Optimizing Deep Transformers for Chinese-Thai Low-Resource Translation

Wenjie Hao[1](✉), Hongfei Xu[1], Lingling Mu[1], and Hongying Zan[1,2]

[1] Zhengzhou University, Henan 450001, China
haowj9977@163.com, hfxunlp@foxmail.com, {iellmu,iehyzan}@zzu.edu.cn
[2] Peng Cheng Laboratory, Shenzhen 518000, China

Abstract. In this paper, we study the use of deep Transformer translation model for the CCMT 2022 Chinese↔Thai low-resource machine translation task. We first explore the experiment settings (including the number of BPE merge operations, dropout probability, embedding size, etc.) for the low-resource scenario with the 6-layer Transformer. Considering that increasing the number of layers also increases the regularization on new model parameters (dropout modules are also introduced when using more layers), we adopt the highest performance setting but increase the depth of the Transformer to 24 layers to obtain improved translation quality. Our work obtains the SOTA performance in the Chinese-to-Thai translation in the constrained evaluation.

Keywords: Low-resource NMT · Deep transformer · Chinese-Thai MT

1 Introduction

Neural machine translation (NMT) has achieved impressive performance with the support of large amounts of parallel data [1,27]. However, in low-resource scenario, its performance is far from expectation [10,12].

To improve the translation performance, previous work either study data augmentation approaches to leverage pseudo data [5,6,16,22,29] or benefit from models pre-trained on large-scale monolingual corpus [18,21].

Instead of introducing more data, in this paper, we explore the effects of different data processing and model settings for the CCMT 2022 Chinese↔Thai low-resource machine translation task inspired by Sennrich and Zhang [24].

Specifically, we adopt the Chinese↔Thai (Zh↔Th) machine translation data from CCMT 2022 of 200k training sentence pairs. We first apply strict rules for data cleaning, and employ the cutting-edge Transformer model [27]. We explore the influence of BPE merge operations on performance, and the effects of different model settings (embedding size, dropout probability). As previous work [2,8,13–15,17,28,30,31,33,36–38] shows that deep Transformers can bring about

T. Xiao and J. Pino (Eds.): CCMT 2022, CCIS 1671, pp. 117–126, 2022.
https://doi.org/10.1007/978-981-19-7960-6_12

improved translation performance, we adopt the setting of highest performance in ablation experiments for the Chinese↔Thai translation task but increase the number of layers to 24. We explore experiment settings with the 6-layer setting but adopt the best one to deeper models, because: 1) exploring the effects of these hyper-parameters with shallow models is more computation-friendly than with deep models, and 2) increasing the number of layers also introduces regularization, as adding new layers also brings dropout modules.

2 Background

2.1 Transformer

Vaswani et al. [27] propose the self-attention based Transformer model, evading the parallelization issue of RNN. Transformer has become the most popular model in NMT field. Transformer consists of one encoder and one decoder module, each of them is formed by several layers, and the multi-layer structure allows it to model complicated functions. The Transformer model also employs residual connection and layer normalization techniques for the purpose of ease optimization.

2.2 Low-Resource NMT

Despite that NMT has achieved impressive performance in high-resource cases [27], its performance drops heavily in low-resource scenarios, even underperforming phrase-based statistical machine translation (PBSMT) [10,12]. NMT normally requires large amounts of auxiliary data to achieve competitive results. Sennrich and Zhang [24] show that this is due to the lack of system adaptation for low-resource settings. They suggest that large vocabularies lead to low-frequency (sub)words, and the amount of data is not sufficient to learn high-quality high-dimensional representations for these low-frequency tokens. Reducing the vocabulary size (14k→2k symbols) can bring significant improvements (7.20→12.10 BLEU). In addition, they show that aggressive (word) dropout (0.1→0.3) can bring impressive performance (13.03→15.84 BLEU), and reducing batch size (4k→1k tokens) may also benefit. Optimized NMT systems can indeed outperform PBSMT.

