TJUNLP:System Description for the WMT23 Literary Task in Chinese to English Translation Direction

Shaolin Zhu and Deyi xiong*
College of Intelligence and Computing, Tianjin University, Tianjin, China
{zhushaolin, dyxiong}@tju.edu.cn

Abstract

This paper introduces the overall situation of the Natural Language Processing Laboratory of Tianjin University participating in the WMT23 machine translation evaluation task from Chinese to English. For this evaluation, the base model used is a Transformer based on a Mixture of Experts (MOE) model. During the model's construction and training, a basic dense model based on Transformer is first trained on the training set. Then, this model is used to initialize the MOE-based translation model, which is further trained on the training corpus. Since the training dataset provided for this translation task is relatively small, to better utilize sparse models to enhance translation, we employed a data augmentation technique for alignment. Experimental results show that this method can effectively improve neural machine translation performance.

1 Introduction

Machine translation, as a core branch of natural language processing, has experienced significant development and received widespread attention in the past few years. Propelled by deep learning and neural networks, architectures like the Transformer(Vaswani et al., 2017) and its derivative models, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), have become mainstream methods for achieving efficient machine translation. These models, by learning underlying representations of language, are able to capture complex relationships and rich semantic information between texts.

Although neural machine translation with dense models has a promising future, it still faces many challenges. One of the main issues with the standard Transformer-based dense multilingual neural machine translation model is the model's capacity bottleneck(Zhu et al., 2021; Fedus et al., 2022b; Cheng et al., 2021). While increasing the model's depth and breadth can effectively enhance its capacity, it severely reduces the model's execution efficiency and increases the hardware requirements for training the model. This often results in the need for large GPU devices, limiting the model's applications. Therefore, in recent years, multilingual neural machine translation based on Mixture-of-Experts (MOE) (Fedus et al., 2022a) has been proposed. Compared to dense models, MOE-based multilingual machine translation activates only a portion of the network parameters during model training and inference (Lepikhin et al., 2021), giving it excellent computational efficiency. Under the same hardware conditions, it can achieve greater model capacity (compared to dense models, capacity can be increased by several tens of times) (Shazeer et al., 2017) and shorter computation time. Therefore, in this translation task evaluation, our basic model framework is based on the MOE Transformer. Furthermore, when there is limited available data, overfitting can easily occur (Wang et al., 2022; Pan et al., 2021). Combining the knowledge of multiple experts can often provide more accurate predictions than a single model. During model training, by allocating experts to focus on different input subsets, MOE can help alleviate the overfitting issue (Szymanski and Lemmon, 1993).

In this paper, we primarily focus on the WMT23 Chinese to English machine translation task. To enhance the model's capacity while maintaining a high computational efficiency, we employ a neural machine translation model based on the MOE Transformer framework. This model can effectively expand the model parameters. Moreover, since it's a domain-specific translation task with limited translation data corpus, we employed a strategy to initialize MOE using dense models effectively. The rest of this paper is organized as follows. In Section 2, we will present the models and methods we designed. Section 3 primarily

[*] Corresponding author.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 307–311
December 6–7, 2023. ©2023 Association for Computational Linguistics

showcases the experimental results and discusses and analyzes the outcomes. Section 4 concludes the paper and provides an outlook.

2 Method Description

To evaluate machine translation from Chinese to English, we need to construct a machine translation model. Therefore, in section 2.1, we first introduce the model's design and initialization strategy. In section 2.2, we primarily discuss the data alignment augmentation method, aiming to further utilize the data to enhance the model's performance. Finally, we introduce the model's training strategy.

2.1 Model Design

Compared to the MOE model, dense models perform better in bilingual settings (Costa-jussà et al., 2022). Given that the WMT23 machine translation evaluation task has relatively limited corpora, in order to enhance the model's performance, we first pretrain a dense model. Then, we use this dense model to initialize the MOE model. The framework of the model is illustrated in the Figure 1.

We first employ a 6x6 Transformer-based encoder-decoder framework to train the dense model. We can then use the parameters of this pretrained model to initialize the MOE-based translation model. The difference between the dense model and the MoE model lies in the fact that some FFN layers are replaced with MoE layers (Lin et al., 2020), while the rest of the structure remains identical. Therefore, we can directly initialize the embedding, Self-Attention, and Cross-Attention using the dense model. As for the MoE layer, it has a routing module and multiple FFN layers of the same size. We take the FFN layer parameters from the corresponding layer in the dense model (Komatsuzaki et al., 2023), add noise to increase the diversity of the initializing parameters, and then use these noisy FFN layer parameters to initialize each FFN in the MoE layer one by one. For the routing module, we initialize it randomly.

