Parallel Mechanism Decoders in Pretrained Language Model-based Neural Machine Translation

Anonymous ACL submission

Abstract

Pre-trained language models (PLMs) have demonstrated their effectiveness in enhancing neural machine translation (NMT) tasks. While researchers have made numerous attempts to enhance the encoder, however, in decoder enhancement, the existing method neglects intra-layer information fusion, potentially resulting in the underutilization of encoder information. In this paper, we propose a model featuring a parallel mechanism decoder, facilitating the integration of PLM enhancements and enabling multi-granularity information fusion in the decoder. We evaluate our proposed method on the IWSLT14 De-En task and obtain significant improvements in model performance with tiny modifications.

1 Introduction

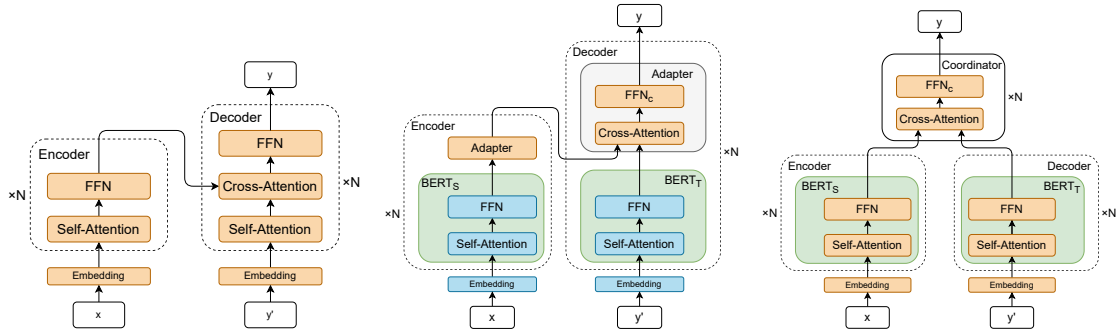
Pre-trained language models, such as GPT (Radford et al., 2018, 2019), BERT (Kenton and Toutanova, 2019), XLM (Conneau and Lample, 2019), have been extensively utilized to improve neural machine translation tasks (Baziotis et al., 2020; Sun et al., 2021; Weng et al., 2022a), and make significant progress. Among these models, BERT has attracted considerable attention from researchers in recent years (Yang et al., 2020; Weng et al., 2020; Hwang and Jeong, 2023) due to its compact design, ease of use, and exceptional performance quality.

From a structural perspective, BERT can be considered as a pre-trained encoder. Since NMT typically employs an "encoder-decoder" framework, efforts to leverage BERT for NMT enhancement generally focus on two key areas: encoder enhancement and decoder enhancement.

Regarding encoder enhancement, Guo et al. (2020) effectively relieves the catastrophic forgetting problem (McCloskey and Cohen, 1989) during fine-tuning by introducing adapters while keeping the pre-trained parameters frozen. Weng et al. (2022b) incorporates multi-task learning and a Layer-wise Coordination Structure to better exploit BERT's encoding capability, thus enhancing the overall quality of the translation model. Duan and Zhao (2023) further enhances the encoder's encoding capability through multi-task fine-tuning and Half-layers Knowledge Distillation techniques.

However, methods to integrate BERT into the decoder are relatively scarce. This may be due to significant differences between the decoder of translation models and BERT in terms of structure, functionality, etc., which limits its effectiveness in the decoder (Ma et al., 2021). Simple fine-tuning methods often fail to yield substantial improvements in the decoder (Weng et al., 2022b). Therefore, Duan and Zhao (2023) first proposed a method that can utilize BERT in the autoregressive decoder. By dividing the encoding and prediction functions of the decoder, they can effectively use BERT to enhance the encoding capability of the decoder. However, this approach overlooks the decoder's ability to integrate information from both the source and target sides. It only uses the outputs of the BERTs for the source and target languages, thereby depriving the decoder of the ability to fuse information from the source language across multiple granularities, ultimately constraining the model's final quality.

In this paper, to effectively utilize BERT in the decoder of translation models while achieving high-quality encoding, information fusion, and prediction capabilities, we propose a parallel mechanism decoder model. Each layer of the decoder consists of two parallel sub-layers: the encoding sub-layer directly employs the BERT of the target language to encode historical prediction information; the fusion sub-layer initializes by the same BERT and integrates information from the encoder output and the decoder state of the previous layer. With this method, our model can effectively leverage BERT's encoding and prediction capabilities while ensur-



(a) Transformer (b) AB-Net (c) External Coordinator

Figure 1: Structures of some existing autoregressive NMT models. The green layer indicates pre-trained BERT, the blue layer indicates that the parameters are frozen, and the orange layer indicates that the parameters are trainable.

ing a comprehensive fusion of information from both source and target languages at each layer, so as to improve the translation quality of the model.