2.3 Parameter Initialization for Deep Transformers

Xu et al. [36] suggest that the training issue of deep Transformers is because that the layer normalization may shrink the residual connections, leading to the gradient vanishing issue. They propose to address this by applying the Lipschitz constraint to parameter initialization. Experiments on WMT14 English-German and WMT15 Czech-English translation tasks show the effectiveness of their simple approach.

2.4 Deep Transformers for Low-Resource Tasks

Previous work shows that deep Transformers generally perform well with sufficient training data [13], and few attempts have been made on training deep Transformers from scratch on small datasets. Xu et al. [37] propose Data-dependent Transformer Fixed-update initialization scheme, called DT-Fixup, and experiment on the Text-to-SQL semantic parsing and the logical reading comprehension tasks. They show that deep Transformers can work better than their shallow counterparts on small datasets through proper initialization and optimization procedure. Their work inspires us to explore the use of deep Transformers for low-resource machine translation.

3 Our Work

3.1 Data Processing

The quality of the dataset affects the performance of NMT. Therefore, We first standardize the texts with the following pipeline:

1. removing sentences with encoding errors;
2. converting Traditional Chinese to Simplified Chinese through OpenCC;[1]
3. replacing full width characters with their corresponding half width characters;
4. converting all named and numeric character HTML references (e.g., &gt;, &#62;, &#x3e) to the corresponding Unicode characters.

For the training of NMT models, we segment Chinese sentences into words using jieba.[2]

We perform independent Byte Pair Encoding (BPE) [23] for Thai and Chinese corpus to address the unknown word issue with the SentencePiece toolkit [11].[3]

As the evaluation does not release the test set, we hold out the last 1000 sentence pairs of the training set for validation.

3.2 Exploration of Training Settings

We explore the influence of different training settings on the low-resource translation task in two aspects:

1. Vocabulary sizes;
2. Model hyper-parameters (embedding size and dropout probabilities).

For our experiment, we employ the Transformer translation model [27] for NMT, as it has achieved the state-of-the-art performance in MT evaluations [1] and conduct our experiment based on the Neutron toolkit [35] system. Neutron is an open source Transformer [27] implementation of the Transformer and its variants based on PyTorch.

[1] https://github.com/BYVoid/OpenCC.

[2] https://github.com/fxsjy/jieba.

[3] https://github.com/google/sentencepiece.

Table 1. Results (BLEU) on CCMT 2022 Th→Zh translation task with different vocabulary sizes.

Merge operations	6k	8k	16k	24k
Thai vocabulary size	5,996	7,997	16,000	23,999
Chinese vocabulary size	5,989	7,984	15,943	23,881
BLEU	27.07	25.70	**29.90**	28.87

Exploration of Vocabulary Sizes. Previous work shows that the effect of vocabulary size on translation quality is relatively small for high-resource settings [7]. While for low-resource settings, reduced vocabulary size (14k→2k) may benefit translation quality [24]. BPE [23] is a popular choice for open-vocabulary translation, which has one hyper-parameter, the number of merge operations, that determines the final vocabulary size. Following Sennrich and Zhang [24], we explore the influence of different vocabulary sizes for the Thai→Chinese translation task.

We train 4 NMT models in Thai→Chinese translation direction with different number of BPE merge operations, and the statistics of resulted vocabularies are shown in Table 1. Specifically, we perform independent BPE [23] for Thai and Chinese corpus with 4k/8k/16k/24k merge operations by SentencePiece [11].

For model settings, we adopted the Transformer with 6 encoder and decoder layers, 256 as the embedding dimension and 4 times of embedding dimension as the number of hidden units of the feed-forward layer, a dropout probability of 0.1. We used relative position [25] with a clipping distance k of 16. The number of warm-up steps was set to $8k$. We used a batch size of around $25k$ target tokens achieved by gradient accumulation, and trained the models for 128 epochs.

For evaluation, we decode with a beam size of 4 with average of the last 5 checkpoints saved in an interval of $1,500$ training steps. We evaluate the translation quality by character BLEU with the SacreBLEU toolkit [20]. Results are shown in Table 1.

