

GLAT: Glancing Transformer for Non-Autoregressive Neural Machine Translation

Lihua Qian^{1*} Hao Zhou² Yu Bao³ Mingxuan Wang²

Lin Qiu¹ Weinan Zhang¹ Yong Yu¹ Lei Li²

¹ Shanghai Jiao Tong University ² ByteDance AI Lab ³ Nanjing University
 {qianlihua, lqiu, wnzhang, yyu}@apex.sjtu.edu.cn
 {zhouhao.nlp, wangmingxuan.89, lileilab}@bytedance.com
 baoys@mail.nju.edu.cn

Abstract

Although non-autoregressive models with one-iteration generation achieve remarkable inference speed-up, they still fall behind their autoregressive counterparts in prediction accuracy. The non-autoregressive models with the best accuracy currently rely on multiple decoding iterations, which largely sacrifice the inference speed of non-autoregressive models. Inspired by the way of learning word dependencies in autoregressive and iterative-decoding models, we propose Glancing Transformer (GLAT) with a glancing language model (GLM), which learns to capture the word dependency gradually. Experiments on three benchmarks demonstrate that our approach can significantly improve the accuracy of non-autoregressive models without multiple decoding iterations. In particular, GLAT achieves state-of-the-art results among non-iterative models and even outperforms top iterative counterparts in some specific benchmarks.

1 Introduction

Non-autoregressive transformer (NAT) has attracted wide attention in neural machine translation (Gu et al., 2018), which generates sentences simultaneously rather than sequentially. To enable parallel decoding, NAT imposes a *conditional independence assumption* among words in the output sentences, which leads to significantly faster inference speed (almost a dozen times speed-up) than the autoregressive Transformer (Vaswani et al., 2017). However, NAT still falls behind autoregressive Transformer (AT) in the quality of output sentences, such as BLEU (Papineni et al., 2002) for machine translation. We blame it for the imposed conditional independence assumption, which prevents NAT models from explicitly learning the *word dependencies* in the output sentence. Note that such word dependency is crucial, and it is explicitly learned in the AT model through the *autoregressive language models* (left-to-right, see Figure 1a).

Recently, Ghazvininejad et al. (2019); Gu et al. (2019) propose to employ the *Masked Language Model* (MLM, Devlin et al., 2019) in NAT, which includes word dependency modeling in an *iterative* fashion (see Figure 1c), therefore yielding quite competitive results compared to AT. Specifically, such iterative models randomly mask words in the reference and predict these masked words conditioned on unmasked ones during training. In this manner, iterative models are trained to explicitly capture the dependencies between masked words and unmasked words. However, these iterative approaches still produce poor results with one decoding iteration and have to perform multiple iterations during inference, namely iteratively refining the generated outputs of the previous iteration. Such iterative process is quite time-consuming, which partly sacrifices the speed merit of NAT. To date, it remains

*The work was done when the first author was an intern at Bytedance.

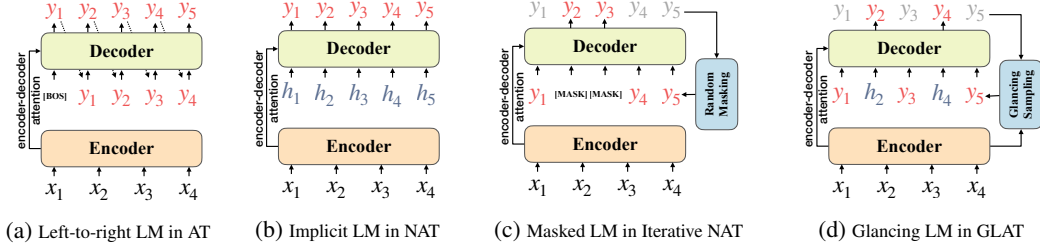


Figure 1: Different language modeling approaches of different text generation models.

an open question as to how the iterative process can be abandoned, while still preserving the benefits of explicitly modeling word dependencies in NAT.

In this paper, we argue that the major culprit of the problem that mask language models have to be used together with iterative inference, is the sampling strategy of masking words in MLM. In particular, MLM employs a fixed uniform strategy for randomly masking words during training, which prevents the model from effectively learning word dependencies for one-iteration generation. For example, at the beginning of training when the NAT model is still poorly tuned, we should mask fewer words. If not, it would be difficult for the NAT model to correctly predict the masked words. On the contrary, if we mask too little words at the end phase of training, the resulting NAT model is rarely trained to predict the whole sentences, and can only predict some sentence fragments. In such a case, to accurately generate the whole sentence in inference, the NAT model has to generate the sentence fragments iteratively. To this end, the sampling strategy is crucial for the training of NAT.