Specifically, our model in this paper adopts three stages. First, we train a basic multilingual neural machine translation model using the Transformer model. Upon successfully training the multilingual machine translation model, we select all of its parameters to initialize the MoE model. We need to create multiple expert sub-networks, and each expert sub-network will replicate the parameters of the corresponding FFN layer.

Next, we use the MoE model for self-supervised learning. Self-supervised learning is an unsupervised learning method that generates its own labels. For the machine translation task, one method of self-supervised learning is to use the original language text as input and then predict its translation. We mask 35% of the input text at random on a per-line basis. Ultimately, we compare the predicted text with the original text, compute the loss, and then update the model parameters. Finally, we use the MoE model, which has undergone self-supervised learning, to initialize a new MoE machine translation model. The expert sub-networks of the new model will replicate the parameters of the self-supervised learning model (Koishekenov et al., 2023). We then continue to train the new model until it meets our performance criteria.

2.2 Training Strategy

We preprocess all the data, removing special characters and standardizing punctuation marks. We uniformly apply SentencePiece (spm) (Kudo and Richardson, 2018) tokenization and construct a unified vocabulary with a size of 32,000. Additionally, we use the fairseq tool (Ott et al., 2019) for binarization. During training and decoding, the vocabulary is shared. We chose the Transformer as the foundational architecture and made improvements upon it to train bilingual models, multilingual dense models, and multilingual MOE models. We uniformly divided the data into training and validation sets. Since there is no test set, the final results are evaluated on the validation set. The model employs Adam (Kingma and Ba, 2015) as the optimizer to update model parameters. Every 30k steps, the model's performance is evaluated using the validation set. We use Polynomial Decay to dynamically adjust the learning rate, with the basic idea being to gradually decrease the learning rate as training progresses. For the dense model, it is trained for 100k steps. For the self-supervised model, we initialize the MOE model parameters using the dense model. We set the number of experts to 32, frequency to 4, expert capacity size to 1.0, and train for 50k steps. For the MOE model, we initialize the MOE model parameters using the self-supervised model. We set the number of experts to 32, frequency to 4, expert capacity size to 1.0, and train for 70k steps. During decoding, we adopt the beam search strategy, and the evaluation metric used is sacrebleu (Post, 2018)

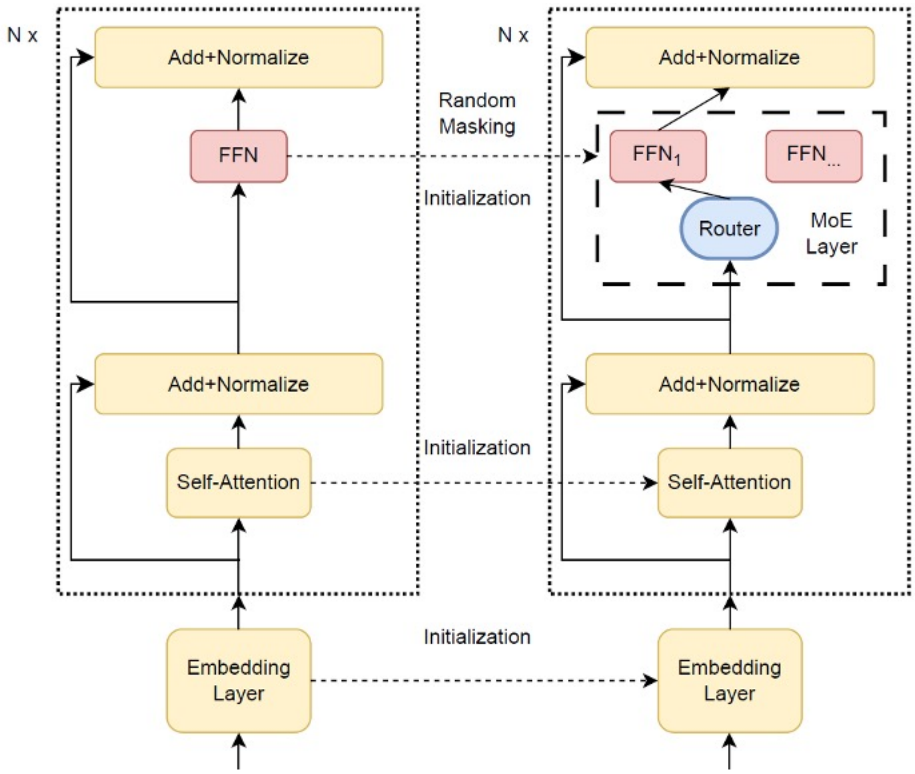

Figure 1: Taking the encoder as an example, the initialization process from the pretrained dense model to MOE is described. The process for the decoder is the same.