Table 1 shows that: 1) in general, the use of more merge operations (16k/24k) is better than fewer ones (6k/8k), and 2) the setting of 16k merge operations leads to the best performance for the Thai→Chinese translation task, achieving 29.90 BLEU points.

Exploration of Hyper-parameter Settings of Model. Hyper-parameters are often re-used across experiments. However, best practices may differ between high-resource and low-resource settings. While the trend in high-resource settings is using large and deep models, Nguyen and Chiang [19] use small models with fewer layers for small datasets, and Sennrich and Zhang [24] show that aggressive dropout is better for low-resource translation. In this paper, we also explore the effects of model sizes (embedding dimension and hidden dimension) and dropout probabilities on the performance.

Table 2. Results (BLEU) on CCMT 2022 Zh→Th translation task with different model settings.

Settings		A	B	C	D	E
Embbeding size	256	√				
	384		√	√		
	512				√	√
Dropout	0.1	√	√		√	
	0.3			√		√
BLEU		6.35	15.02	5.30	**24.42**	7.73

We train 5 NMT models in Chinese→Thai translation direction with different training settings as shown in Table 2. We set the number of BPE merge operations to 16k based on Table 1.

We experimented the Transformers with 6 encoder and decoder layers, 256/384/512 as the embedding dimension and 4 times of embedding dimension as the number of hidden units of the feed-forward layer, dropout probabilities of 0.1 or 0.3. We used relative position [25] with a clipping distance k was 16 and GeLU as the activation function. The number of warm-up steps was set to $8k$. We used a batch size of around $25k$ target tokens achieved by gradient accumulation, and trained the models for 128 epochs.

For evaluation, we decode with a beam size of 4, and evaluate the translation quality with the SacreBLEU toolkit [20] with the average of the last 5 checkpoints saved in an interval of $1,500$ training steps. Results are shown in Table 2.

Table 2 shows that: 1) large embedding dimension is beneficial to translation performance, 2) aggressive dropout (0.3 in this paper) does not benefit the task, and 3) Setting D with 512 as the embedding dimension and 0.1 as the dropout probability is the best option, achieving a BLEU score of 24.42 in the Chinese→Thai translation task.

3.3 Deep Transformers for Low-Resource Machine Translation

To obtain good translation quality, we adopt the setting D, but use 24 encoder and decoder layers for better performance [2,8,13–15,17,28,30,31,33,36–38]. Parameters were initialized under the Lipschitz constraint [36] to ensure the convergence. We used the dynamical batch size strategy which dynamically determines proper and efficient batch sizes during training [34].

We use the best experiment setting explored with the 6-layer models for deeper models, because: 1) training shallow models are much faster than deep models, and 2) adding new layers also introduces regularization, as dropout modules are also introduced with these layers.

We train two models on the whole training set, which takes about 75 h to train one model on a nvidia RTX3090 GPU. We averaged the last 20 checkpoints saved with an interval of $1,500$ training steps.

Table 3. Results on the CCMT 2022 Zh↔Th test set. The computation of BLEU scores for the test set are different from that for Tables 1 and 2.

	Th→Zh	Zh→Th
BLEU4	/	9.06
BLEU5	4.85	/

We decode the CCMT 2022 Zh↔Th test set consisting of 10k sentences for each direction with a beam size of 4. Results are shown in Table 3.

Table 3 shows that the CCMT 2022 Chinese-Thai low-resource translation task is still a quite challenging task and there is a quite large space for improvements. But to date, our study establishes the SOTA performance in the Chinese-to-Thai translation in the constrained evaluation.

4 Related Work

As data scarcity is the main problem of low-resource machine translation, making most of the existing data is a popular research direction to address this issue in previous work. There are two specific types: 1) data augmentation, and 2) using pre-trained language models.

Data augmentation is to add training data, normally through modifications of existing data or the generation of new pseudo data. In machine translation, typical data enhancement methods include back-translating external monolingual data [5,22], obtaining pseudo bilingual data by modifying original bilingual data, such as adding noise to training data [6,29] or by paraphrasing which takes into the diversity of natural language expression into account [16], and mining of bilingual sentence pairs from comparable corpus [32] (comparable corpus is a text that is not fully translated from the source language to the target language but contains with rich knowledge of bilingual contrast).