To address the above issues, we propose a simple yet effective approach called *Glancing Transformer* (GLAT), which is equipped with the proposed *Glancing Language Model* (GLM) for non-iterative parallel text generation, achieving significant improvements upon strong baselines. Intuitively, GLM adopts a *adaptive glancing sampling* strategy, which glances at some fragments of the reference if the reference is too difficult to fit in the training of NAT. Correspondingly, when the model is well tuned, it will adaptively reduce the percentage of glancing sampling, making sure that the resulting model could learn to generate the whole sentence in the one-iteration fashion.

Specifically, our proposed GLM differs from MLM in two aspects. Firstly, GLM proposes an adaptive glancing sampling strategy, which enables GLAT to generate sentences in a one-iteration way, working by gradual training instead of iterative inference (see Figure 1d). Generally, GLM is quite similar to curriculum learning (Bengio et al., 2009) in spirit, namely first learning to generate some fragments and gradually moving to learn the whole sentences (from easy to hard). To achieve the adaptive glancing sampling, GLM performs decoding twice in training. The *first decoding* is the same as the vanilla NAT, and the prediction accuracy indicates whether current reference is “difficult” for fitting. In the second decoding, GLM gets words of the reference via glancing sampling according to the first decoding, and learn to predict the remaining words that are not sampled. Note that only the second decoding will update the model parameters. Secondly, instead of using the [MASK] token, GLM directly use representations from the encoder at corresponding positions, which is more natural and could enhance the interactions between sampled words and signals from the encoder.

Experimental results show that GLAT obtains significant improvements (about 5 BLEU) on standard benchmarks compared to the vanilla NAT, without losing inference speed-up. GLAT achieves competitive results against iterative approaches like Mask-Predict (Ghazvininejad et al., 2019), even outperforming the Mask-Predict model on WMT14 DE-EN and WMT16 RO-EN. Compared to the strong AT baseline, GLAT can still close the performance gap within 1 BLEU point while keeping $7.9\times$ speed-up. Empirically, we find that GLAT outperforms AT when the source input length is less than 20 on WMT14 DE-EN. We speculate this is because GLM could capture bidirectional context while left-to-right LM is unidirectional, which indicates the potential of parallel generation models.

2 Text Generation via Conditional Language Modeling

In this section, we compare different language models used in different text generation approaches. Formally, considering a sequence-to-sequence model (Cho et al., 2014; Bahdanau

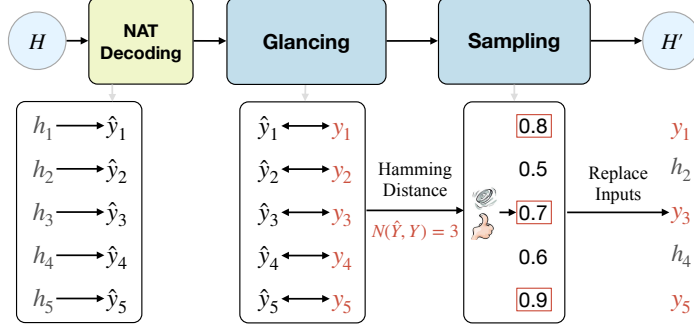


Figure 2: Illustration of the procedure of glancing sampling

et al., 2014; Vaswani et al., 2017) for predicting $Y = \{y_1, y_2, \dots, y_T\}$ given the input sentence $X = \{x_1, x_2, \dots, x_N\}$. In the AT model, the training objective is maximizing the log-likelihood with autoregressive decomposition:

$$\mathcal{L}_{AT} = \sum_{t=1}^T \log p(y_t | y_{<t}, X; \theta), \quad (1)$$

where the word y_t is conditioned on the target prefix $y_{<t} = \{[\text{BOS}], y_1, \dots, y_{t-1}\}$ and the source input X . AT models sentences from left-to-right, therefore word dependencies are learned in a unidirectional way. NAT, on the other hand, incorporates conditional independence assumption among words in a sentence with the aim of enabling parallel generation:

$$\mathcal{L}_{NAT} = \sum_{t=1}^T \log p(y_t | X; \theta). \quad (2)$$

Note that, in NAT, different y_t in Y are predicted simultaneously, which removes the interactions among target words in the NAT modeling. Thus, to generate fluent and faithful sentences, NAT has to model the target word dependencies implicitly, which makes the learning process quite challenging.

To explicitly model the target word dependencies, previous iterative approaches, such as Mask-Predict, introduce mask language models (MLM) in NAT. MLM models word dependencies by learning to predict the masked words conditioned on the unmasked ones:

$$\mathcal{L}_{MLM} = \sum_{y_t \in \mathbb{RM}(Y)} \log p(y_t | \Phi(Y, \mathbb{RM}(Y)), X; \theta). \quad (3)$$

Here $\mathbb{RM}(Y)$ returns some randomly sampled words from Y , and Φ replaces these sampled words in Y with the $[\text{MASK}]$ token. For example, if $\mathbb{RM}(Y) = \{y_2, y_3\}$, $\Phi(Y, \mathbb{RM}(Y)) = \{y_1, [\text{MASK}], [\text{MASK}], y_4, \dots\}$. To this end, the training objective is to predict the masked words $\mathbb{RM}(Y)$ given the source sentence X and the unmasked target words.