We experimented on the IWSLT14 German-English dataset, and our model achieved 37.30 BLEU, representing a significant improvement of 0.82 BLEU over the baseline model. Furthermore, we conducted a series of comparative experiments to validate the effectiveness of our method.

2 Related Work

2.1 Transformer

The vanilla Transformer (Vaswani et al., 2017) works as an encoder-decoder model, wherein the encoder encodes the source language, and the decoder generates the target language translation word by word in an autoregressive fashion, relying on the encoder output and historical prediction.

Given a parallel sentence pair $\{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x}$ and $\mathbf{y}$ represent sentences in the source and target languages, respectively. The translation process of the Transformer can be expressed as $\mathbf{y} = \textbf{DEC}(\textbf{ENC}(\mathbf{x}))$, where the encoder **ENC** and decoder **DEC** are each composed of multiple layers. Illustrated in Figure 1(a), the output R^n of each layer of the encoder can be calculated by:

$$R^n = \textbf{FFN}^n(\textbf{S-ATT}^n(R^{n-1})) \quad (1)$$

where **S-ATT** and **FFN** are self-attention network and feed-forward network, respectively. Similarly, the output H^n of each layer of the decoder can be calculated by:

$$H^n = \textbf{FFN}^n(\textbf{C-ATT}^n(\textbf{S-ATT}^n(H^{n-1}), R^N)) \quad (2)$$

where **C-ATT** is cross-attention network.

Finally, the training objective L of the model is to minimize the negative log-likelihood, defined as:

$$L = -\textbf{log}P(\mathbf{y}|\mathbf{x};\theta) \quad (3)$$

$$P(\mathbf{y}|\mathbf{x}) = \textbf{softmax}(\textbf{Linear}(H^N)) \quad (4)$$

Where θ represents the model parameters, and **Linear** denotes a linear layer that transforms the last hidden states H^N into the vocabulary dimension for prediction.

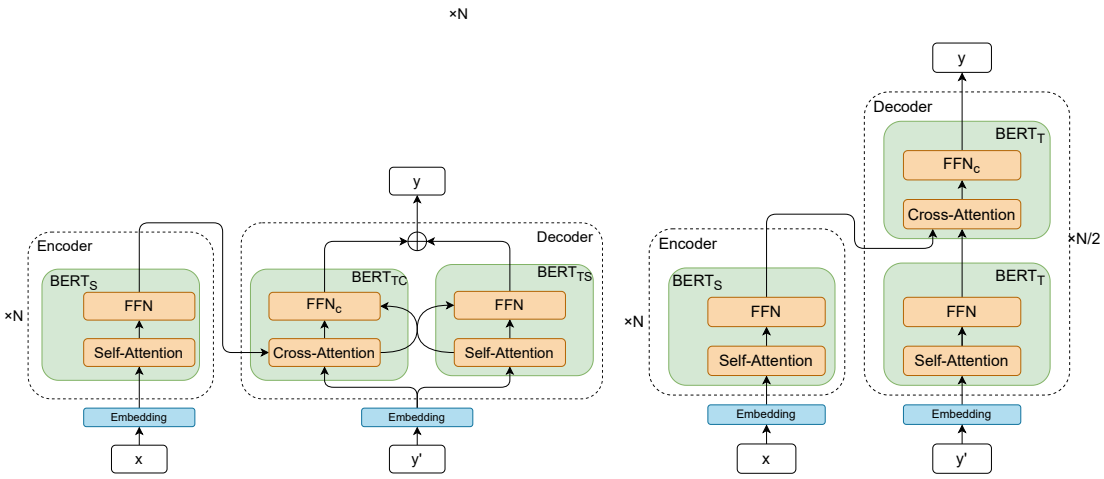
2.2 AB-Net

To enhance the quality of neural machine translation models leveraging pre-trained language models, Guo et al. (2020) introduced AB-Net. Illustrated in Figure 1(b), they utilize BERT models of the source language and target language as the primary components of the encoder and decoder, respectively. Through the incorporation of adapters (Bapna and Firat, 2019) at each layer and the freezing of pre-trained BERT parameters, they effectively mitigate the issue of catastrophic forgetting.

