Dynamic Mask Curriculum Learning for Non-Autoregressive Neural Machine Translation

Yisong Wang, Hongxu Hou(✉), Shuo Sun, Nier Wu, Weichen Jian,
Zongheng Yang, and Pengcong Wang

National & Local Joint Engineering Research Center of Intelligent Information
Processing Technology for Mongolian, Inner Mongolia Key Laboratory of Mongolian
Information Processing Technology, College of Computer Science,
Inner Mongolia University, Hohhot, China
wangyisong06@126.com, cshhx@imu.edu.cn

Abstract. Non-autoregressive neural machine translation is gradually becoming a research hotspot due to its advantages of fast decoding. However, the increase of decoding speed is often accompanied by the loss of model performance. The main reason is that the target language information obtained at the decoder side is insufficient, and the mandatory parallel decoding leads to a large number of mistranslation and missing translation problems. In order to solve the problem of insufficient target language information, this paper proposes a dynamic mask curriculum learning approach to provide target side language information to the model. The target side self-attention layer is added in the pre-training phase to capture the target side information and adjust the amount of information input at any time by way of curriculum learning. The fine-tuning and inference phases disable the module in the same way as the normal NAT model. In this paper, we experiment on two translation datasets of WMT16, and the BLEU improvement reaches 4.4 without speed reduction.

Keywords: Non-autoregressive model · Curriculum learning · Mask ratio

1 Introduction

Neural machine translation (NMT) [1–3] has become a popular direction of research and has achieved great results. However, the mainstream autoregressive neural machine translation (AT) models have high decoding latency and exist in exposure bias [4]. Therefore, Gu et al. [5] proposed non-autoregressive neural machine translation (NAT), which uses parallel decoding to generate all tokens at once and improves the decoding speed significantly. However, this method can't obtain enough contextual information for the model to learn, and the generated translations suffer from a large number of mistranslation, missing translation and multi-modality problems.

T. Xiao and J. Pino (Eds.): CCMT 2022, CCIS 1671, pp. 72–81, 2022.
https://doi.org/10.1007/978-981-19-7960-6_8

Ding et al. [6] proposed that there are differences between the distillation data and the raw data, and simply using distillation data in one direction will result in poor translation of low-frequency words. Therefore, adding the knowledge distillation data in the opposite direction, which utilizes the target side data and solves the low-frequency word problem, but generating distillation data using only the target side data does not allow the decoder to obtain more information on the target side. Ran et al. [7] proposed that the decoding stage makes use of reordering information. Reorder the source copy token so that the position of each token is aligned with the target language token. Although makes use of word alignment information at the target side, but semantic information is not sufficiently obtained. Guo et al. [8] proposed fine-tuning by curriculum learning (FCL-NAT), which transfers the knowledge learned from the AT model to the NAT model by way of curriculum learning. However, this approach requires training the AT model first and then fine-tuning it using curriculum learning. This approach greatly increases the training time and consumes a lot of resources.

Obtain more linguistic information at the target side, some researchers have proposed a semi-autoregressive model with multiple iterations of decoding. Therefore, Gu et al. [9] proposed Levenshtein Transformer (LevT), which modifies the translation by three operations: delete, insert, and replace placeholders. More contextual information can be obtained during the translation adjustment process. The mask prediction method proposed by Ghazvininejad et al. [10] replaces a token with a lower probability with a mask and re-predicts it after each generation. It stops after two iterations unchanged or after reaching the maximum number of iterations. Although the above method can provide enough target side information for the model by multiple iterations, the increase in the number of iterations is accompanied by a decrease in the decoding speed, which can even degrade to the autoregressive model level and lose the advantage of NAT. Qian et al. [11] proposed GLAT, which uses the token of partial ground truth translation to replace the source copy token, and the model obtained by training in this way can achieve better performance. It is illustrated that, the performance of the model can be improved without losing speed by incorporating more target side information based on the model decoded in a single iteration.