As mentioned in the Introduction, though MLM can explicitly model target word dependencies, it can hardly generate satisfactory sentences without multiple iterations. Better language modeling approach should be explored for parallel text generation in the one-iteration way.

3 The Proposed Glancing Transformer

In this section, we describe GLAT in detail. Generally, GLAT differs from vanilla NAT in that it explicitly models word dependencies via the proposed glancing language model (GLM), which leads to significant accuracy improvements. Additionally, compared to these iterative approaches that adopt MLM, the proposed GLM employs an adaptive glancing sampling strategy, which enables GLAT to boost the performance of one-iteration generation and generate promising sentences without multiple iterations. And GLAT further strengthen the decoder inputs with the source representation.

3.1 The Glancing Language Model

Formally, given the training data $D = \{X, Y\}_{i=1}^L$, the task is predicting $Y = \{y_1, y_2, \dots, y_T\}$ with the input sentence $X = \{x_1, x_2, \dots, x_N\}$. By employing GLM, the training objective of GLAT is:

$$\mathcal{L}_{\text{GLAT}} = - \sum_{\{X, Y\} \in D} \sum_{y_t \in \{Y \setminus \mathbb{GS}(Y, \hat{Y})\}} \log p(y_t | \mathbb{GS}(Y, \hat{Y}), X; \theta). \quad (4)$$

Here, \hat{Y} is the predicted sentences in the first decoding pass, and $\mathbb{GS}(Y, \hat{Y})$ is the set of sampled target words by the sampling strategy of \mathbb{GS} (described in detail in the next section). The sampling strategy samples more words from reference Y if prediction \hat{Y} is less accurate, and samples fewer words for the opposite case. Additionally, $\{Y \setminus \mathbb{GS}(Y, \hat{Y})\}$ is the difference set, representing the remaining words except for these sampled words. Then we perform the second decoding, updating the model parameters θ , by maximizing the likelihood of these remaining words with cross-entropy.

Specifically, GLM performs decoding twice in its learning procedure. In the first decoding pass, given the encoder $\mathbb{F}_{\text{encoder}}$ and decoder $\mathbb{F}_{\text{decoder}}$, $H = \{h_1, h_2, \dots, h_T\}$ is the encoded output sequence gathered from the input X , and $\hat{Y} = \mathbb{F}_{\text{decoder}}(H, \mathbb{F}_{\text{encoder}}(X; \theta); \theta)$ is the predicted sentence in the first decoding pass. With the predicted sentence \hat{Y} , $\mathbb{GS}(Y, \hat{Y})$ is the set of sampled words from reference Y , according to our adaptive sampling strategy \mathbb{GS} that will be introduced in the next section. Note that we use the attention mechanism to form the decoder inputs with the input X . Previous work adopts *Uniform Copy* (Gu et al., 2018) or *SoftCopy* (Wei et al., 2019) instead. But empirically, we find that they produce almost the same results in our setting.

In the second decoding pass, we cover the original decoding inputs H by the embeddings of words from $\mathbb{GS}(Y, \hat{Y})$ to get the new decoding inputs $H' = \mathbb{F}_{\text{cover}}(E_{y_t \in \mathbb{GS}(Y, \hat{Y})}(y_t), H)$, where $\mathbb{F}_{\text{cover}}$ covers the corresponding positions. Namely, if we have a sampled word at one position, we use its word embedding to replace the original decoding input at the same position. Here the word embeddings are obtained from the softmax embedding matrix of the decoder. With the mix of encoding signals and reference words from glancing sampling H' as decoder inputs, the training objective of our proposed *glancing language model* can be written as Equation 4, where the probabilities of remaining words on each position $p(y_t | \mathbb{GS}(Y, \hat{Y}), X; \theta)$ are computed by $\mathbb{F}_{\text{decoder}}(H', \mathbb{F}_{\text{encoder}}(X; \theta), t; \theta)$.

3.2 The Glancing Sampling Strategy

We adopt the glancing sampling in GLM to adaptively sample words from the reference. Intuitively, the sampling strategy of our proposed glancing sampling method will sample many words at the start of the training, when the model is not yet well tuned. As the model gets better progressively, the sampling strategy will sample fewer words to enable the model to learn the simultaneous generation of the whole sentence. Note that the sampling strategy is crucial in the training of NAT. Our adaptive sampling strategy guide the model to first learn the generation of fragments and then gradually turn to the whole sentences.

