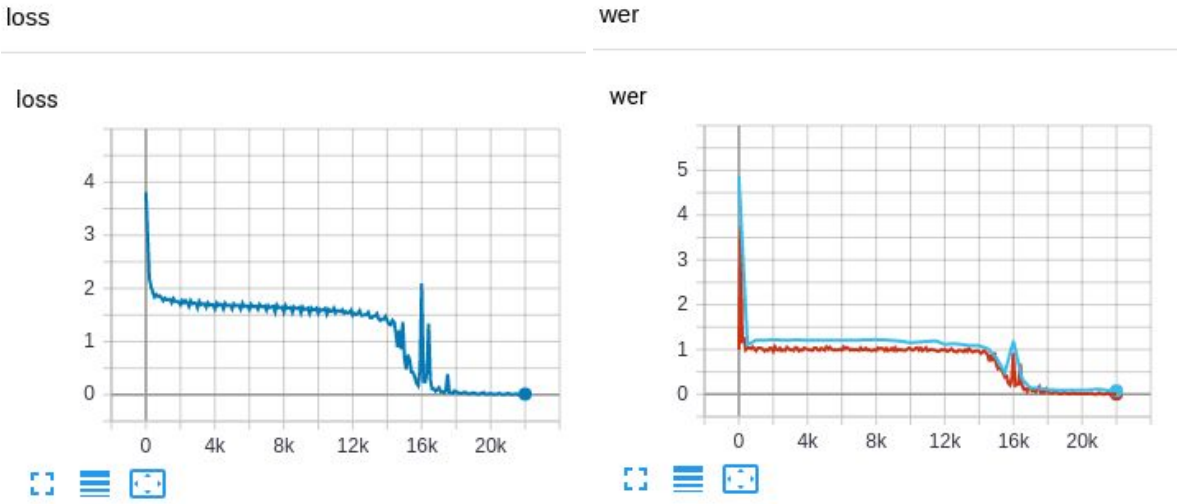
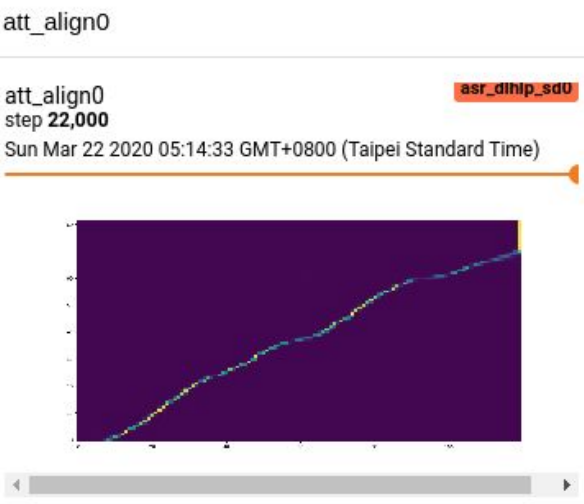
Homework 1 - End-to-end Speech Recognition

學號：R08922A02 系級： AI所碩一 姓名：張碩恩
學號：R08922130 系級： 資工所碩一 姓名：丁杰

---

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.



