Context-Aware Cross-Attention for Non-Autoregressive Translation

Liang Ding[†*] **Longyue Wang**[‡] **Di Wu**[§] **Dacheng Tao**[†] **Zhaopeng Tu**[‡]
[†]The University of Sydney
ldin3097@uni.sydney.edu.au dacheng.tao@sydney.edu.au
[‡]Tencent AI Lab [§]Peking University
{vinnylywang,zptu}@tencent.com inbath@163.com

Abstract

Non-autoregressive translation (NAT) significantly accelerates the inference process by predicting the entire target sequence. However, due to the lack of target dependency modelling in the decoder, the conditional generation process heavily depends on the cross-attention. In this paper, we reveal a localness perception problem in NAT cross-attention, for which it is difficult to adequately capture source context. To alleviate this problem, we propose to enhance signals of neighbour source tokens into conventional cross-attention. Experimental results on several representative datasets show that our approach can consistently improve translation quality over strong NAT baselines. Extensive analyses demonstrate that the enhanced cross-attention achieves better exploitation of source contexts by leveraging both local and global information.

1 Introduction

Different from autoregressive translation (Bahdanau et al., 2015; Vaswani et al., 2017, AT) models that generate each target word conditioned on previously generated ones, non-autoregressive translation (Gu et al., 2018, NAT) models break the autoregressive factorization and produce the target words in parallel. Given a source sentence $\mathbf{x}$, the probability of generating its target sentence $\mathbf{y}$ with length T is defined by NAT as: $p(\mathbf{y}|\mathbf{x}) = p_L(T|\mathbf{x};\theta)\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x};\theta)$, where $p_L(\cdot)$ is a separate conditional distribution to predict the length of target sequence. As NAT models can predict all tokens independently and simultaneously, recent works have fully investigated their superiority on decoding efficiency (Lee et al., 2018; Ghazvininejad et al., 2019; Gu et al., 2019; Kasai et al., 2020; Sun et al., 2019; Shu et al., 2020; Ran et al., 2019). However, there still exists a gap between AT and NAT models in terms of effectiveness.

In encoder-decoder frameworks, the cross-attention module dynamically selects relevant source-side information (key) given a target-side token (query) (Yang et al., 2020; Wang and Tu, 2020). Through qualitative and quantitative analyses, we found that it is difficult for the NAT decoder to adequately capture the source context due to the lack of autoregressive factorization. As shown in Table 1, when translating the Chinese word "交往", the source context word "女孩" should play a significant role in predicting the candidate word "dating". However, the NAT model inappropriately generates "socializing with", resulting in lexical choice errors. As seen, the AT model gives relatively higher attention weights to local contexts on the source side while the NAT model pays less attention on them (0.15 vs 0.04). We make further statistical analysis in Section 2 to prove the universality of this localness perception problem. Similar to our findings, Li et al. (2019) showed that distributions of cross-attention in NAT models are more ambiguous than those in AT ones.

To alleviate this localness perception problem in NAT, we propose a context-aware cross-attention to model both local and global contexts simultaneously. For local attention, we limit the scope of cross-attention to adjacent tokens surrounding the source word with the maximum alignment probability. We then combine the local attention weights with the original global ones by a gating mechanism (in Section 3).

[*] Work done when interning at Tencent AI Lab.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4396–4402
Barcelona, Spain (Online), December 8-13, 2020

Input	弗兰克 找到 一间 公寓 ， 同时 在 跟 一个 女孩 交往 。
Reference	Frank found an apartment and was **dating** a girl at the same time.
NAT Output	Frank found an apartment and was *socializing with* a girl.
Attention	交往$_{0.68}$ 弗兰克$_{0.18}$ 女孩$_{0.04}$
AT Output	Frank found an apartment and was **dating** a girl.
Attention	交往$_{0.81}$ 女孩$_{0.15}$ 弗兰克$_{0.03}$
Ours Output	Frank found an apartment and was **dating** a girl.
Attention	交往$_{0.69}$ 女孩$_{0.11}$ 弗兰克$_{0.09}$

Table 1: Case study of localness perception problem. “NAT Output” and “AT Output” are generated by NAT and AT models, respectively. “Attention” shows top-3 cross-attention probabilities when generating the target word “dating” or other equivalents.

