Cost-Aware Learning Rate for Neural Machine Translation

Yang Zhao, Yining Wang, Jiajun Zhang, and Chengqing Zong(✉)

National Laboratory of Pattern Recognition, Institute of Automation, CAS University of Chinese Academy of Sciences, Beijing, China
zhaoyang2015@ia.ac.cn
{yining.wang,jjzhang,cqzong}@nlpr.ia.ac.cn

Abstract. Neural Machine Translation (NMT) has drawn much attention due to its promising translation performance in recent years. The conventional optimization algorithm for NMT sets a unified learning rate for each gold target word during training. However, words under different probability distributions should be handled differently. Thus, we propose a cost-aware learning rate method, which can produce different learning rates for words with different costs. Specifically, for the gold word which ranks very low or has a big probability gap with the best candidate, the method can produce a larger learning rate and vice versa. The extensive experiments demonstrate the effectiveness of our proposed method.

Keywords: Neural machine translation · Cost-aware learning rate

1 Introduction

Neural Machine Translation (NMT) based on the encoder-decoder architecture proposed by [4,10] can achieve promising translation performance for several language pairs, such as English-to-German and English-to-French [1,14,23,25].

In general, we can train a NMT system by using maximum likelihood estimation with stochastic gradient descent and back propagation through time. However, this kind of optimization algorithm has a drawback that it sets a unified learning rate for each gold target word during training.

Actually, for a parallel sentence pair, gold target words have different prediction probability distributions (costs). Figure 1 shows an example. Given the gold target sentence $\{y_1, y_2, y_3, y_4\}$ during training, each gold target word has a different probability distribution. y_1 ranks first and this is the ideal case. y_2 ranks second and the gap with the best candidate is quite small (0.01), making y_2 only need a small boost. In contrast, although there exists a small gap (0.05) between y_3 and the top one, its ranking is relatively low (lower than the beam size in decoding). The ranking of y_4 is high (second) while there exists a huge gap (0.85) with the top one. Intuitively, y_3 and y_4 need a big boost to increase the ranking or reduce the gap.

We believe that it is more reasonable to assign different learning rates to words under different costs. Therefore, we propose a cost-aware learning rate

M. Sun et al. (Eds.): CCL 2017 and NLP-NABD 2017, LNAI 10565, pp. 85–93, 2017.
https://doi.org/10.1007/978-3-319-69005-6_8

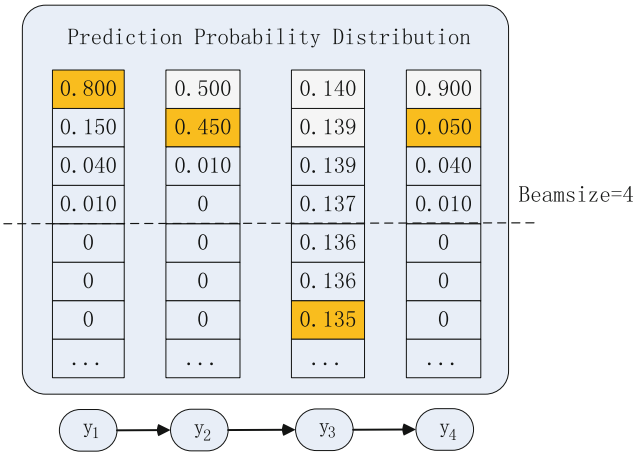

Fig. 1. For the target sentence y_1, y_2, y_3, y_4 in a parallel sentence pair, each gold target word (highlighted in yellow) has a different prediction probability distribution, and ideally dynamic learning rate is needed. In this example, we assume the beam size in decoding is set to 4. (Color figure online)

method, which can produce dynamic learning rates according to the probability distribution of the gold target words. More specifically, for the gold words which have a low probability ranking (lower than the beam size which is set in decoding) or have a big probability gap with the top candidate, the method will produce a larger learning rate and vice versa.

In this paper, we make the following contributions:

(1) To best of our knowledge, this is the first effort to propose a cost-aware learning rate to improve the training procedure of neural machine translation. According to the prediction probability distributions, we design different strategy to produce dynamic learning rates.
(2) Our empirical experiments on Chinese-English translation tasks show the efficacy of our methods and we can obtain an average improvement of 0.87 BLEU score on multiple evaluation datasets.