As illustrated in Figure 2, the glancing sampling could be divided into two steps: first deciding a sampling number N adaptively, and then *randomly* selecting N words from the reference. The sampling number N will be larger when the model is poorly trained and decreases along the training process. Note that we choose to randomly select the N words from the reference. The random reference word selection is simple and yields good performance empirically.

Formally, given the input X , its predicted sentence \hat{Y} and its reference Y , the goal of glancing sampling function $\mathbb{GS}(Y, \hat{Y})$ is to obtain a set of sampled words S , where S is a subset of Y :

$$\mathbb{GS}(Y, \hat{Y}) = \mathcal{R}(Y, N(Y, \hat{Y})) \quad (5)$$

Here, $\mathcal{R}(Y, N(Y, \hat{Y}))$ means randomly selecting N words from Y , and N is computed as:

$$N(Y, \hat{Y}) = f_{\text{ratio}} \cdot d(Y, \hat{Y}) \quad (6)$$

where $d(Y, \hat{Y})$ is a metric for measuring the differences between Y and \hat{Y} . We adopt the Hamming distance (Hamming, 1950) as the metric, which is computed as $d(Y, \hat{Y}) = \sum_{t=1}^T (y_t \neq \hat{y}_t)$. With $d(Y, \hat{Y})$, the sampling number can be decided adaptively considering the training status of the model.

Note that $d(Y, \hat{Y})$ could be other distances such as Levenshtein distance (Levenshtein, 1966), but we find the Hamming distance achieves the best result empirically. Additionally, to better control the glancing sampling process, we include a hyper-parameter f_{ratio} to adjust the number of sampled words more flexibly.

4 Experiments

In this section, we first introduce the settings of our experiments, then report the main results compared with several strong baselines. Ablation studies and further analysis are also included to verify the effects of different components used in GLAT.

4.1 Experimental Settings

Datasets We conduct experiments on three machine translation benchmarks: WMT14 EN-DE (4.5M translation pairs), WMT16 EN-RO (610k translation pairs), and IWSLT16 DE-EN (150K translation pairs). These datasets are tokenized and segmented into subword units using BPE encodings (Sennrich et al., 2016). We preprocess WMT14 EN-DE by following the data preprocessing in Vaswani et al. (2017). For WMT16 EN-RO and IWSLT16 DE-EN, we use the processed data provided in Lee et al. (2018).

Distillation Following previous work (Gu et al., 2018; Lee et al., 2018; Wang et al., 2019), we also use sequence-level knowledge distillation for all datasets. We employ the transformer with base setting in Vaswani et al. (2017) as the teacher for knowledge distillation. Then, we train our models on distilled data for each task.

Inference GLAT only modifies the training procedure and performs one-iteration non-autoregressive generation as the vanilla NAT in Gu et al. (2018). Before decoding, GLAT first predict the target lengths for outputs and the length prediction is implemented as in Ghazvininejad et al. (2019). An additional [LENGTH] token is add to the source input, and the encoder output for the [LENGTH] token is used to predict the length.

Besides generation with one predicted target length, we also consider the common practice of noise parallel decoding (Gu et al., 2018; Lee et al., 2018; Guo et al., 2019a; Wang et al., 2019), which generates several decoding candidates in parallel and selects the best via re-scoring with a pre-trained autoregressive model. For GLAT, we first predict m target length candidates, then generate output sequences with argmax decoding for each target length candidate. Then we use the pre-trained transformer to rank these sequences and identify the best overall output as the final output.

Implementation We adopt the vanilla model which copies source input uniformly in Gu et al. (2018) as our base model (NAT-base) and replace the *UniformCopy* with attention mechanism using positions. For WMT datasets, we follow the hyperparameters of the base Transformer in Vaswani et al. (2017). And we choose a smaller setting for IWSLT16, considering that IWSLT16 is a smaller dataset. For IWSLT16, we use 5 layers for encoder and decoder and set the model size d_{model} to 256. We train the model with batches of 64k/8k tokens for WMT/IWSLT datasets, respectively. We use Adam optimizer (Kingma & Ba, 2014) with $\beta = (0.9, 0.999)$. For WMT datasets, the learning rate warms up to $5e - 4$ in 4k steps and gradually decays according to inverse square root schedule in Vaswani et al. (2017). As for IWSLT16 DE-EN, we adopt linear annealing (from $3e - 4$ to $1e - 5$) as in Lee et al. (2018). For the hyper-parameter f_{ratio} , we adopt linear annealing from 0.5 to 0.3 for WMT datasets and a fixed value of 0.5 for IWSLT16. The final model is created by averaging the 5 best checkpoints chosen by BLEU scores on the validation set.