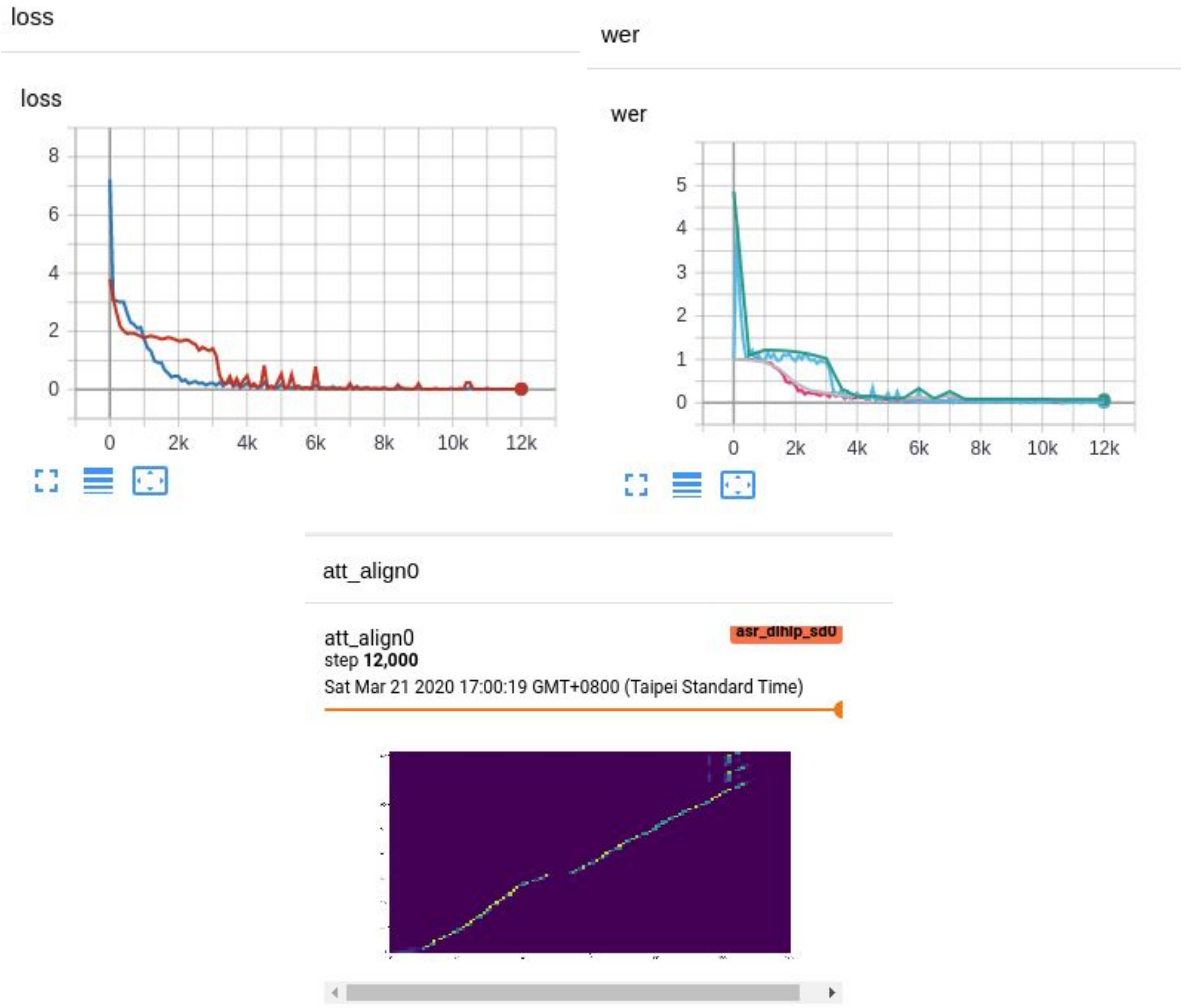



```
============  Result of result/decode_dlhlp_dev_output.csv ============
 ------------------------------------------------------------------------
| Statics              |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars      |  66.99        |  66.76        |  0.48         |
| Avg. # of words      |  17.14        |  17.11        |  0.03         |
 ------------------------------------------------------------------------

 ------------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------------
| Character     | 2.4852        | 2.94          | 0.00/37.50    |
| Word          | 7.9416        | 7.87          | 0.00/75.00    |
 ------------------------------------------------------------------
```

kaggle score =1.50600

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




```
============  Result of dev.csv ============
 ------------------------------------------------------------------------
| Statics               |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars       |  66.99        |  66.92        |  0.37         |
| Avg. # of words       |  17.14        |  17.12        |  0.02         |
 ------------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 2.1197        | 2.50          | 0.00/31.25    |
| Word          | 7.1834        | 7.59          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.254

Joint CTC-attention model converges much faster than only seq2seq model. We think it is because there are more parameters in seq2seq attention-based ASR model (one encoder, one decoder and an attention module) need to be optimized than CTC model (only encoder and a classifier need to be trained). So if we using joint CTC-attention, the CTC model can help the whole model to find the better parameter faster.

On the other hands, compare to seq2seq model, CTC doesn't need using character inter-dependencies. It can also be a reason why joint CTC-attention model can converge faster than only seq2seq model.

In the loss changing during training procedure (the figure above for example), we can see that seq2seq (red line) always converge after CTC (blue line) converge, which can prove for our opinion.

---

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.

