Residual Stacking of RNNs for Neural Machine Translation

Raphael Shu
The University of Tokyo

Akiva Miura
Nara Institute of Science and Technology

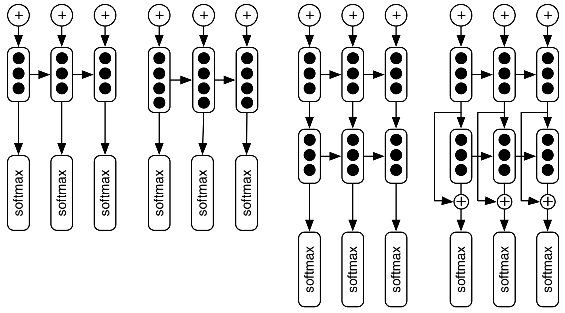
Overview (Abstract)

To enhance Neural Machine Translation models, several obvious ways such as enlarging the hidden size of recurrent layers and stacking multiple layers of RNN can be considered. Surprisingly, we observe that using naively stacked RNNs in the decoder slows down the training and leads to degradation in performance. In this paper, We demonstrate that applying residual connections in the depth of stacked RNNs can help the optimization, which is referred to as residual stacking. In empirical evaluation, residual stacking of decoder RNNs gives superior results compared to other methods of enhancing the model with a fixed parameter budget. Our submitted systems in WAT2016 are based on a NMT model ensemble with residual stacking in the decoder. To further improve the performance, we also attempt various methods of system combination in our experiments.

Evaluation results in English-Japanese task

	RIBES	BLEU
Baseline decoder	79.49	29.32
Enlarged decoder	79.60	30.24
Stacked decoder	79.25	29.07
Residual stacking of decoder RNNs	**79.88**	**30.75**

Experiments

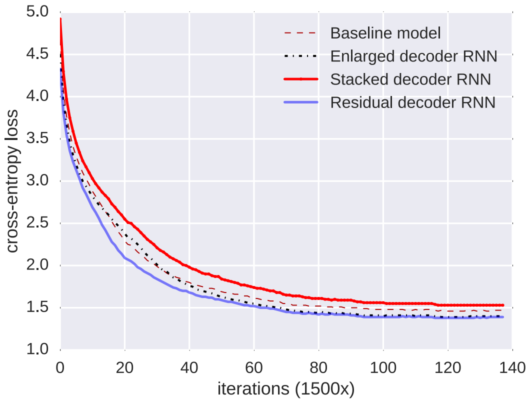


(a)
Baseline decoder

(b)
Enlarged decoder

(c)
Stacked decoder

(d)
Residual stacking



Residual stacking of decoder RNNs

In our experiment, we designed three kinds of decoders to enhance NMT models with almost same amount of extra parameters.

- Baseline decoder: single-layer LSTM with 1000 units
- Enlarged decoder: single-layer LSTM with 1400 units
- Stacked decoder: two-layer stacked LSTMs with 1000 units each
- Residual stacking: two-layer residual stacking of LSTMs with 1000 units each

For residual stacking, the second LSTM computes a residual of the first one. We found that residual stacking of decoder RNNs achieves the best performance among all three variations.

Stacking decoder RNNs naively hurts performance

Surprisingly, found the naively stacked decoder with two-layer LSTMs gives the worst performance. The training is significantly slowed down from the beginning. The final translation accuracy is even worse than baseline single-layer decoder.

Systems submitted in WAT2016

	RIBES	BLEU	HUMAN
Online A	71.52	19.81	49.57
Ensemble of 2 NMT models with residual stacking of decoder RNNs	81.72	33.38	30.50
+ System combination with T2S SMT	81.44	34.77	29.75

Acknowledgment

We thank Weblio Inc. for financial support of this research.