Competitors We compare our method with strong representative baselines, including fully non-iterative models: our vanilla NAT-base model, the NAT with fertility (Gu et al., 2018, NAT-FT), the NAT imitating AT (Wei et al., 2019, imit-NAT), the Flow-based NAT (Ma et al., 2019, Flowseq), the NAT with hint-based training (Li et al., 2019, NAT-HINT), Imputer (Saharia et al., 2020), the NAT with CRF (Lafferty et al., 2001) decoding (Sun et al., 2019, NAT-DCRF), and the NAT with iterative refinement: NAR-IR (Lee et al., 2018), LevT (Gu et al., 2019), Mask-Predict (Ghazvininejad et al., 2019), and JM-NAT (Guo et al., 2020). For all our tasks, we obtain other NAT models' performance by directly using the performance figures reported in their papers if they are available on our datasets.

Table 1: Performance on WMT14 EN-DE/DE-EN and WMT16 EN-RO/RO-EN benchmarks. I_{dec} is the number of decoding iterations and m is the number of reranking candidates.

Models		I_{dec}	WMT14		WMT16		Speed Up
			EN-DE	DE-EN	EN-RO	RO-EN	
AT Models	Transformer (Vaswani)	N	27.30	/	/	/	/
	Transformer (ours)	N	27.48	31.27	33.70	34.05	1.0×
Iterative NAT	NAT-IR	10	21.61	25.48	29.32	30.19	1.5×
	LevT	6+	27.27	/	/	33.26	4.0×
	Mask-Predict	10	27.03	30.53	33.08	33.31	/
	JM-NAT	10	27.31	31.02	/	/	5.7×
Non-iterative NAT	NAT-FT	1	17.69	21.47	27.29	29.06	15.6×
	imit-NAT	1	22.44	25.67	28.61	28.90	18.6×
	NAT-HINT	1	21.11	25.24	/	/	30.2×
	Flowseq	1	23.72	28.39	29.73	30.72	/
	Imputer	1	25.8	28.4	/	/	/
	NAT-base (ours)	1	20.36	24.81	28.47	29.43	15.3×
	GLAT (ours)	1	25.21	29.84	31.19	32.04	15.3×
Non-iterative NAT w/ Reranking	NAT-FT (m=100)	1	19.17	23.20	29.79	31.44	2.4×
	imit-NAT (m=7)	1	24.15	27.28	31.45	31.81	9.7×
	NAT-HINT (m=9)	1	25.20	29.52	/	/	17.8×
	Flowseq (m=30)	1	25.31	30.68	32.20	32.84	/
	NAT-DCRF (m=9)	1	26.07	29.68	/	/	6.1×
	GLAT (m=7, ours)	1	26.55	31.02	32.87	33.51	7.9×

4.2 results

The main results on the benchmarks are presented in Table 1. Obviously, GLAT significantly improves the translation quality and outperforms strong baselines by a large margin. Our method introduces explicit dependency modeling for the decoder and gradually learns simultaneous generation of whole sequences, enabling the model to better capture the underlying data structure. Compared to models with iterative decoding, our method completely maintains the inference efficiency advantage of fully non-autoregressive models, since GLAT generate with only one-iteration. Compared with the competitors, we will highlight our empirical advantages:

- The performance of GLAT is surprisingly good. Compared with the vanilla NAT-base models, GLAT obtains significant improvements (about 5 BLEU) on EN-DE/DE-EN. Additionally, GLAT also outperforms other fully non-autoregressive models with a substantial margin (almost +3 BLEU score on average). The results are even very close to those of the AT model, which shows great potential.
- GLAT is simple and can be applied to other NAT models flexibly, as we only modify the training process by reference glancing while keeping inference unchanged. For comparison, imitate-NAT introduces additional AT models as teachers; NAT-DCRF utilizes CRF to generate sequentially; NAT-IR and Mask-Predict models need multiple decoding iterations.
- Note that Imputer and GLAT use different methods to determine the best target length. Based on CTC (Graves et al., 2006), Imputer sets the max target length twice the length of the source input and determines the best length by removing blanks and contiguous repetitive words after generation. Thus, it is non-trivial to apply target length reranking in Imputer, while GLAT can be further improved from 25.2 to 26.5 with AT reranking on WMT14 EN-DE, which outperforms the Imputer model.

We also present a scatter plot in Figure 3, displaying the trend of speed-up and BLEU scores with different NAT models. It is shown that the point of GLAT is located on the top-right of the competing methods. Obviously, GLAT outperforms our competitors in BLEU if speed-up is controlled, and in speed-up, if BLEU is controlled. This indicates that GLAT outperforms previous state-of-the-art NAT methods. Although iterative models like Mask-Predict achieves competitive BLEU scores, they only maintain minor speed advantages over AT. In contrast, fully non-autoregressive models remarkably improve the inference speed.

