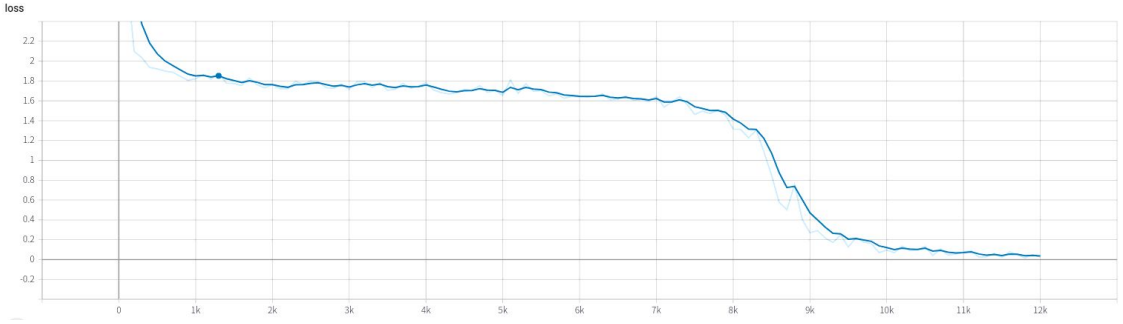
Homework 1 : End-to-end Speech Recognition

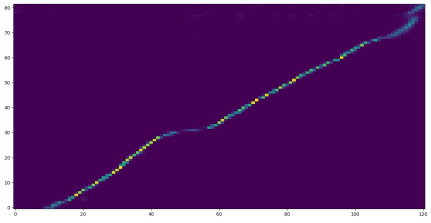
學號：R08942087, R08921040 系級： 電信一, 電機一 姓名：吳彬睿, 徐均筑

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.

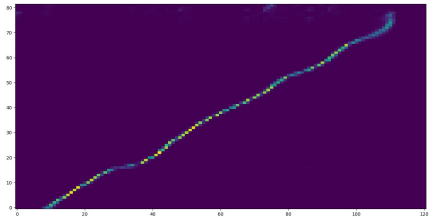
- **Learning curve**



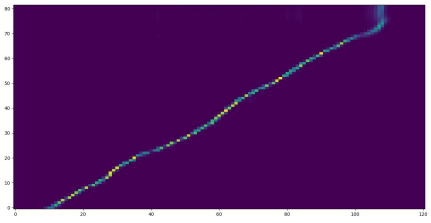
- **Alignment plot**



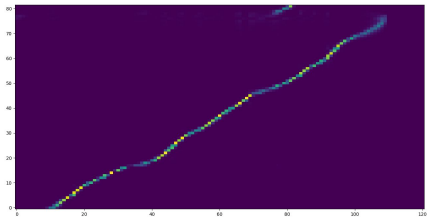
(att_align0)



(att_align1)



(att_align2)



(att_align3)

- **CER/WER**

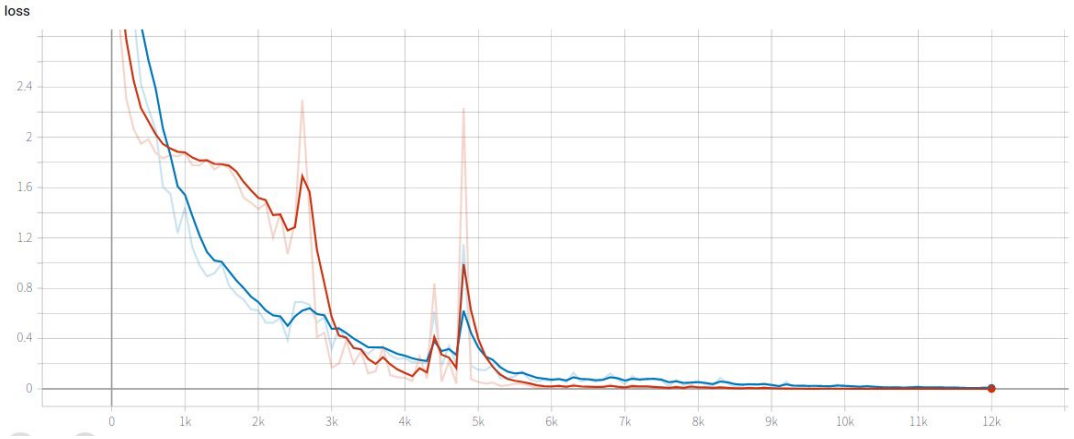
Error Rate (%)	Mean	Std.	Min./Max.
Character	3.1271	5.08	0.00/134.15
Word	9.7909	9.03	0.00/131.82