2 Neural Machine Translation

The goal of machine translation is to transform a sequence of source words $X = \{x_1, x_2, ..., x_{Tx}\}$ into a sequence of target words $Y = \{y_1, y_2, ..., y_{Ty}\}$. The NMT contains two parts, encoder and decoder. As the name suggests, encoder transforms the source sentence X into context vectors C. And decoder generates target translation Y from the context vectors C by maximizing the probability of $p(y_i|y_{<i}, C)$. And Fig. 2 shows the framework of the attention-based NMT proposed by Luong et al. [14], which utilizes stacked Long Short Term Memory (LSTM) [29] layers for both encoder and decoder.

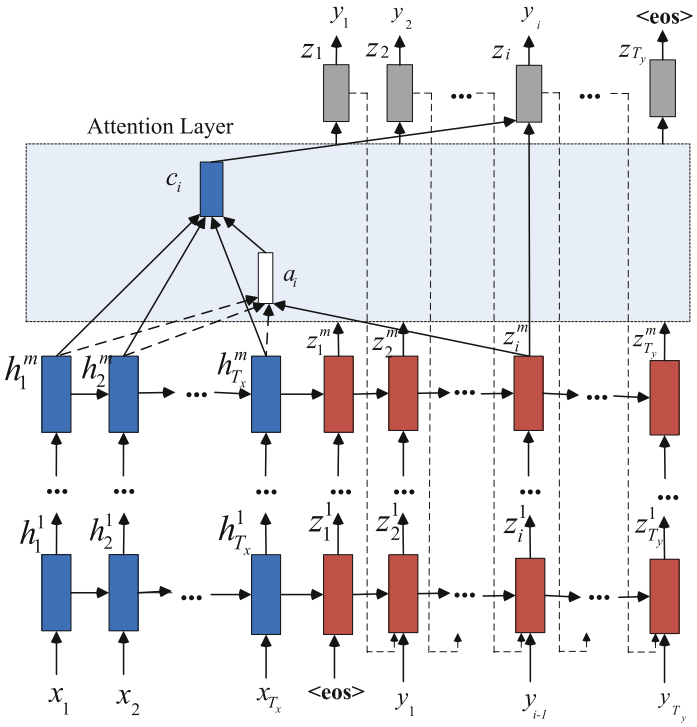

Fig. 2. The encoder-decoder framework for NMT.

Given the sentence aligned bilingual training data, the cost functions can be defined as the following conditional log-likelihood[1]:

$$L(\theta, D) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i}^{T_y} log(p(y_i^{(n)} | p(y_{<i}^{(n)}, X^{(n)}, \theta)) \tag{1}$$

Then, we can use maximum likelihood estimation with stochastic gradient descent and back propagation through time to get the optimal parameters as follows:

$$\theta \leftarrow \theta + \eta * \bigtriangledown L(\theta, D) \tag{2}$$

where η is the learning rate, $\bigtriangledown L(\theta, D)$ is the gradient direction, and can be calculated as the sum of the gradients of the sentences in minibatch B:

$$\bigtriangledown L(\theta, D) = \sum_{n=1}^{B} \bigtriangledown L(\theta, (X^{(n)}, Y^{(n)})) \tag{3}$$

[1] Recently, evaluation metric oriented cost functions are investigated Shen et al. [21] and Wu et al. [25] and the cost-aware learning rate can also be applied. In this paper, we use log-likelihood costs as a case study.

The gradients of each sentences can be calculated as a sum of gradients per-step:

$$\bigtriangledown L(\theta, (X^{(n)}, Y^{(n)})) = \sum_{i=1}^{T_y} \bigtriangledown log(p(y_i^{(n)}|y_{<i}^{(n)}, X^{(n)}, \theta)) \tag{4}$$

From Eqs. (2)–(4), the final parameter optimization method can be defined as follows:

$$\theta \leftarrow \theta + \eta * \sum_{n=1}^{B} \sum_{i=1}^{T_y} \bigtriangledown log(p(y_i^{(n)}|y_{<i}^{(n)}, X^{(n)}, \theta)) \tag{5}$$

It is easy to see from Eq. (5) that the current optimization algorithm sets a unified learning rate η for each step gradient $\bigtriangledown log(p(y_i^{(n)}|y_{<i}^{(n)}, X^{(n)}, \theta))$.