In this paper, we propose a **dynamic mask** method based on **curriculum learning** (DMCL) to generate ground truth translations with mask for model training, so that the decoder can obtain more linguistic information on the target side. Specifically, the number of masks for the ground truth translations is dynamically increased in each training phase by means of curriculum learning, and the ground truth translations with mask are input to the decoder side. The target self-attention layer is added at the decoder side to obtain the target language information and fuse it with the self-attention layer information. The target language information provided can be limited by the mask ground truth token to prevent relying too much on the target self-attention part in the training phase. The number of masks is dynamically adjusted using a curriculum learning approach so that the model can be trained from easy to difficult, and the training process is smoother and achieves better model performance. In the fine-tuning phase, the target self-attention is removed, and the model is identical to the common NAT model. The experimental results show that the maximum improvement of BLEU value is more than 4.4 without losing decoding speed. It is noted that the DMCL approach in this paper is also applicable to the model with multiple iterations of decoding.

2 Background

2.1 Non-autoregressive Neural Machine Translation

The non-autoregressive model is based on the hypothesis that all words in the target language are independent of each other, and generates all target language words in parallel [5]. The generation process can be expressed as follows:

$$P(y|x) = P(T_y|x) \cdot \prod_{t=1}^{T_y} P(y_t|x, z) \tag{1}$$

where T_y denotes the length of the target sentence, x denotes the source language sentence, and y denotes the target language sentence. From the Eq. (1), it can be seen that although the hidden variable of z is involved in the decoding stage, the latent variables are also derived from the source side language. Therefore, this approach does not fully utilize the target side language information in the training phase, but forcibly decodes the translation based on the latent variables. In contrast, the DMCL proposed in this paper can provide part of the target side information in the pre-training stage, so that the model learns richer target side information.

2.2 Curriculum Learning

Curriculum learning is a strategy to train a model from easy to difficult. This asymptotic training approach allows the model to be smoother during the training phase while achieving better results. Platanios et al. [12] proposed a new training framework that decides the next phase of input to the model based on the difficulty of the training data and the current model capabilities. There are two important metrics under this training framework, data difficulty and model competence. The data difficulty can be calculated based on the sentence length or the average word occurrence probability. The model competence uses a predefined incremental function. The input data difficulty at each stage is less than the current model competence. In this paper, the same idea is adopted, and DMCL determines the amount of target side language information provided in the next step based on the current status of the model. The DMCL strategy adjusts the amount of target side language information provided to enable the model to achieve better results compared to the strategy that doesn't use the course learning approach.

3 Method

In this section, a detailed description of the model structure of DMCL-NAT and the dynamic mask curriculum learning training strategy will be illustrated.

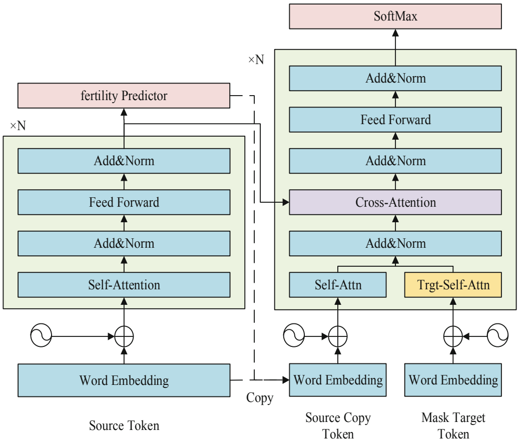

Fig. 1. The model structure of DMCL-NAT. Where trgt-self-attention is added to this paper. Residual connectivity is dispensed with in the figure.

3.1 Model

The encoder side of the model is identical to the Transformer's encoder, and the fertility predictor is added to the encoder side to predict the target sentence length. DMCL -NAT adds language information on the target side mainly at the decoder side. It also gradually reduces the amount of incorporated information in a curriculum learning manner, thus achieving an easy-to-hard training strategy. Firstly, the symbolic representation is defined, and the source language token sequence is denoted as $X = \{x_1, x_2, x_3, ..., x_n\}$, and the target language token sequence is denoted as $Y = \{y_1, y_2, y_3, ..., y_T\}$. The main structure of the model is shown in Fig. 1. The input to the decoder side has two parts, one part of the copy from the source language according to the fertility predictor denoted as $X^* = \{x_1, x_2, x_2...x_T\}$, and the other part replaces the token in the ground truth translation with the mask according to the mask ratio. The input is the mask target token denoted as $Y^* = \{y_1, y_2, [MASK], [MASK], ...y_T\}$. The DMCL strategy will be described in detail on the next section. The two part inputs are embedding and their respective self-attention modules:

$$H_{self-attn} = MultiHead(Emb(x^*), Emb(x^*), Emb(x^*)) \tag{2}$$

$$H_{tart-self-attn} = MultiHead(Emb(y^*), Emb(y^*), Emb(y^*)) \tag{3}$$

where $Emb(\cdot)$ denotes the word embedding. After obtaining the self-attention results of the two parts, the two parts are fused and expressed as:

$$H = 0.5 * (H_{self-attn} + H_{tart-self-attn}) \tag{4}$$

where H is the result of the fusion of the two parts of the self-attention. In the model fine-tuning and inference phase then the target self-attention layer is disabled, returning to the original structure of the NAT model.

