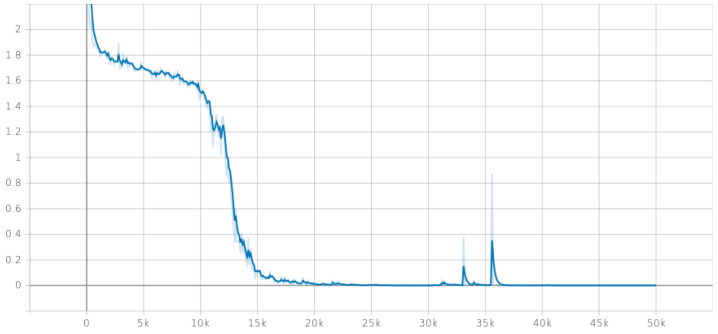
Homework 1 - End-to-end Speech Recognition

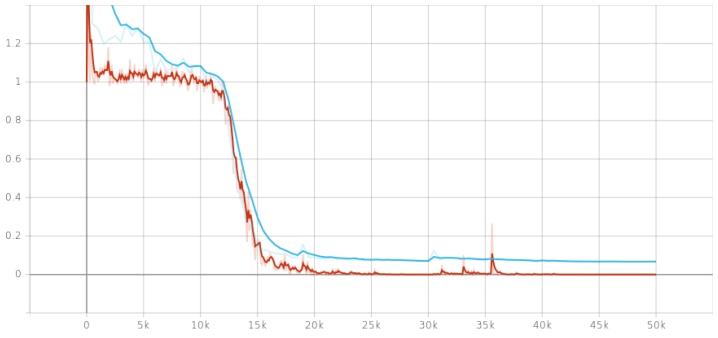
學號：r08944008 系級：網媒所碩一 姓名：簡　義
學號：r07922104 系級：資工所碩二 姓名：林傳祐
學號：r08944024 系級：網媒所碩一 姓名：陳品媛
學號：r06922089 系級：資工所碩二 姓名：邱淳浩

NOTE: reproduce.sh是reproduce作為private評分的那筆submission

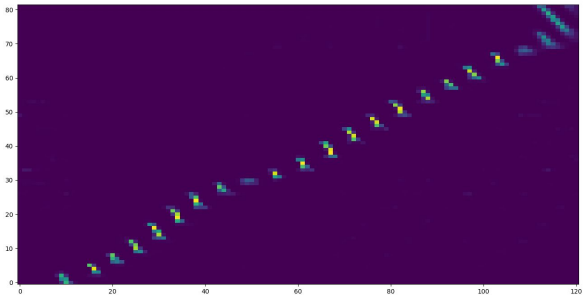
1. (2%) Train a seq2seq attention-based ASR model. Paste the learning curve and alignment plot from tensorboard. Report the CER/WER of dev set and kaggle score of testing set.



training loss.



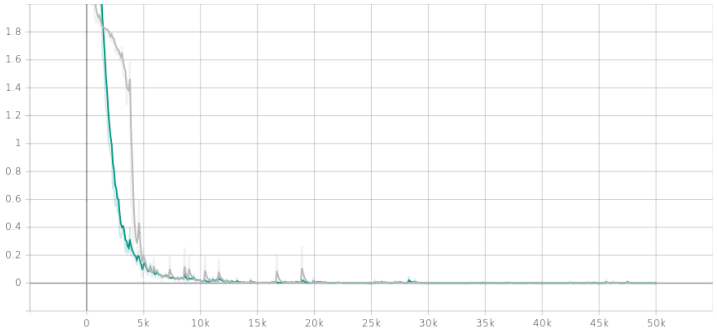
training wer (blue) and dev wer (red).



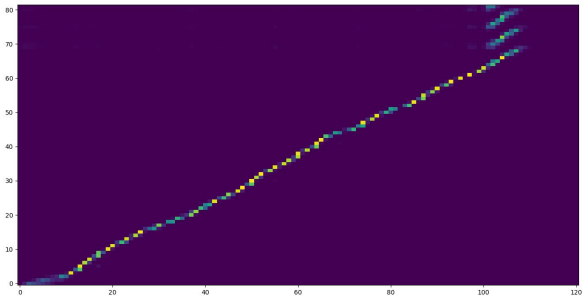
alignment.

dev - CER	2.0366
dev - WER	6.6903
kaggle score	1.26000

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



training loss (green: ctc, grey: att).



alignment.

dev - CER	1.7438
dev - WER	5.8834
kaggle score	1.07600

Joint CTC-attention ASR model收斂較快，CTC的訓練目標是將encoder的輸出直接decode成文字，因此該embedding相較於沒有joint訓練的model，更有利於seq2seq的decoder辨識。

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.

dev - CER	1.8159
dev - WER	6.2916
kaggle score	1.08200

Decoding with seq2seq decoder效果較好。CTC的decode只會考慮到當下時間點的feature vector，缺乏不同時間點間的資訊。而seq2seq在decode時會參考到過去時間點的輸出，進而影響到這個時間點的輸出。此外，seq2seq的decode使用到了attention的技術，能夠參考到所有feature vector，因此效果較好。

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.

Corpus: aishell、matbn200 (僅取70個字元的句子)、tcc300、thchs30，由以上四個文本整併起來作為external language model的訓練文本。
lm_weight: 0.5

dev - CER	1.9241
dev - WER	6.1633
kaggle score	1.05000

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?

	beam=2	beam=5	beam=10	beam=20
dev - CER	1.9241	1.9241	1.8842	1.8842
dev - WER	6.1633	6.1633	6.0353	6.0353
kaggle score	1.05000	1.05000	1.02600	1.02600

Bonus: (1%)

1. CTC decode with beam search and language model

reference:

1. https://towardsdatascience.com/beam-search-decoding-in-ctc-trained-neural-networks-5a889a3d85a7
2. http://placebokkk.github.io/asr/2020/02/01/asr-ctc-decoder.html

	greedy	beam=2, w/o lm	beam=2, w/ lm
dev - CER	1.8153	1.8159	1.5621
dev - WER	6.2836	6.2916	5.2335
kaggle score	-	1.08200	0.93400

2. RNN-T with beam search

使用Q2 joint CTC-attention model的encoder作為RNN-T encoder的pretrain，internal language model作為RNN-T decoer的pretrain。

reference:

1. RNN-T loss: https://github.com/HawkAaron/warp-transducer
2. Beam search: Alex Graves, Sequence Transduction with Recurrent Neural Networks, ICML workshop, 2012
3. Beam search: https://github.com/ZhengkunTian/rnn-transducer

	greedy	beam=5
dev - CER	1.5565	1.5481
dev - WER	5.2152	5.1940
kaggle score	0.93600	0.93800