For the use of pre-trained language models in NMT, leveraging the target-side language model is the most straightforward way to use monolingual data [26]. Other work [18,21] directly uses word embeddings pre-trained on monolingual data to initialize the word embedding matrix of NMT models. More recently, some studies leverage pre-trained models to initialize the model parameters of the encoder of NMT [3,4,9].

Fore-mentioned studies require large amounts of auxiliary data. Low-resource NMT without auxiliary data has received comparably less attention [19,24]. In this work, we revisit this point with deep Transformers, and focus on techniques to adapt deep Transformers to make most of low-resource parallel training data, exploring the vocabulary sizes and model settings for NMT.

5 Conclusion

In this paper, we explore the influence of different settings for the use of deep Transformers on the CCMT 2022 Zh↔Th low-resource translation task.

We first test the effects of the number of BPE merge operations, embedding dimension and dropout probabilities with 6-layer models, then adapt the best setting to the 24-layer model, under the motivation that: 1) shallow models are fast to train, and 2) increasing the number of layers also introduces regularization for these added layers.

Acknowledgments. We thank anonymous reviewers for their insightful comments. We acknowledge the support of the National Social Science Fund of China (Grant No. 17ZDA138 and Grant No. 14BYY096).

References

1. Akhbardeh, F., et al.: Findings of the 2021 conference on machine translation (WMT21). In: Proceedings of the Sixth Conference on Machine Translation, pp. 1–88. Association for Computational Linguistics (2021). https://aclanthology.org/2021.wmt-1.1
2. Bapna, A., Chen, M., Firat, O., Cao, Y., Wu, Y.: Training deeper neural machine translation models with transparent attention. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3028–3033. Association for Computational Linguistics (2018). https://aclweb.org/anthology/D18-1338
3. Clinchant, S., Jung, K.W., Nikoulina, V.: On the use of BERT for neural machine translation. In: Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, pp. 108–117. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-5611. https://aclanthology.org/D19-5611
4. Edunov, S., Baevski, A., Auli, M.: Pre-trained language model representations for language generation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, vol. 1 (Long and Short Papers), pp. 4052–4059. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1409. https://aclanthology.org/N19-1409
5. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 489–500. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1045. https://aclanthology.org/D18-1045
6. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada (vol. 2: Short Papers), pp. 567–573. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-2090. https://aclanthology.org/P17-2090
7. Haddow, B., et al.: The University of Edinburgh's submissions to the WMT18 news translation task. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, pp. 399–409. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-6412. https://aclanthology.org/W18-6412
8. Huang, X.S., Perez, F., Ba, J., Volkovs, M.: Improving transformer optimization through better initialization. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning

Research, vol. 119, pp. 4475–4483. PMLR (2020). https://proceedings.mlr.press/v119/huang20f.html

9. Imamura, K., Sumita, E.: Recycling a pre-trained BERT encoder for neural machine translation. In: Proceedings of the 3rd Workshop on Neural Generation and Translation, Hong Kong, pp. 23–31. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-5603. https://aclanthology.org/D19-5603
10. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, pp. 28–39. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/W17-3204. https://aclanthology.org/W17-3204
11. Kudo, T., Richardson, J.: SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, pp. 66–71. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-2012. https://aclanthology.org/D18-2012
12. Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M.: Phrase-based & neural unsupervised machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 5039–5049. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1549. https://aclanthology.org/D18-1549
13. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. CoRR abs/1909.11942 (2019). https://arxiv.org/abs/1909.11942
14. Li, B., et al.: Learning light-weight translation models from deep transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 15, pp. 13217–13225 (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17561
15. Li, B., et al.: Shallow-to-deep training for neural machine translation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 995–1005. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.72. https://aclanthology.org/2020.emnlp-main.72
16. Mallinson, J., Sennrich, R., Lapata, M.: Paraphrasing revisited with neural machine translation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, vol. 1, Long Papers, pp. 881–893. Association for Computational Linguistics (2017). https://aclanthology.org/E17-1083
17. Mehta, S., Ghazvininejad, M., Iyer, S., Zettlemoyer, L., Hajishirzi, H.: Delight: deep and light-weight transformer. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=ujmgfuxSLrO
18. Neishi, M., Sakuma, J., Tohda, S., Ishiwatari, S., Yoshinaga, N., Toyoda, M.: A bag of useful tricks for practical neural machine translation: embedding layer initialization and large batch size. In: Proceedings of the 4th Workshop on Asian Translation (WAT 2017), Taipei, Taiwan, pp. 99–109. Asian Federation of Natural Language Processing (2017). https://aclanthology.org/W17-5708
19. Nguyen, T., Chiang, D.: Improving lexical choice in neural machine translation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, vol. 1 (Long Papers), pp. 334–343. Association for Computational Lin-