- **Kaggle score**

Submission and Description	Public Score	Use for Final Score
answer.csv 4 days ago by BinRay Wu	2.20400	☐

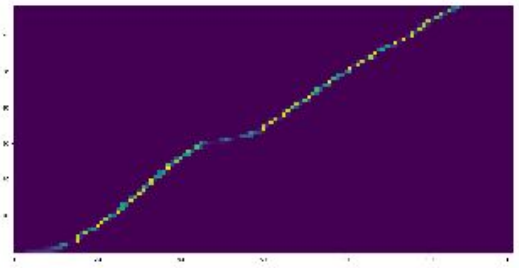
2. (2%) Repeat 1. by training a joint CTC-attention ASR model (decoding with seq2seq decoder). Which model converges faster? Explain why.

- **ctc_weight is set to 0.1 .**
- **Learning curve**

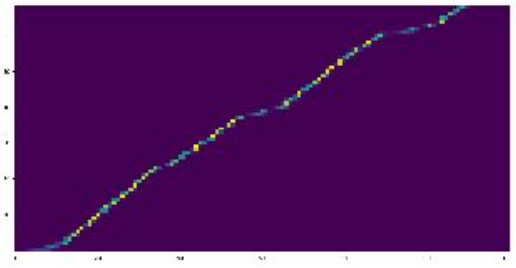


(red) : att
(blue): ctc

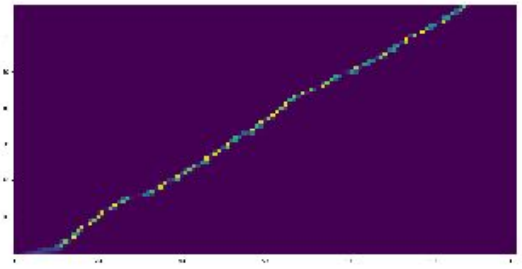
- **Alignment plot**



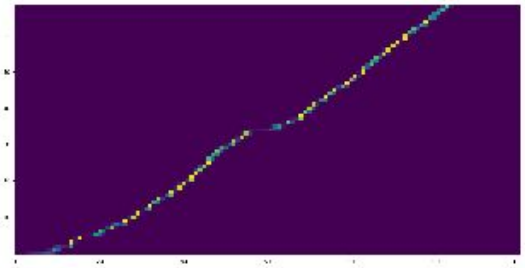
(att_align0)



(att_align1)



(att_align2)



(att_align3)

- **CER/WER**

Error Rate (%)	Mean	Std.	Min./Max.
Character	2.1861	2.28	0.00/19.44
Word	7.4822	7.23	0.00/50.00

Note : If the text unit is phoneme, WER = PER and CER is meaningless.

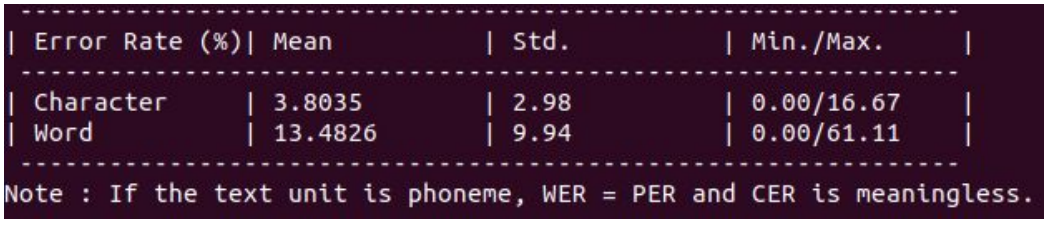
- **Kaggle score**

- **Which model converges faster? Explain why.**

  The attention model is easily affected by noises, and generates misalignments because the model does not have any constraint that guides the alignments be monotonic as in CTC. Compared with the attention model, a joint CTC-attention model is more robust and achieves fast convergence.