```
============  Result of dev.csv ============
 ---------------------------------------------------------------------
| Statics              |  Truth       |  Prediction   | Abs. Diff.    |
 ---------------------------------------------------------------------
| Avg. # of chars      |  66.99       |  66.93        |  0.43         |
| Avg. # of words      |  17.14       |  17.12        |  0.02         |
 ---------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean         | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 2.2493       | 2.44          | 0.00/22.73    |
| Word          | 7.8013       | 7.93          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.31

2>3>1, it is easy to find out that joint CTC-attention have a better performance. Using multi-task method can give a better performance, the reason is that in the similar target task. One task can also help the others, which is called "Solidarity".

Compared between LAS method and CTC method, LAS method gives the better performance. According to Kim, Suyoun *et al.*[1], since the attention model does not use any conditional independence assumption, it has often shown to improve Character Error Rate (CER) than CTC when no external language model is used.

---

[1] Kim, Suyoun et al. "Joint CTC-attention based end-to-end speech recognition using multi-task learning." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016): 4835-4839.

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.

```
===========  Result of result/decode_dlhlp_dev_output.csv ===========
 ------------------------------------------------------------------------
| Statics              |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars      |  66.99        |  66.94        |  0.35         |
| Avg. # of words      |  17.14        |  17.12        |  0.02         |
 ------------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean         | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 1.9094       | 2.46          | 0.00/33.33    |
| Word          | 6.3315       | 7.27          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.13399

---

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

beam size = 2

```
===========  Result of result/decode_dlhlp_dev_output.csv ===========
 ------------------------------------------------------------------------
| Statics              |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars      |  66.99        |  66.94        |  0.35         |
| Avg. # of words      |  17.14        |  17.12        |  0.02         |
 ------------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean         | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 1.9094       | 2.46          | 0.00/33.33    |
| Word          | 6.3315       | 7.27          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.13399

---

beam size = 5

```
============  Result of result/decode_dlhlp_dev_output.csv ============
 ------------------------------------------------------------------------
| Statics               |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars       |  66.99        |  66.94        |  0.35         |
| Avg. # of words       |  17.14        |  17.12        |  0.02         |
 ------------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 1.8776        | 2.45          | 0.00/33.33    |
| Word          | 6.2310        | 7.22          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.122

---

beam size = 10

```
============  Result of dev.csv ============
 ------------------------------------------------------------------------
| Statics               |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars       |  66.99        |  66.94        |  0.35         |
| Avg. # of words       |  17.14        |  17.12        |  0.02         |
 ------------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 1.8776        | 2.45          | 0.00/33.33    |
| Word          | 6.2310        | 7.22          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.122

---

beam size = 20

```
============  Result of dev.csv ============
 ---------------------------------------------------------------------
| Statics              |  Truth        |  Prediction   | Abs. Diff.    |
 ---------------------------------------------------------------------
| Avg. # of chars      |  66.99        |  66.93        |  0.43         |
| Avg. # of words      |  17.14        |  17.12        |  0.02         |
 ---------------------------------------------------------------------
 ------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------
| Character     | 2.2493        | 2.44          | 0.00/22.73    |
| Word          | 7.8013        | 7.93          | 0.00/75.00    |
 ------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

kaggle score = 1.122

Beam size = 5 is the best.
But after beam size=5, larger doesn't makes the performance better. We think that because for ㄅㄆㄇ, only at most four character have a dependence. So, larger beam size doesn't mean the better performance.

Bonus: (1%)

We implemented the beam search algorithm for the prediction of only CTC model (ctc weight = 1.0). Following is how I implemented it. (refer to the code in src/decode.py)

1. Sort the scores among the predictions of the first character and choose the largest k, which is the same as beam size.
2. Store any possible paths that concatenate to next word and compute the scores, and merge the scores together when the last two word of these paths represent the same meaning (ㄅ, _ or _ㄅ, or ㄅㄅ)
3. Sort the scores among the paths and choose the largest k. If this isn't the last word, Go to 2.
4. return the max score of the paths.

When we used the beam_size = 2 for the prediction, we found that this didn't give our prediction any improvement, so as the beam_size = 5.