3 Cost-Aware Learning Rate

In Sect. 2 we described the current optimization algorithm (Eq. (5)), which sets a unified learning rate for each gold target word during training. In fact, the probability distributions vary dramatically for gold target words in different training steps. Ideally, a gold target word should be penalized much more if it ranks very low or has a big gap with the best candidate. Accordingly, dynamic learning rate is needed. To achieve this goal, we design cost-aware learning rate as follows:

Step1: For each gold target word y_i in Y, we can get the ranking (denoted as r) of y_i based on the prediction probability distribution $p(V_t|y_{<i}, X, \theta)$, where V_t is all the target candidates.

Step2: We can also get the probability gap between y_i and the word with the maximum probability as follows:

$$g = max(p(V_t|y_{<i}, X, \theta)) - p(y_i|y_{<i}, X, \theta) \tag{6}$$

Step3: Then, we can calculate the cost aware learning rate λ_i for y_i as follows:

$$\begin{aligned} \lambda_i &= \alpha * f(r) + \beta * g + \gamma \\ f(r) &= \begin{cases} 1 & if\ r > b \\ 0 & if\ r \leq b \end{cases} \end{aligned} \tag{7}$$

where r is derived from Step 1, g is calculated as Eq. (6), b is the beam size which is set in decoding, α, β and γ are hyper parameters that can be used to adjust the respective weights. As shown in Eq. (7), our method has two functions:

(1) It can produce a larger learning rate for the word whose ranking is lower than the beam size. Here we explain the reason why we design our method like this. During decoding, we will use beam search to get the best target sentence. To make it possible, we should first guarantee that each gold target word could rank before beam size during training. Therefore, the algorithm will set a larger learning rate for the word whose ranking is lower than the beam size to boost its ranking.
(2) It can also produce a larger learning rate for the words which have a big probability gap with the top one. The process of involving g is important because we want to reduce the gap between the gold target words and the candidates with the maximum prediction probability.

After getting the cost-aware learning rate λ_i for y_i, our final parameter optimization method, extended from Eq. (5), can be described as follows:

$$\theta \leftarrow \theta + \eta * \sum_{n=1}^{B} \sum_{i=1}^{T_y} \lambda_i * \bigtriangledown log(p(y_i^{(n)} | y_{<i}^{(n)}, X^{(n)}, \theta)) \tag{8}$$

where λ_i is calculated as Eq. (7). We retain the unified learning rate η. When $\alpha = 0$, $\beta = 0$ and $\gamma = 1$ (Eq. (7)), our method falls back to the original optimization method.

4 Experimental Settings

4.1 Dataset

We test the proposed methods on Chinese-to-English with two data sets: (1) small data set, which includes 0.63 M[2] sentence pairs; (2) large-scale data set, which contains about 2.1 M sentence pairs. For validation, we choose NIST 2003 (MT03) dataset. For testing, we use NIST2004 (MT04), NIST 2005 (MT05), NIST 2006 (MT06) and NIST 2008 (MT08) datasets.

4.2 Training and Evaluation Details

We use the Zoph_RNN toolkit[3] to implement our described methods. The encoder and decoder include two stacked LSTM layers. The word embedding dimension and the size of hidden layers are all set to 1,000. We use a minibatch size of $B = 128$. We limit the vocabulary to 30 K most frequent words for both the source and target languages. Other words are replaced by a special symbol UNK. At test time, we employ beam search with beam size $b = 12$. We use case-insensitive 4-gram BLEU score as the automatic metric [18] for translation quality evaluation.

[2] LDC2000T50, LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2003E14, LDC2003T17, LDC2004T07.
[3] https://github.com/isi-nlp/Zoph_RNN. We extend this toolkit with global attention.