guistics (2018). https://doi.org/10.18653/v1/N18-1031. https://aclanthology.org/N18-1031

20. Post, M.: A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, pp. 186–191. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-6319. https://aclanthology.org/W18-6319
21. Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., Neubig, G.: When and why are pretrained word embeddings useful for neural machine translation? In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, vol. 2 (Short Papers), pp. 529–535. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-2084. https://aclanthology.org/N18-2084
22. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Berlin, Germany, pp. 86–96. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-1009. https://aclanthology.org/P16-1009
23. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Berlin, Germany, pp. 1715–1725. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-1162. https://aclanthology.org/P16-1162
24. Sennrich, R., Zhang, B.: Revisiting low-resource neural machine translation: a case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 211–221. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1021. https://aclanthology.org/P19-1021
25. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, vol. 2 (Short Papers), pp. 464–468. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-2074. https://aclanthology.org/N18-2074
26. Stahlberg, F., Cross, J., Stoyanov, V.: Simple fusion: return of the language model. In: Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, pp. 204–211. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/W18-6321. https://aclanthology.org/W18-6321
27. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
28. Wang, Q., et al.: Learning deep transformer models for machine translation. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, Florence, Italy, pp. 1810–1822. Association for Computational Linguistics (2019). https://www.aclweb.org/anthology/P19-1176
29. Wang, X., Pham, H., Dai, Z., Neubig, G.: SwitchOut: an efficient data augmentation algorithm for neural machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 856–861. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1100. https://aclanthology.org/D18-1100

30. Wei, X., Yu, H., Hu, Y., Zhang, Y., Weng, R., Luo, W.: Multiscale collaborative deep models for neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 414–426. Association for Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.acl-main.40
31. Wu, L., et al.: Depth growing for neural machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 5558–5563. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1558. https://www.aclweb.org/anthology/P19-1558
32. Wu, L., et al.: Machine translation with weakly paired documents. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 4375–4384. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1446. https://aclanthology.org/D19-1446
33. Xiong, R., et al.: On layer normalization in the transformer architecture. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 10524–10533. PMLR (2020). https://proceedings.mlr.press/v119/xiong20b.html
34. Xu, H., van Genabith, J., Xiong, D., Liu, Q.: Dynamically adjusting transformer batch size by monitoring gradient direction change. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3519–3524. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.323. https://aclanthology.org/2020.acl-main.323
35. Xu, H., Liu, Q.: Neutron: an implementation of the transformer translation model and its variants. CoRR abs/1903.07402 (2019). https://arxiv.org/abs/1903.07402
36. Xu, H., Liu, Q., van Genabith, J., Xiong, D., Zhang, J.: Lipschitz constrained parameter initialization for deep transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 397–402. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.38. https://aclanthology.org/2020.acl-main.38
37. Xu, P., et al.: Optimizing deeper transformers on small datasets. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (vol. 1: Long Papers), pp. 2089–2102. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.163. https://aclanthology.org/2021.acl-long.163
38. Zhang, B., Titov, I., Sennrich, R.: Improving deep transformer with depth-scaled initialization and merged attention. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 898–909. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1083. https://www.aclweb.org/anthology/D19-1083