3.2 Dynamic Mask Curriculum Learning

Due to the feature of parallel decoding of NAT model, if all the target language information is directly introduced at the decoder side, the model will completely rely on the target self-attention part in the training phase, and the originally self attention part can't be adequately trained, resulting in the model losing the ability to generate translations after the target self-attention is removed in the inference phase. Therefore, it is necessary to limit the amount of information provided in the target language.

Inspired by BERT [13], Replace some words in the target sentence with mask tokens. To prevent the problem of over-fitting and not decoding properly, the mask ratio should be more than 50%. Therefore, this paper adopts the curriculum learning method to dynamically adjust the proportion of tokens in mask, and its value range should be $[0.5, 1]$.

The mind of curriculum learning is to let the model train from easy to difficult. When the mask is less, more contextual semantic information can be provided to the model, and as the mask ratio keeps increasing, the ground truth translation information that the model can refer to keeps decreasing. Therefore, the adjustment function for the mask ratio should be an increasing function overall. Platanios et al. [12] proposed a function taking values between 0 and 1 and increasing with the number of training steps:

$$ratio(t) = min(1, \sqrt[p]{\frac{t}{T}(1 - c_0^p) + c_0^p}) \tag{5}$$

where c_0 is the starting value, t is the current number of training steps, and T is the total number of curriculum learning steps. When $p = 1$, it is a linear increasing function, and when $p = 2$, $ratio(t)$ increases gradually less as t increases. From the existing course learning experience, generally $p = 2$ works best.

However, this way of taking values still has drawbacks. The main reason is that the mask ratio cannot be adjusted in time for different training conditions and can only be trained in a predefined way. Therefore, dynamically adjusting the mask ratio according to the current training condition of the model can make the model achieve better results. Therefore, this paper proposes a dynamic adjustment strategy that follows the change of last step loss. The equation is as follows:

$$ratio(loss) = min(1, \sqrt[2]{\frac{loss_{min}}{loss}(1 - c_0^2) + c_0^2}) \tag{6}$$

where $loss$ denotes the loss obtained by the model for the previous stage of calculation, and $loss_{min}$ denotes the loss when the autoregressive model reaches the convergence state, c_0 is the minimum mask ratio. As can be seen from the Eq. (6), as the loss decreases indicates that the current model can reach a better learning state, so the mask ratio can be increased appropriately, and when the loss increases during the training process indicates that the current stage is difficult to train, the mask ratio can be reduced appropriately.

3.3 Train and Inference

Train: The model is divided into pre-training and fine-tuning phases during training. The pre-training process is shown in Fig. 2. In the pre-training phase, which can also be called the mask curriculum learning phase, the mask ratio is dynamically adjusted according to the current training status of the model, and some of the target side language information is added so that the model can learn more target side language information. Can be expressed as:

$$P(y|x, y^*) = P(T_y|x) \cdot \prod_{t=1}^{T_y} P(y_t|x, y^*, z) \tag{7}$$

where y^* denotes the ground truth translation with mask token and T_y denotes the target sentence length. However, due to the existence of target self-attention there is still the problem of partial information leakage. Therefore, the fine-tuning phase removes target self-attention completely and doesn't introduce the ground truth translation information. The fine-tuning process can be expressed as Eq. (1).

Inference: The method in this paper only modifies the model structure in the training phase, and the fine-tuned model no longer relies on the target side information provided by target self-attention. The inference stage is the same as in Eq. (1). The sentence length is obtained according to the fertility prediction and the decoder needs the latent variable z copied from the source language. In addition, this paper also uses the noise parallel decoding (NPD) [5] method to generate the translation, the candidate set is increased according to the sentence length in the inference stage, and then the optimal result is selected from all the candidate sets as the final translation, which can make a better decision on the sentence length. Therefore, the inference stage is the same as the ordinary NAT model, and the model performance is further improved without affecting the decoding speed.