3 Experimental Results and Analysis

We first introduce the parameter settings of our trained models, and then analyze the experimental results.

3.1 Setting

The model is improved upon fairseq (MoE version) [1]. The training precision is uniformly set to fp16. Both the encoder and decoder are set to 6 layers with 8 attention heads each. The word embedding size is 512, and the hidden layer size is 1024. The loss function used is the cross-entropy function, and the optimizer is Adam, with beta1 set to 0.9 and beta2 set to 0.98. During the pretraining phase, the learning rate is set to 2e-4. A polynomial learning rate scheduling strategy is employed to optimize the learning rate, with warmup set to 4000. Dropout is set to 0.1. Each batch has a maximum of 4096 tokens, and gradients are updated every 4 accumulated batches.

3.2 Experimental Results

For this evaluation task, we did not compare our system with the current state-of-the-art NMT systems. The reason is that the organizers fixed the training data and system configurations to ensure a fair comparison among all participants.We use the Test1 and Test2 provided by the organizers as evaluation targets.

Model	Dense-MOE	Dense
Test1	21.59	19.08
Test2	17.89	15.48

Table 1: Evaluation Results for Dense-MOE and Dense.

In the experiments, we used sacrebleu as the evaluation metric. From Table 1, we can first observe that the method we employed in this paper achieved better performance compared to the dense model. After training, the dense model has already learned the basic patterns of the dataset. Using these parameters to initialize the MOE model allows the MOE model to start from a more optimal initial state, thereby converging quickly. Using the parameters of the dense model as initial values ensures that the MOE model has already grasped the basic features of the data at the onset of training. This provides a stable starting point for the MOE model, reducing the risks of instability and overfitting during training. Each expert in the MOE model can specifically handle certain distinct patterns or features in the

[1] https://github.com/facebookresearch/fairseq/tree/moe

data. By utilizing the pretrained dense model parameters, each expert in the MOE model can more rapidly identify its area of expertise, leading to a more efficient decomposition of model tasks. Even on the same dataset, due to its structural characteristics, the MOE model can capture more complex patterns in the data. With the initialization from the dense model's parameters, the MOE model can further optimize on this foundation, enhancing the model's expressive capability.

4 Conclusion

This paper introduces the main techniques and methods used for the WMT23 Chinese to English neural machine translation evaluation task. We employ a multilingual neural machine translation model based on the MOE Transformer framework. This model effectively achieves a vast and efficient parameterization. Moreover, given that it's a domain-specific translation task with limited translation data corpus, we utilized an effective strategy of initializing the MOE model using a dense model. This ensures that the MOE model has already grasped the fundamental features of the data at the start of training, providing a stable foundation for the MOE model and reducing the risks of instability and overfitting during training. Experimental results demonstrate that these methods can significantly improve the translation quality of neural machine translation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yong Cheng, Wei Wang, Lu Jiang, and Wolfgang Macherey. 2021. Self-supervised and supervised joint training for resource-rich machine translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1825–1835. PMLR.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William Fedus, Jeff Dean, and Barret Zoph. 2022a. A review of sparse expert models in deep learning. *CoRR*, abs/2209.01667.

William Fedus, Barret Zoph, and Noam Shazeer. 2022b. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3567–3585. Association for Computational Linguistics.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Taku Kudo and John Richardson. 2018. 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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018, pages 66–71. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2649–2663. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *CoRR*, abs/2105.09501.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Peter T. Szymanski and Michael D. Lemmon. 1993. Adaptive mixtures of local experts are source coding solutions. In *Proceedings of International Conference on Neural Networks (ICNN'88), San Francisco, CA, USA, March 28 - April 1, 1993*, pages 1391–1396. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael R. Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2591–2600. Association for Computational Linguistics.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2812–2823. Association for Computational Linguistics.