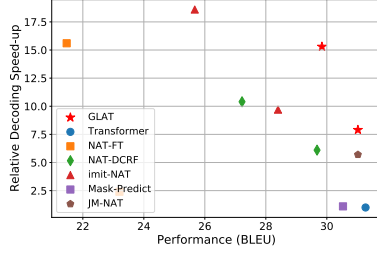


Figure 3: The trade-off between speed-up and BLEU on WMT14 DE-EN

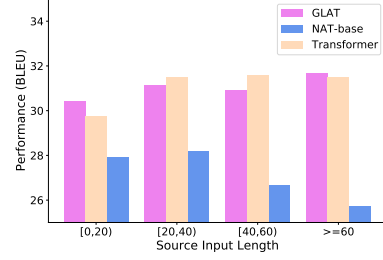


Figure 4: Performance under different source input length on WMT14 DE-EN

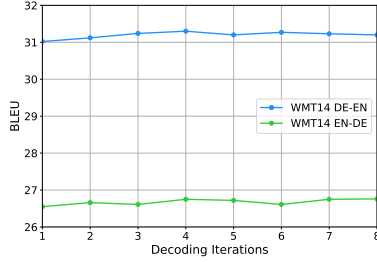


Figure 5: The BLEU scores of GLAT with different decoding iterations

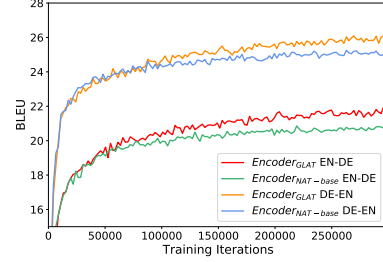


Figure 6: Training NAT with different initialized encoder

4.3 Analysis

Effect of Source Input Length To analyze the effect of source input length on the models' performance, we split the source sentences into different intervals by length after BPE and compute the BLEU score for each interval. The histogram of results is presented in Figure 4. NAT-base's performance drops sharply for long sentences, while the gradual learning process enables GLAT to boost the performance by a large margin, especially for long sentences. We also find that GLAT outperforms autoregressive Transformer when the source input length is smaller than 20.

GLAT Achieves Strong Results without Multiple Iterations We conduct experiments of GLAT with more than one decoding iteration in inference. We adopt the inference algorithm in Ghazvininejad et al. (2019) for multiple-iteration decoding. The results are shown in Figure 5. We find that GLAT can achieve decent performances with only one decoding iteration, while further iterations only obtain minor improvements of 0.2~0.3 BLEU.

GLAT Learns Better Encoder To validate whether GLAT strengthens encoders, we train NAT models with different initialization for comparison. We choose NAT-base and GLAT for comparison and the only difference between them is the training procedure. Specifically, we initialize the encoders of NAT models with the parameter of trained GLAT and trained NAT-base, respectively. The parameters of encoders are fixed during training. The models are all trained using the traditional MLE loss, and the results are presented in Figure 6. Obviously, models initialized with encoders of GLAT outperform those with NAT-base encoders, indicating that GLAT learns better encoders.

Training Speed of GLAT Training on 8 V100 GPUs for WMT14 EN-DE, NAT-base takes about 45 hours and GLAT takes about 56 hours. Compared to NAT-base, the training of GLAT is about 1.2 times slower. Because we only add one forward pass of the decoder, the overhead is relatively small.

4.4 Ablation Study

Effectiveness of the Adaptive Sampling Number To validate the effectiveness of the adaptive sampling strategy for the sampling number $N(Y, \hat{Y})$, we also introduce two fixed approaches for comparison. The first one decides the sampling number with $\lambda * T$, where T is the length of Y , and

Table 2: Performances on IWSLT16 with fixed sampling ratio.

Sampling Method	λ	BLEU
Fixed	0.0	24.66
	0.1	24.91
	0.2	27.12
	0.3	24.98
	0.4	22.96
Adaptive	-	29.61

Table 3: Performances on IWSLT16 with decreasing sampling ratio.

Sampling Method	Schedule		BLEU
	λ_s	λ_e	
Fixed	0.5	0	27.80
	0.5	0.1	28.21
	0.5	0.2	27.15
	0.5	0.3	23.37
Adaptive	-	-	29.61

Table 4: Performance on WMT14 EN-DE with different reference word selection strategies.

Selection Strategy	random	p_{ref}	$1 - p_{\text{ref}}$	most certain	most uncertain
GLAT	25.21	24.87	25.37	24.99	24.86
GLAT (w/ reranking m=7)	26.55	25.83	26.52	26.22	26.13

Table 5: Ablation study on WMT14 EN-DE and WMT14 DE-EN.

