ViVi\_NMT: An Open Source Toolkit for

Neural Machine Translation

# Abstract

This paper introduces ViVi\_NMT in detail, a neural machine translation toolkit developed by Center for Speech and Language Technologies, Tsinghua University.On top of Tensorflow, ViVi\_NMT implements the attention-based encoder-decoder framework. Experiments on Chinese-English datasets and English-German dataset show that ViVi\_NMT has a good performance and achieves the exceptive goal.

# 1 Introduction

Neural Machine Translation (NMT) has been shown to have highly promising performance, particularly when a large amount of training data is available (Wu et al., 2016; Johnson et al., 2016; Shi et al., 2016; Mi et al., 2016). Although there are different model architectures (Sutskever et al.,2014; Bahdanau et al., 2015), the common principle behind the NMT approach is the same: encoding the meaning of the input into a concept space and performing translation based on this encoding. This ‘meaning encoding’ principle leads to a deeper understanding and learning of the translation rules, hence a better translation than conventional statistic machine translation (SMT) that considers only surface forms, i.e., words and phrases (Koehn et al., 2003).

This paper detailed introduces ViVi\_NMT, an open source toolkit for neural machine translation developed by Center for Speech and Language Technologies, Tsinghua University. On top of Tensorflow, this toolkit mainly reproduced the RNNsearch model (Bahdanau et al., 2015).We compare ViVi\_NMT with the statistical machine translation system Moses and THUMT(Zhang et al., 2002)，the experiments on Chinese-English datasets and English-German dataset show that ViVi\_NMT has a good performance.

# 2 ViVi\_NMT

On top of Tensorflow, ViVi\_NMT implements the attention-based neural machine translation model( Bahdanau et al., 2015）. Model details are clearly described in the paper( Bahdanau et al., 2015） ,please refer to it.

## 2.1 Training Criteria

To save model training time,ViVi\_NMT only uses maximum likelihood estimation as training criterion, which aims to find a set of model parameters that maximizes the likelihood of training data.

## 2.2 Optimization

Optimization is a key part in neural machine translation which directly affects the training time and translation performance of neural model. ViVi\_NMT uses Adam optimization (Kingma and Ba, 2015) which computes individual learning rate for different parameters.

# 3 Experiments

## 3.1 Data

The experiments were conducted for Chinese-English translation and English-German translation using there datasets, the relatively small IWSLT dataset, the much larger NIST dataset and the standard WMT 2014 English-German dataset.

**The IWSLT05 corpus** The training data consists of 44K sentences from the tourism and travel domain.The development set was composed of the ASR devset 1 and devset 2 from IWSLT 2005, and testing used the IWSLT 2005 test set.

**The NIST03 corpus** The training data consists of 1M sentence pairs with 19M source tokens and 24M target tokens from the LDC corpora of LDC2002E18, LDC2003E07, LDC2003E14, and Hansard’s portion of LDC2004T07, LDC2004T08 and LDC2005T06. We use the NIST 2005 test set as the development set and the NIST 2003 test set as the test set.

**The standard WMT 2014 English-German dataset** The training data consists of about 4.5 million sentence pairs from the Common Crawl corpus,new News Commentary corpus and version 7 of the Europarl corpus. The development set and the test set are from the WMT 2014 development sets and the WMT 2014 cleaned test sets respectively.

## 3.2 Systems and Settings

We use the conventional SMT system Moses and an open source neural machine translation toolkit THUMT as the baselines. The translation performance was evaluated using the case-insensitive BLEU score (Papineni et al., 2002).

For a fair comparison, the configurations of the models in the ViVi\_NMT system and the THUMT system were intentionally set to be identical.For all languages,we set the vocabulary size to 30K，configure the dimension of hidden layers and the word embedding dimension to 1000,620,respectively. In the training process, the batch size of the SGD algorithm was set to 80 and the step sizes of Adam optimizer was set to 0.0005.The decoding is implemented as a beam search,where the beam size was set to be 10.We train models on a single GPU device GeForce GTX 1070.

## 3.3 Experimental Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Toolkit | Criterion | Optimizer | IWSLT05 | NIST03 | WMT 2014 English-German |
| Moses |  |  | 53.99 | 28.18 | 15.3 |
| THUMT | MLE | Adam | 45.47 | OOM | OOM |
| ViVi\_NMT | MLE | Adam | 47.93 | 31.85 | 14.55 |

Table 1: Comparison between Moses,THUMT and ViVi\_NMT.

Table 1 shows the BLEU results with different systems. Firstly, it can be observed that with the small IWSLT05 dataset, the Moses performs much better than the other two NMT toolkit(this is unsurprising as neural models often need more training data),and ViVi\_NMT got comparable(even slightly better) performance than THUMT. Secondly, when the much larger NIST dataset and WMT 2014 dataset were used in training process,OOM(out of memory error) occurred from THUMT toolkit due to insufficient memory of GPU. So we can only compare the other two toolkits.As we can see in the table, ViVi\_NMT performs better than Moses with the much larger NIST dataset,and the translation performance of the two toolkits are approximately the same when WMT 2014 dataset were used.

# 4 Conclusion

In this paper,we have introduced ViVi\_NMT,an open source toolkit for neural machine translation. Experiments on Chinese-English datasets and English-German dataset show that ViVi\_NMT has a good performance and achieves the exceptive goal.The toolkit is freely available at <https://github.com/CSLT-THU/CSLT_NMT>.

# References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.2015. Neural machine translation by jointly learning to align and translate. In ICLR’2015.

Philipp Koehn, Franz Josef Och, and Daniel Marcu.2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics,pages 48–54.

Diederik Kingma and Jimmy Ba.2015.Adam:A method for stochastic optimization. In Proceedings of ICLR.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu:a method for automatic evaluation of machine translation.In Proceedings of ACL.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014.Sequence to sequence learning with neural networks.In Advances in neural information processing systems. pages 3104–3112.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. arXiv preprint arXiv:1605.03148.

Xing Shi, Kevin Knight, and Deniz Yuret. 2016. Why neural translations are the right length. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pages 2278–2282.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat,Fernanda Vi′egas, MartinWattenberg, Greg Corrado,et al. 2016. Google’s multilingual neural machine translation system: Enabling zero-shot translation.arXiv preprint arXiv:1611.04558.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey,Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 .

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, Yang Liu. 2017. THUMT: An Open Source Toolkit for Neural Machine Translation. arXiv:1706.06415.