依賴語言模型的修正。

4. (2%) Train an external language model. Use it to help the model in 1. to decode. Report the CER/WER of dev set and kaggle score of testing set.

CER/WER of dev set:

```
============  Result of 4_dev.csv ============
 -----------------------------------------------------------------------
| Statics                |  Truth          |  Prediction    | Abs. Diff.    |
 -----------------------------------------------------------------------
| Avg. # of chars        |  66.94          |  66.99         |  0.40         |
| Avg. # of words        |  17.11          |  17.14         |  0.03         |
 -----------------------------------------------------------------------
 ---------------------------------------------------------------
| Error Rate (%)| Mean           | Std.           | Min./Max.      |
 ---------------------------------------------------------------
| Character     | 1.9814         | 2.37           | 0.00/20.00     |
| Word          | 6.5786         | 7.18           | 0.00/75.00     |
 ---------------------------------------------------------------
```

Score of testing set:

4.csv 1.15399

9 minutes ago by r08942085

plus lm

5. (2%) Try decoding the model in 4. with different beam size (e.g. 2, 5, 10, 20, 50). Which beam size is the best?

2 is the best

beam size =2:

4.csv 1.15399

9 minutes ago by r08942085

plus lm

beam size =5:

5_beam.csv 1.54800

just now by r08942085

plus lm beam_size=5

beam size =20:

20_beam.csv 1.54800

3 minutes ago by r08942085

add submission details

Bonus: (1%)