	WMT14 EN-DE	WMT14 DE-EN
GLAT w/ sampling strategy of Mask-Predict	19.16	23.56
GLAT w/ decoder inputs of Mask-Predict	24.99	29.48
GLAT	25.21	29.84

λ is a constant ratio. The second one is relatively flexible, which sets a start ratio of λ_s and a end ratio λ_e , and linearly reduce the sampling number from $\lambda_s * T$ to $\lambda_e * T$ along the training process.

As shown in Table 2 and Table 3, clearly, our adaptive approach (Adaptive in the table) outperforms the competitors with big margins. The results confirm our intuition that the sampling schedule affects the generation performance of our NAT model. The sampling strategy, which first offers relatively easy generation problems and then turns harder, benefits the final performance. Besides, even with the simplest constant ratio, GLAT still achieves remarkable results. When set $\lambda = 0.2$, it even outperforms the baseline $\lambda = 0.0$ by 2.5 BLEU score.

The experiments potentially support that it is beneficial to learn the generation of fragments at the start and gradually transfer to the whole sequence. The flexible decreasing ratio method works better than the constant one, and our proposed adaptive approaches achieve the best results.

Influence of Reference Word Selection To analyze how the strategies of selecting reference words affect glancing sampling, we conduct experiments with different selection strategies. By default, we assume all the words in the reference are equally important and randomly choose reference words for glancing. Besides the random strategy, we devise four other selection methods considering the prediction of first decoding. For p_{ref} and $1 - p_{\text{ref}}$, the sampling probability of each reference word is proportional to the output probability for the reference word p_{ref} and the probability $1 - p_{\text{ref}}$, respectively. Similar to the word selection strategy for masking words during inference in Mask-Predict, we also add two strategies related to the prediction confidence: "most certain" and "most uncertain". We choose the positions where predictions have higher confidence for "most certain", and vice versa for "most uncertain". The results for different selection methods are listed in Table 4.

In comparisons, the model with the selection strategy $1 - p_{\text{ref}}$ outperforms the one with p_{ref} , indicating that words hard to predict are more important for glancing in training. And we find that the random strategy performs a little better than the two confidence-based strategies. We think this indicates that introducing more randomness in sampling could enable GLAT to explore more dependencies among target words. We adopt the random strategy for its simplicity and good performance.

Advantages of GLAT over Mask-Predict To study the effects of sampling strategy and decoder inputs of GLAT, we report the results of replacing these two modules in GLAT with the corresponding

Table 6: Performance on WMT14 EN-DE and WMT14 DE-EN with different distances.

	WMT14 EN-DE		WMT14 DE-EN	
	Hamming	Levenshtein	Hamming	Levenshtein
GLAT	25.21	24.56	29.84	28.96
GLAT (w/ reranking m=7)	26.55	26.21	31.02	30.85

part in Mask-Predict, respectively in Table 5. GLAT employs glancing sampling strategy instead of the uniform sampling strategy used in Mask-Predict, and replaces the [MASK] token with source representation from the encoder. The results show that the glancing sampling strategy outperforms the uniform sampling strategy by 5~6 BLEU scores, and feeding representations from the encoder as the decoder input could still improve the strong baseline by 0.2~0.3 BLEU scores after adopting glancing sampling. To sum up, the adaptive glancing sampling approach contributes the most to the final improvement, and the use of representations from the encoder also helps a bit.

Comparison of Different Distances for Glancing Sampling We conduct experiments with two distances for comparing the predictions of the first decoding and references, and the results are presented in Table 6. Experimental results show that both distances can be used to improve the quality of one-iteration generation, and GLAT with Hamming distance is better than GLAT with Levenshtein distance (Levenshtein, 1966). We think Hamming distance is more strict than Levenshtein distance because only the same words on the corresponding positions are regarded as correct, which is more consistent with the training of GLAT.

5 Related Work

Fully Non-Autoregressive Models A line of work introduces various forms of latent variables to reduce the model’s burden of dealing with dependencies among output words (Gu et al., 2018; Ma et al., 2019; Bao et al., 2019; Ran et al., 2019). Another branch of work considers transferring the knowledge from autoregressive models to non-autoregressive models (Wei et al., 2019; Li et al., 2019; Guo et al., 2019b). Besides, there are also some work that apply different training objectives to train non-autoregressive models (Libovický & Helcl, 2018; Shao et al., 2020; Ghazvininejad et al., 2020) or add regularization terms (Wang et al., 2019; Guo et al., 2019a).

Non-Autoregressive Models with Structured Decoding To model the dependencies between words, Sun et al. (2019) introduces a CRF inference module in NAT and performs additional sequential decoding after the non-autoregressive computation in inference. Deng & Rush (2020) proposes cascaded CRF decoding. Since GLAT only performs one-iteration non-autoregressive generation, our approach is orthogonal to the method proposed in Sun et al. (2019). We can also combine our approach with the structured decoding method.