Considering the disparities between the parallel nature of BERT pre-training tasks and the autoregressive decoding process, they choose to construct a non-autoregressive translation model. In this model, the encoder predicts the decoding length additionally, and parallel decoding is executed in the decoder. Consequently, the training objective was adjusted to Masked-Prediction (Ghazvininejad et al., 2019):

$$L = -\sum_{t=1}^{m} P(\mathbf{y}_m|\mathbf{y}_r, \mathbf{x}; \theta_a) \quad (5)$$

where m represents the number of masked tokens, $\mathbf{y}_m$ and $\mathbf{y}_r$ respectively denote the masked tokens



(a) Parallel Decoder (b) Serial Decoder

Figure 2: Structures of parallel decoder model (Our) and serial decoder model. Each serial decoder layer is initialize with two adjacent $\mathbf{BERT}_T$ layers following Ma et al. (2021).

and the remaining tokens in $\mathbf{y}$, and θ_a represents the parameters contained in all adapters in the encoder and decoder.

2.3 External Coordinator

Duan and Zhao (2023) first proposed a method to integrate BERT into the autoregressive decoder, termed the External Coordinator, by partitioning the self-attention and cross-attention networks of the Transformer decoder into two distinct components. As illustrated in Figure 1(c), they grouped the self-attention network **S-ATT** and feed-forward network **FFN** of the decoder as the the decoder, initialized using BERT to encode historical prediction information. Furthermore, they introduced a coordinator (He et al., 2018), which takes the outputs of encoder and the decoder as inputs. Each layer of this coordinator comprised a cross-attention network $\mathbf{C\text{-}ATT}_c$ and a new feed-forward network $\mathbf{FFN}_c$, responsible for prediction. Expanding on this framework, they further enhanced the translation quality of the model by incorporating a series of auxiliary tasks to raise the capabilities of each part.

3 Our Model

Duan and Zhao (2023) achieved successful integration of BERT into the decoder of the translation model by splitting the decoder. However, this approach disrupted the multi-granularity interaction of encoding and decoding information. Motivated by previous works such as Guo et al. (2020); Ma et al. (2021); He et al. (2021), we take the self-attention network and cross-attention network in segmentation within the decoder and combine them in a parallel mechanism. Our method aims to preserve the capacity for intra-layer interaction between encoding and decoding information while separating the functionalities of self-attention and cross-attention as accomplished in Duan and Zhao (2023).

Specifically, our model is shown in Figure 2(a). In this architecture, the encoder employs the source language BERT. Thus, the output R^n of each layer of the encoder can be represented as:

$$\begin{aligned} R^n &= \mathbf{BERT}_s^n(x) \\ &= \mathbf{FFN}^n(\mathbf{S\text{-}ATT}^n(R^{n-1})) \end{aligned} \tag{6}$$

where $\mathbf{BERT}_s^n(x)$ represents the nth layer of the source language BERT. In each layer of the decoder, we consider the self-attention network **S-ATT** and the feed-forward network **FFN** as the encoding sub-layer to encode historical information. Similarly, the cross-attention network **C-ATT** and the newly introduced feed-forward network $\mathbf{FFN}_c$ are treated as the fusion sub-layer, so as to integrate the context representation R^n of the encoder output (referred to as the fusion module). Therefore, the output H^n of each layer of the decoder can be represented as:

$$H^n = \mathbf{BERT}_{tc}^n(H^{n-1}) + \mathbf{BERT}_{ts}^n(H^{n-1}) \tag{7}$$

$$\mathbf{BERT}_{tc}^n(H^{n-1}) = \mathbf{FFN}_c^n(\mathbf{ATT}^n(H^{n-1})) \tag{8}$$

$$\mathbf{BERT}_{ts}^n(H^{n-1}) = \mathbf{FFN}^n(\mathbf{ATT}^n(H^{n-1})) \tag{9}$$

$$\mathbf{ATT}^n(H^{n-1}) = \mathbf{S\text{-}ATT}^n(H^{n-1}) + \mathbf{C\text{-}ATT}^n(H^{n-1}, R^N) \tag{10}$$

#	Model	#Trainable Parameters	BLEU
1	Transformer	248M	34.15
2	Transformer + BERT init	200M	35.76
3	External Coordinator (Duan and Zhao, 2023)	329M	36.48
4	Serial Decoder	172M	36.65
5	Our	257M	**37.30**
6	Our + Freeze encoding sub-layer	171M	37.08

Table 1: The BLEU scores of the proposed model (Our) and the baseline methods on the IWSLT14 En-De task. The number of trained parameters are also reported.

