Homework 1 - End-to-end Speech Recognition

學號：r08922080　系級：資工碩一　姓名：簡仲明
學號：r08921062　系級：電機碩一　姓名：黃健祐
學號：b04501127　系級：土木四　姓名：凌于凱

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.

Show data download links

Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing

0

Horizontal Axis

STEP RELATIVE WALL

Runs

Write a regex to filter runs

asr_dlhlp_sd0

asr_dlhlp_sd0/loss_tr_att

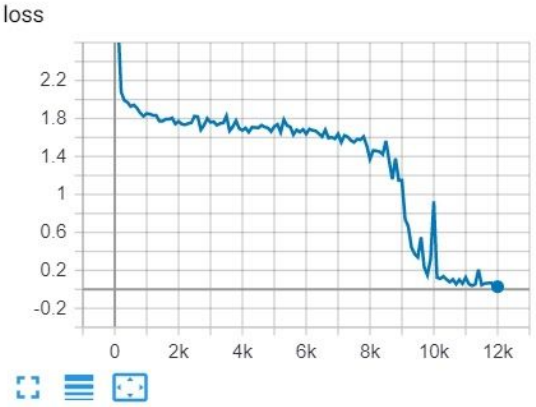
asr_dlhlp_sd0/wer_tr_att

asr_dlhlp_sd0/wer_dv_att

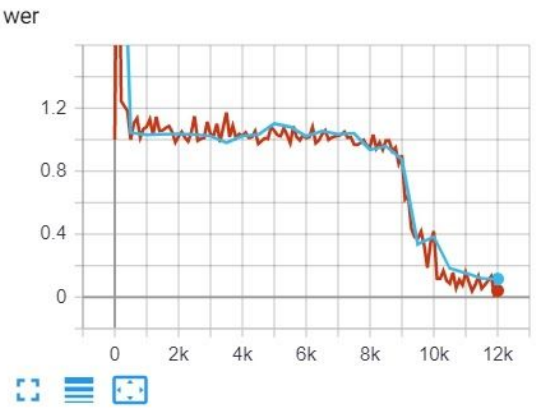
TOGGLE ALL RUNS

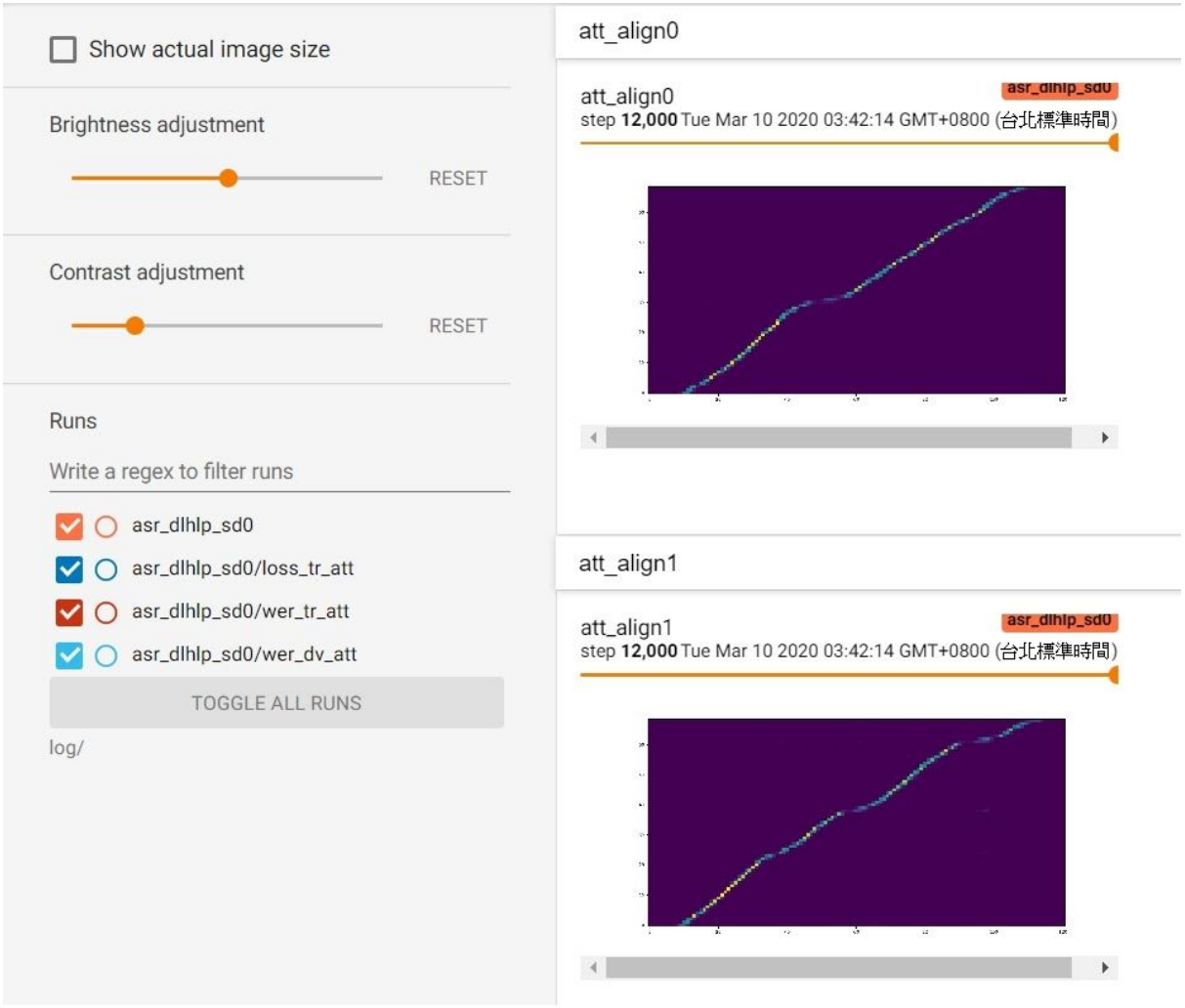
log/

loss



wer





CER/WER of dev set: 3.4286/11.0433
kaggle score of testing set: 2.088

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

Show data download links

Ignore outliers in chart scaling

Tooltip sorting method: default

Smoothing

0

Horizontal Axis

STEP RELATIVE WALL

Runs

Write a regex to filter runs

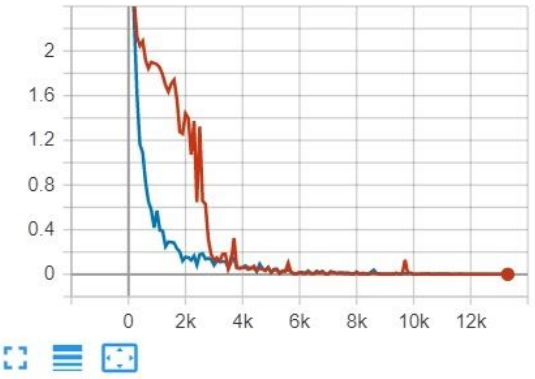
- asr_sd0/loss_tr_ctc
- asr_sd0/loss_tr_att
- asr_sd0/wer_tr_att
- asr_sd0/wer_tr_ctc
- asr_sd0/wer_dv_att
- asr_sd0/wer_dv_ctc

TOGGLE ALL RUNS

log/

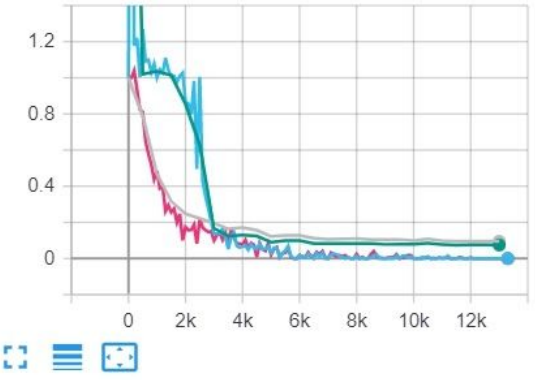
loss

loss



wer

wer



CER/WER of dev set: 2.1502/7.3091
kaggle score of testing set: 1.242

CTC loss helps LAS-ASR model converge faster. Since CTC loss is computed at the output of every encoder time step, it provides clear direction for the training of the encoder model. However, if CTC loss is not used, the only connection between the encoder model and the seq-to-seq loss is the attention layer, which is rather difficult to train.

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: 2.7122/9.4719
kaggle score of testing set: 1.502

The model in 2. performs the best. We have compared the difference between 1. and 2. in the previous problem. For 3., the model is trained with LAS loss and CTC loss together, thus the model outperforms the model in 1., as expected. Better performance of the model in 2. compared with the model in 3. can be credited to the recurrent decoder architecture of LAS, where the generation of an output token is conditioned on the previous decoded tokens. In

the CTC architecture, the tokens are independently generated by a dense network, which takes the encoder states as input.

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: 3.0082/8.9777
kaggle score of testing set: 1.790

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?

I think there is no significant difference between beam sizes larger than 5.

Beam size	2	5	10	20
Dev CER	3.1396	3.0082	2.9854	3.1768
Dev WER	9.3498	8.9777	8.9194	9.0589
Kaggle score	1.856	1.790	1.790	1.788

Bonus: (1%)