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**



Error Rate (%)	Mean	Std.	Min./Max.
Character	3.8035	2.98	0.00/16.67
Word	13.4826	9.94	0.00/61.11

Note : If the text unit is phoneme, WER = PER and CER is meaningless.

- **Kaggle score**

- **Which model performs better in 1. 2. 3.? Explain why.**

  The model in 2. achieves the best performance. Due to the lack of left-to-right constraints as used in CTC, the attention model is too flexible to predict proper alignments. The CTC objective helps guide the attention model during training to be more robust and effective, and produce a better model for speech recognition.

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.

- **Parameter**

  lm_weight = 0.3

- **CER/WER**

```
-------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
-------------------------------------------------------------
| Character     | 2.4586        | 2.61          | 0.00/31.25    |
| Word          | 7.9218        | 7.36          | 0.00/50.00    |
-------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

- **Kaggle score**

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?

- **Parameters**

  lm_weights = 0.3

- **Experiments**

Beam size	CER(%)	WER(%)
2	2.7140 %	8.3334 %
5	2.6484 %	8.1146 %
10	2.6295 %	8.0646 %
20	**2.6269 %**	**8.0596 %**
30	2.6269 %	8.0596 %
50	-	-

Since the vocabulary size is 45, it's impossible to have 50 beams.

- **Which beam size is the best?**

  From the chart above, we observe that the larger beam size can achieve the better result. Moreover, the performance starts to saturate when the beam size reaches 20.

Bonus: (1%)

Best Submition:

- **parameter:**
  train:

  BLSTM * 4 (dropout:0.3)
  ctc_weight: 0.2

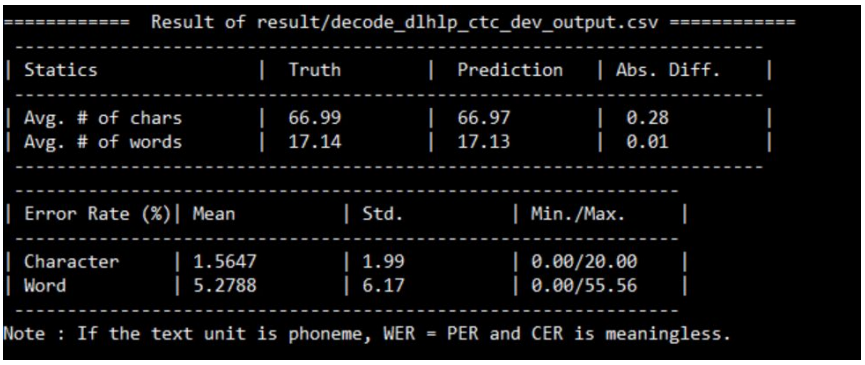
Step: 30001

decode:

beam_size: 10

lm_weight: 0.3

- **CER/WER**



```
============  Result of result/decode_dlhlp_ctc_dev_output.csv ============
 ------------------------------------------------------------------------
| Statics              |  Truth        |  Prediction   | Abs. Diff.    |
 ------------------------------------------------------------------------
| Avg. # of chars      |  66.99        |  66.97        |  0.28         |
| Avg. # of words      |  17.14        |  17.13        |  0.01         |
 ------------------------------------------------------------------------

 ------------------------------------------------------------------
| Error Rate (%)| Mean          | Std.          | Min./Max.     |
 ------------------------------------------------------------------
| Character     | 1.5647        | 1.99          | 0.00/20.00    |
| Word          | 5.2788        | 6.17          | 0.00/55.56    |
 ------------------------------------------------------------------
Note : If the text unit is phoneme, WER = PER and CER is meaningless.
```

References

[1] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in jointCTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in Interspeech, 2017

[2] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in Acoustics, Speech and Signal Processing (ICAS-SP), 2017 IEEE International Conference on. IEEE, 2017