4.3 Translation Methods

In the experiments, we compare our method with the conventional Statical Machine Translation (SMT) model and the baseline NMT model trained with the unified learning rate. We list all the translation methods as follows:

(1) **Moses:** It is the state-of-the-art phrase-based SMT system [11]. Our system is built using the default settings.
(2) **U_lr:** It is the baseline attention-based NMT system [14,28] trained with the unified learning rate. The initial unified learning rate η is set to 0.1, and the learning rate decay for η is set to 0.5.
(3) **C_lr:** It is similar to **U_lr** except that it is trained by our cost-aware learning rate method. The hyper parameters in Eq. (7) are respectively set to 0.2 (α), 0.8 (β) and 1 (γ), which are tuned on the validation dataset.

5 Translation Results

Table 1 reports the detailed translation results for different methods. Comparing the **Moses** and **U_lr**, it is very obvious that the attention-based NMT system **U_lr** substantially outperforms the phrase-based SMT system **Moses** on both small and large data, where average improvement on small data is up to 3.99 BLEU points (32.71 vs. 28.72) and on large data is 1.85 (36.74 vs. 34.89).

Table 1. Translation results (BLEU) for different methods on small and large-scale data. "*" indicates that it is statistically significant better ($p < 0.05$) than "**U_lr**" and "†" indicates $p < 0.01$.

Method	MT03	MT04	MT05	MT06	MT08	Ave
Moses(small)	28.35	30.02	29.10	32.92	23.20	28.72
U_lr(small)	34.20	36.96	32.60	33.85	25.96	32.71
C_lr(small)	35.01*	37.74*	33.71†	34.93†	26.51*	33.58
Moses(large)	38.54	39.01	36.55	35.59	24.76	34.89
U_lr(large)	39.07	40.49	37.26	38.04	28.83	36.74
C_lr(large)	39.73*	41.06*	38.24*	38.88*	29.49*	37.48

Compared to **U_lr** (row 2 in Table 1), our method (**C_lr**) improves the translation quality on all test sets and the average improvement is up to 0.87 BLEU points (33.58 vs. 32.71). It indicates that our cost-aware learning rate method can learn better network parameters for neural machine translation.

As our method tries to make more gold target words rank before beam size and reduce the probability gap between the gold word and the best candidate with the maximum prediction probability, we calculate in the training data the proportion of the gold target words which rank in top beam size and the average

Table 2. The proportion of gold target words whose rankings lie in the top beam size and the average gap between the gold target words and the best candidate.

Method	Ranking in top beam size	Gap
U_lr	92.73%	0.067
C_lr	95.92%	0.046

gap between the gold target word and the top one. As shown in Table 2, our method can indeed boost the rankings of the gold target words and narrow the gap with the best candidate.

Besides that, we conduct another experiment to find out whether or not is our method still very effective when we have much more bilingual data. As shown by the last two rows in Table 1, our model can also improve the NMT translation quality on all of the test sets and the average improvement can be up to 0.74 BLEU points (37.48 vs. 36.74).

6 Related Work

In order to get better parameters for NMT, most of the existing works mainly focus on using more monolingual data [3,7,20,27] or adding additional prior knowledge besides bilingual data [5,17,24,26], and designing better attentional mechanisms [2,6,13–16].

Our work attempts to improve the network parameter tuning for neural machine translation when the log-likelihood objective function is employed. There are two closely related studies: one resorts to redesign the loss functions and the other tries to optimize the beam search algorithm.

Shen et al. [21] applies the minimum risk training for NMT and achieves a significant improvement. Sam and Alexander [19] proposes a model using beam search training scheme to get sequence-level scores.

Several researchers improved the beam search method in decoding. He et al. [8] and Stahlberg et al. [22] rerank target word candidates with additional features. Li and Jurafsky [12] rescore the translation candidates on sentence-level by using the mutual information between target and source sides. Hoang et al. [9] converts the decoding from a discrete optimization problem to a continuous optimization problem.

The significant difference between our work and these studies lies in that our work focuses on improving the NMT training procedure from another perspective. We design a cost-aware learning rate method and set different learning rates for the words with different costs.

7 Conclusions and Future Work

In order to improve the current NMT optimization algorithm that used a unified learning rate for each gold target word during the whole training procedure, we

proposed a cost-aware learning rate method, which aims at producing different learning rates for the gold target words under different probability distributions. The extensive experiments show that our method can achieve statistical significantly improvements on translation quality.

In the future, we plan to design more effective methods to calculate more appropriate learning rates for NMT training.