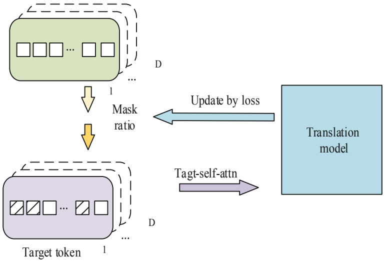

Fig. 2. Dynamic mask curriculum learning process. The shaded area indicates the token that was masked

4 Experiment

4.1 Data Preparation

In this paper, experiments were conducted on two sets of language pairs. The WMT16 RO-EN dataset is v6 version containing a total of 220K sentence pairs. For the WMT16 EN-DE dataset (2M sentence pairs), 1M sentence pairs were selected as the training set. On the RO-EN and EN-DE tasks, all the corpus was preprocessed by Byte Pair Encoding (BPE) [14], and the BPE dictionary size was set to 32K for both.

4.2 Configuration

For all the above datasets, the experimental model configurations all follow the settings of Vaswani et al. [2]. The decoder and encoder were each set to n-layers = 6, where the attention module d-model = 512, n-heads = 8. The warm-up steps were all set to 4000. The learning rate was set to 0.0005, and the learning rate update followed the inverse square root annealing algorithm. For the RO-EN dataset, a total of 6W steps are trained, and for the EN-DE dataset, a total of 30W steps are trained, where the DMCL pre-training are set to half of the total training steps.

4.3 Baseline

The experimental baseline in this paper is derived from the autoregressive model, the non-autoregressive model with single decoding, and the semi-autoregressive model with multiple iterations of decoding.

Transformer [2]: Autoregressive model strong baseline.

NAT [5]: The NAT model proposed by Gu et al. assumes parallel decoding with individual tokens directly independent of each other.

Mask Predict [10]: The token with lower probability in each generated translation is replaced with mask and re-predicted. The final translation is generated after several iterations.

All the above baseline and methods in this paper are implemented based on fairseq [15]. Choose the BLEU [16] value to evaluate the model performance.

4.4 Results

The main results of the experiments are shown in Table 1. The DMLM approach can significantly improve the performance of the NAT model. Compared with the multiple iteratives decoding model, the method in this paper retains the original fast decoding advantage of the NAT model and significantly reduces the performance difference with the multiple iteratives decoding model. Compared with the vanilla NAT model, significant improvements are obtained on the RO-EN dataset, with a maximum performance gain of more than 4.4. In addition, great potential is shown on large corpora of millions.

Table 1. Results on the WMT16 RO-EN and EN-DE benchmarks. m denotes the noise parallel decoding window size. DMCL-NAT is the method proposed in this paper. The bolded results indicate the best performance of single decoding.

Models		WMT2016			Speedup
		RO-EN	EN-RO	EN-DE	
AT Model	Transformer	36.64	34.65	22.13	1.0×
Iterative NAT	Mask-Predict (iter = 10)	35.22	32.96	18.43	2.6×
Fully NAT	NAT	26.46	25.32	12.56	15.7×
	NAT (m = 5)	28.83	27.07	12.92	7.6×
Ours	DMCL-NAT	30.47	28.81	14.74	15.7×
	DMCL-NAT (m = 5)	**33.27**	**31.76**	**15.44**	7.6×

In this paper, only the target self-attention layer is added to the pre-training process of the model, and the amount of target language information input is adjusted by adjusting the number of masks in the input ground truth translation. Therefore the method can be applied to a variety of NAT models.

5 Analysis

5.1 Mask Strategy

In this paper, two points of view are evaluated to verify the effectiveness of DMCL. Firstly, the mask ratio is fixed to a certain value. In addition, a strategy of mask ratio adjustment with the idea of curriculum learning is adopted, which gradually increases from 0.5 to 1. There are four incremental functions set in Table 3.

The experimental results are shown in Table 2 and Table 3. The optimal performance is reached when the fixed mask ratio is 0.7. When the incremental function is used, the inverse quadratic incremental function achieves the maximum value and is stronger than the model performance when the ratio is fixed. But this DMCL strategy proposed in this paper obtains a significantly better model performance than all the above approaches. The main reason is that dynamically adjusting the amount of information input allows the model to obtain the most appropriate amount of information and achieve better training results. So the DMCL strategy proposed in this paper is effective.