Model	BLEU
AB-Net (serial decoder)	36.65
AB-Net (parallel decoder)	36.99

Table 2: The BLEU scores of the replicated AB-Net (Guo et al., 2020) and with parallel mechanism decoder.

4 Experiment

4.1 Datasets

We evaluated our model on the IWSLT14 English-German task, with a training set containing 160k sentences pairs, and development and test sets including 2k and 5k sentences pairs, respectively. We preprocessed the data following Ma et al. (2021), and filtered out training data with lengths exceeding 400.

4.2 Model Configurations

For the pre-trained BERT models used, following Duan and Zhao (2023), we employed 'bert-base-uncased' for English and 'dbmdz/bert-base-german-uncased' for German. Our model parameters are consistent with those of the pre-trained models used, and we fully utilized the tokenization and vocabulary of the pre-trained models.

4.3 Results

Main experimental results are presented in Table 1. We used a Transformer baseline with a configuration similar to the pre-trained model (line 1) and compared the model initialized using those pre-trained model (line 2). For existing methods, we reproduced the External Coordinator (Duan and Zhao, 2023) but omitted their extensive auxiliary tasks and embedding adjustments to ensure a relatively fair comparison of model structures (line 3). Our model (line 5) achieved 37.30 BLEU , which represents an improvement of 3.15 and 1.54 BLEU over random initialization and BERT initialization for the Transformer, respectively, and a 0.82 BLEU improvement over the External Coordinator. Our results indicate that our method can further enhance model quality by multi-granularity encoding-decoding information interaction, besides relieve catastrophic forgetting.

To further verify the effectiveness of the parallel mechanism, as shown in Figure 2(b), we implemented a serial mechanism decoder referring to Ma et al. (2021) (line 4). The results reveal that the parallel mechanism yielded a more larger improvement compared to the serial structure, with an increase of 0.65 BLEU. Even when freezing the encoding sub-layers to eliminate the influence of trainable parameters (line 6), the parallel mechanism still outperformed by 0.43 BLEU.

Finally, considering the similarity between our model and AB-Net (Guo et al., 2020), we replicated their work on non-autoregressive translation tasks and adapted the serial adapter in the decoder to a parallel mechanism. The results presented in Table 2 indicate that even in non-autoregressive tasks, the parallel mechanism can achieve a 0.34 BLEU improvement compared to the serial structure, further underscoring the effectiveness of our proposed parallel mechanism.

5 Conclusion

In this paper, we propose a parallel mechanism decoder, which encodes historical predication information and fuses source language information in the decoder. It allows for more effective utilization of pre-trained BERT and more comprehensive fusion of encoding and decoding information, to enhance the quality of the model. Comparative experiments with existing methods and a series of evaluations demonstrate the effectiveness of our model.

6 Limitations

Although our work has achieved some success, there are still existing the following limitations:

- Due to time constraints, we just performed experiments on one translation task. We will further validate our method on translation tasks in different languages and data scale.
- No attempts were combined with the multi-task framework proposed by existing methods. Hereafter, we will further try to combine the multi-task framework to explore more effective model structure and training methods.
- Just did try on the BERT pre-trained model. In the future, we will do research on other different types of pre-trained language models.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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)*, pages 1538–1548.

Christos Baziotis, Barry Haddow, and Alexandra Birch-Mayne. 2020. Language model prior for low-resource neural machine translation. In *The 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7622–7634. Association for Computational Linguistics (ACL).

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Sufeng Duan and Hai Zhao. 2023. Encoder and decoder, not one less for pre-trained language model sponsored nmt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3602–3613.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. 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)*, pages 6112–6121.

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *Advances in Neural Information Processing Systems*, 33:10843–10854.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. *Advances in Neural Information Processing Systems*, 31.

Soon-Jae Hwang and Chang-Sung Jeong. 2023. Integrating pre-trained language model into neural machine translation. *arXiv preprint arXiv:2310.19680*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Zewei Sun, Mingxuan Wang, and Lei Li. 2021. Multilingual translation via grafting pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2735–2747.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rongxiang Weng, Wensen Cheng, Qiang Wang, Changfeng Zhu, and Min Zhang. 2022a. Seq2seq pre-training with dual-channel recombination for translation.

Rongxiang Weng, Heng Yu, Shujian Huang, Shanbo Cheng, and Weihua Luo. 2020. Acquiring knowledge from pre-trained model to neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9266–9273.

Rongxiang Weng, Heng Yu, Weihua Luo, and Min Zhang. 2022b. Deep fusing pre-trained models into neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11468–11476.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.