Acknowledgments. The research work has been supported by the Natural Science Foundation of China under Grant No. 61403379 and No. 61402478.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of ICLR 2015 (2015)
2. Cheng, Y., Shen, S., He, Z., He, W., Wu, H., Sun, M., Liu, Y.: Agreement-based joint training for bidirectional attention-based neural machine translation. In: Proceedings of IJCAI 2016 (2016)
3. Cheng, Y., Wei, X., He, Z., He, W., Hua, W., Sun, M., Liu, Y.: Semi-supervised learning for neural machine translation. In: Proceedings of ACL 2016, pp. 1965–1974 (2016)
4. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of EMNLP 2014, pp. 1724–1734 (2014)
5. Cohn, T., Vu Hoang, C.D., Vymolova, E., Yao, K., Dyer, C., Haffari, G.: Incorporating structural alignment biases into an attentional neural translation model. In: Proceedings of NAACL 2016, pp. 876–885 (2016)
6. Feng, S., Liu, S., Li, M., Zhou, M.: Implicit distortion and fertility models for attentionbased encoder-decoder NMT model. arXiv preprint arXiv:1601.03317 (2016)
7. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., Ma, W.: Dual learning for machine translation. In: Proceedings of NIPS 2016 (2016)
8. He, W., He, Z., Hua, W., Wang, H.: Improved neural machine translation with Smt features. In: Proceedings of AAAI 2016, pp. 151–157 (2016)
9. Vu Hoang, C.D., Haffari, G., Cohn, T.: Decoding as continuous optimization in neural machine translation. arXiv preprint arXiv:1701.02854 (2017)
10. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of EMNLP 2013, pp. 1700–1709 (2013)
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL 2007, pp. 177–180 (2007)
12. Li, J., Jurafsky, D.: Mutual information and diverse decoding improve neural machine translation. arXiv preprint arXiv:1601.00372 (2016)
13. Liu, L., Utiyama, M., Finch, A., Sumita, E.: Neural machine translation with supervised attention. In: Proceedings of COLING 2016, pp. 3093–3102 (2016)
14. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention based neural machine translation. In: Proceedings of EMNLP 2015, pp. 1412–1421 (2015)
15. Meng, F., Zhengdong, L., Li, H., Liu, Q.: Interactive attention for neural machine translation. In: Proceedings of COLING 2016, pp. 2174–2185 (2016)

16. Mi, H., Sankaran, B., Wang, Z., Ittycheriah, A.: A coverage embedding model for neural machine translation. In: Proceedings of EMNLP 2016, pp. 955–960 (2016)
17. Mi, H., Wang, Z., Ittycheriah, A.: Supervised attentions for neural machine translation. In: Proceedings of EMNLP 2016, pp. 2283–2288 (2016)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL 2002, pp. 311–318 (2002)
19. Wiseman, S., Rush, A.M.: Sequence-to-sequence learning as beam-search optimization. In: Proceedings of EMNLP 2016, pp. 1296–1306 (2016)
20. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of ACL 2016, pp. 86–96 (2016)
21. Shen, S., Cheng, Y., He, Z., He, W., Hua, W., Sun, M., Liu, Y.: Minimum risk training for neural machine translation. In: Proceedings of ACL 2015, pp. 1683–1692 (2015)
22. Stahlberg, F., Hasler, E., Waite, A., Byrne, B.: Syntactically guided neural machine translation. arXiv preprint arXiv:1605.04569 (2016)
23. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of NIPS 2014, pp. 3104–3112 (2014)
24. Tang, Y., Meng, F., Lu, Z., Li, H., Yu, P.L.H.: Neural machine translation with external phrase memory. arXiv preprint arXiv:1606.01792 (2016)
25. Yonghui, W., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
26. Zhang, J., Zong, C.: Bridging neural machine translation and bilingual dictionaries. arXiv preprint arXiv:1610.07272 (2016)
27. Zhang, J., Zong, C.: Exploiting source-side monolingual data in neural machine translation. In: Proceedings of EMNLP 2016, pp. 1535–1545 (2016)
28. Zoph, B., Knight, K.: Multi-source neural translation. In: Proceedings of NAACL 2016, pp. 30–34 (2016)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)