Table 2. Performance on WMT16 RO-EN with fixed mask ratio.

Mask ratio	BLEU
0.5	29.78
0.6	29.91
0.7	30.00
0.8	29.57
0.9	28.36
Ours	**30.47**

Table 3. Performances on WMT16 RO-EN with incremental mask ratio.

Functions	Description	BLEU
Linear	$\frac{t}{T}(1-c_0)+c_0$	29.85
Sqrt	$\sqrt[2]{\frac{t}{T}(1-c_0^2)+c_0^2}$	30.23
Exponent	$e^{-log\frac{t}{T}(1-c_0)+c_0}$	29.72
Ladder-like	$\lfloor\frac{5*t}{T}\rfloor*0.1+c_0$	29.82
Ours		**30.47**

Table 4. Performance on WMT16 RO-EN when DMCL applied to Mask Predict.

Model	BLEU
Transformer	36.64
Mask Predict (iter = 10)	35.22
Mask Predict+DMLM (iter = 10)	35.87

5.2 Method Generality

Since the method in this paper is to add target self-attention at the decoder side and then pre-train the model by DMCL. So the method is also applicable to the non-autoregressive translation model with multiple iterations of decoding. To test this hypothesis, the multiple iterations decoding model Mask Predict was chosen as the base model and experiments were conducted on the RO-EN task. The experimental results are shown in the Table 4, after adding the method of this paper to Mask Predict, the BLEU value has improved by 0.65. The reason why the improvement is not as large as that of the model with single decoding is that DMCL provides linguistic information on the target side in the training phase, while the same information on the target side is available during the iteration of Mask Predict. Therefore, the impact of DMCL is weakened.

6 Conclusion

In this paper, we propose a new method that can incorporate the target side language information in the NAT model, while dynamically adjusting the ratio of mask substitution in the ground truth translation in a curriculum learning approach, and controlling the amount of target language information provided by the ground truth translation can achieve a progressive learning process from easy to difficult. The method significantly improves the performance of the single decoding model without speed loss. Also, experiments are conducted in this paper based on the Mask Predict model, and it is demonstrated that the method is also applicable to models with multiple iterations of decoding. Providing target side information at the decoder side can effectively improve NAT model performance, and future research will focus on exploring more appropriate curriculum learning strategies and ways to apply the approach to other generative tasks.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Vaswani, A., et al.: Attention is all you need. Adv. Neural. Inf. Process. Syst. **30**, 5998–6008 (2017)
3. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning, pp. 1243–1252. PMLR (2017)
4. Yu, L., Zhang, W., Wang, J., Yu, Y.: SeqGAN: sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, No. 1 (2017)
5. Gu, J., Bradbury, J., Xiong, C., Li, V.O., Socher, R.: Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281 (2017)
6. Ding, L., Wang, L., Liu, X., Wong, D.F., Tao, D., Tu, Z.: Rejuvenating low-frequency words: making the most of parallel data in non-autoregressive translation. arXiv preprint arXiv:2106.00903 (2021)
7. Ran, Q., Lin, Y., Li, P., Zhou, J.: Guiding non-autoregressive neural machine translation decoding with reordering information. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, No. 15, pp. 13727–13735 (2021)
8. Junliang Guo, X., Tan, L.X., Qin, T., Chen, E., Liu, T.-Y.: Fine-tuning by curriculum learning for non-autoregressive neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 7839–7846 (2020)
9. Gu, J., Wang, C., Zhao, J.: Levenshtein transformer. arXiv preprint arXiv:1905.11006 (2019)
10. Ghazvininejad, M., Levy, O., Liu, Y., Zettlemoyer, L.: Mask-predict: parallel decoding of conditional masked language models. arXiv preprint arXiv:1904.09324 (2019)
11. Qian, L., et al.: Glancing transformer for non-autoregressive neural machine translation. arXiv preprint arXiv:2008.07905 (2020)
12. Antonios Platanios, E., Stretcu, O., Neubig, G., Poczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. arXiv preprint arXiv:1903.09848 (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
15. Ott, M., et al.: fairseq: a fast, extensible toolkit for sequence modeling. In: Proceedings of NAACL-HLT 2019: Demonstrations (2019)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318 (2002)