An Empirical Study of Mini-Batch Creation Strategies for Neural Machine Translation

Makoto Morishita[1], Yusuke Oda[2], Graham Neubig[3,2], Koichiro Yoshino[2,4], Katsuhito Sudoh[2], Satoshi Nakamura[2]

[1]NTT Communication Science Laboratories, [2]Nara Institute of Science and Technology, [3]Carnegie Mellon University, [4]PRESTO Japan Science and Technology Agency

Innovative R&D by NTT

Abstract

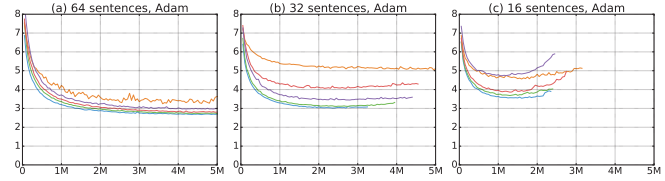
- There are a lot of different mini-batch creation strategies
- We did the experiments to survey the effect of mini-batch creation strategies
- The choice of a mini-batch creation strategy has a large effect on NMT training
- Our results suggest that we should:
  - Use larger mini-batch size
  - For Adam: sort the corpus based on the source sentence length or just shuffle before making mini-batches
  - For SGD: sort using the target sentence length, break ties by sorting by source sentence length

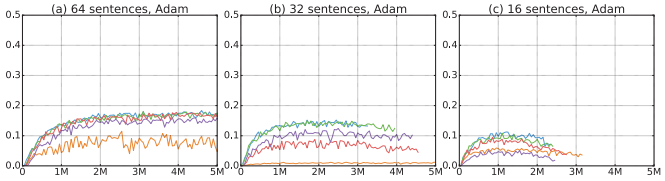
Effect of Mini-Batch Size

Q Are larger mini-batches better?

A Yes!

- Hypothesis
  - Larger mini-batches make the gradients more stable.
  - They also increase efficiency with parallel computation.
- Experimental Result
  - Mini-batch size can affect the final accuracy
  - The larger mini-batch size seems to be better.





Experimental Settings

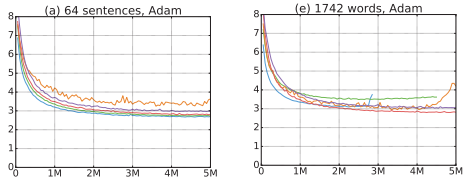
- 1 Layer LSTM with 512 Units
- Dropout 30% for all vertical connections
- Parameters are identically initialized between experiments
- Optimization function
  - Adam (α=0.001)
  - SGD (η=0.1)
- Corpus
  - ASPEC (English-Japanese)
  - WMT 2016 (English-German)
- Corpus Sorting Method
  - SHUFFLE: shuffle the corpus randomly before creating mini-batches, with no sorting
  - SRC: sort based on the source sentence length
  - TRG: sort based on the target sentence length
  - SRC_TRG: sort using the source sentence length, break ties by sorting by target sentence length
  - TRG_SRC: converse of SRC_TRG

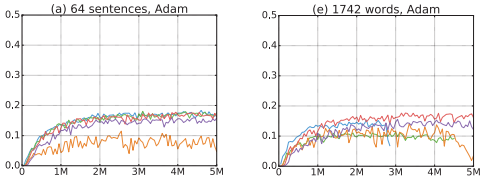
Effect of Mini-Batch Unit

Q Is there any differences by mini-batch units?

A No differences.

- Hypothesis
  - Most NMT toolkits create mini-batches with a constant number of sentences.
  - This leads to vary the scale of the losses since the loss function for the mini-batch is the sum of the word level losses.
  - Creating mini-batches by keeping the number of target words may lead to more stable convergence.
- Experimental Result
  - Mini-batch units do not effect to the training process





Effect of Corpus Sorting Method

Q Should we sort the corpus?

A Sometimes, not.

- Hypothesis
  - If the sentence length varies in the mini-batch, we need to pad the tokens to adjust their length to the length of the longest one.
  - Many NMT toolkits implement length-based sorting for reducing the amount of padding required.
- Experimental Result
  - When using Adam, the TRG and TRG_SRC sorting methods, do not always work well, use SHUFFLE or SRC sorting method.
  - When using SGD, use TRG_SRC (it process one iteration faster.)