Non-Autoregressive Models with Iterative Refinement A series of work are devoted to semi-autoregressive models that combine the strength of both types of models by iteratively refining the former outputs. Lee et al. (2018) proposed a method of iterative refinement based on denoising autoencoder. Gu et al. (2019) utilized insertion and deletion to refine the generation. Ghazvininejad et al. (2019) trained the model in the way of the masked language model, and the model iteratively replaces the mask tokens with new outputs. Despite the relatively better accuracy, the multiple decoding iterations vastly reduces the inference efficiency of non-autoregressive models.

6 Conclusion

In this paper, we propose Glancing Transformer with a glancing language model to improve the performance of non-iterative NAT. With the glancing language model, the model starts from learning the generation of sequence fragments and gradually moving to whole sequences. Experimental results show that our approach significantly improves the performance of non-autoregressive machine translation with one-iteration generation. As non-autoregressive models are efficient and have great potential in multiple tasks, we plan to apply our approach to other tasks.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. [arXiv preprint arXiv:1409.0473](#), 2014.
- Yu Bao, Hao Zhou, Jiangtao Feng, Mingxuan Wang, Shujian Huang, Jiajun Chen, and Lei Li. Non-autoregressive transformer by position learning. [arXiv preprint arXiv:1911.10677](#), 2019.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pp. 41–48, 2009.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pp. 1724–1734, 2014.
- Yuntian Deng and Alexander M Rush. Cascaded text generation with markov transformers. [arXiv preprint arXiv:2006.01112](#), 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP-IJCNLP*, pp. 6114–6123, 2019.
- Marjan Ghazvininejad, Vladimir Karpukhin, Luke Zettlemoyer, and Omer Levy. Aligned cross entropy for non-autoregressive machine translation. [arXiv preprint arXiv:2004.01655](#), 2020.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pp. 369–376, 2006.
- Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *ICLR*, 2018.
- Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. In *NeurIPS*, pp. 11179–11189, 2019.
- Junliang Guo, Xu Tan, Di He, Tao Qin, Linli Xu, and Tie-Yan Liu. Non-autoregressive neural machine translation with enhanced decoder input. In *AAAI*, volume 33, pp. 3723–3730, 2019a.
- Junliang Guo, Xu Tan, Linli Xu, Tao Qin, Enhong Chen, and Tie-Yan Liu. Fine-tuning by curriculum learning for non-autoregressive neural machine translation. [arXiv preprint arXiv:1911.08717](#), 2019b.
- Junliang Guo, Linli Xu, and Enhong Chen. Jointly masked sequence-to-sequence model for non-autoregressive neural machine translation. In *ACL*, pp. 376–385, 2020.
- Richard W Hamming. Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160, 1950.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. [arXiv preprint arXiv:1412.6980](#), 2014.
- John D Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.
- Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *EMNLP*, pp. 1173–1182, 2018.
- Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pp. 707–710, 1966.
- Zhuohan Li, Di He, Fei Tian, Tao Qin, Liwei Wang, and Tie-Yan Liu. Hint-based training for non-autoregressive translation. In *EMNLP-IJCNLP*, 2019.

- Jindřich Libovický and Jindřich Helcl. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. In *EMNLP*, pp. 3016–3021, 2018.
- Xuezhe Ma, Chunting Zhou, Xian Li, Graham Neubig, and Eduard Hovy. FlowSeq: Non-autoregressive conditional sequence generation with generative flow. In *EMNLP-IJCNLP*, pp. 4273–4283, Hong Kong, China, November 2019. doi: 10.18653/v1/D19-1437.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.
- Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. Guiding non-autoregressive neural machine translation decoding with reordering information. *arXiv preprint arXiv:1911.02215*, 2019.
- Chitwan Saharia, William Chan, Saurabh Saxena, and Mohammad Norouzi. Non-autoregressive machine translation with latent alignments. *arXiv preprint arXiv:2004.07437*, 2020.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, pp. 1715–1725, 2016.
- Chenze Shao, Jinchao Zhang, Yang Feng, Fandong Meng, and Jie Zhou. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation. In *AAAI*, pp. 198–205, 2020.
- Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. Fast structured decoding for sequence models. In *NeurIPS*, pp. 3016–3026, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.
- Yiren Wang, Fei Tian, Di He, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Non-autoregressive machine translation with auxiliary regularization. In *AAAI*, 2019.
- Bingzhen Wei, Mingxuan Wang, Hao Zhou, Junyang Lin, and Xu Sun. Imitation learning for non-autoregressive neural machine translation. In *ACL*, 2019.