Experiments are conducted on four commonly-cited datasets on translation task (i.e. WMT16 Romanian⇒English, WAT17 Japanese⇒English, WMT14 English⇒German and WMT17 Chinese⇒English) and show that our approach can consistently improve translation quality by around 0.5 BLEU point over advanced NAT models (in Section 4). Further analyses reveal that our method can enhance abilities of NAT to learn syntactic and semantic information as well as phrase patterns (in Section 5).

2 Localness Perception Problem

To validate our motivation, we conduct a statistical analysis. Following Tu et al. (2014), we employ the locality entropy to measure how the cross-attention concentrate around a source word that corresponds with $\mathbf{y}_t$. As shown in Table 1, when generating the target side word “dating”, the concentrated source word is “交往” according to the maximum probability of attention. And AT’s attention distribution is obviously concentrated than NAT’s, thereby have a lower entropy. In our case, given a sentence pair $\{f_1, f_2, \dots, f_n; e_1, e_2, \dots, e_m\}$, for each decoding position $pos \in [1, m]$, we can obtain a probability distribution $\mathbf{P}^i_{pos} = \{P^i(f_1|pos), \dots, P^i(f_n|pos)\}$ by calculating cross-attention in the i-th decoding layer. Thus, the locality entropy of one certain sentence is LE $= -\frac{1}{6m}\sum_{i\in[1,6]}\sum_{pos\in[1,m]}\mathbf{P}^i_{pos} log_2 \mathbf{P}^i_{pos}$. Finally, we average all sentence-level LE to get the corpus-level one. The lower LE means the more concentrated attention on source-side localness and vice versa.

We compare the locality entropy of NAT and AT models on En-De and Zh-En. As shown in Table 2, the locality entropy “*LE*” of NAT model is higher than that of AT, showing that the localness perception problem in NAT is more severe. With the help of our method (in Section 3), this problem can be alleviated (*LE*↓), leading to better translation quality (BLEU ↑). This observation confirms the universality and side effect of localness perception problem in NAT, validating our hypothesis in Section 1.

Models	**En-De**		**Zh-En**	
	LE	BLEU	*LE*	BLEU
NAT	1.66	27.0	2.65	24.0
+Ours	1.62	27.5	2.51	24.6
AT	1.46	29.2	2.12	25.3

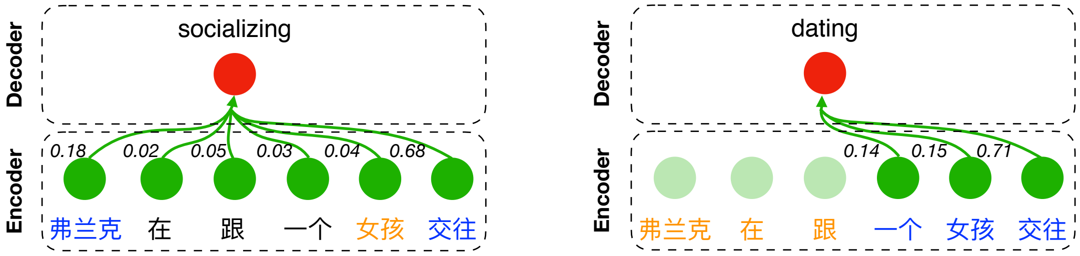
Table 2: The locality entropy *LE* of NAT and AT models as well as our proposed method.

3 Context-Aware Cross-Attention for NAT

In this section, we introduce the detail of our proposed context-aware cross-attention networks (CCAN), which perceives the original and local cross-attention simultaneously.

Original Cross-Attention For the target-side query Q, source-side key K and value V. The i-th original cross-attention ψ_i can be calculated with dot-product: $\psi_i = Q_i K^T$. The original attention of the i-th element is the weighted sum of values $\textsc{Att}(\psi_i, V) = softmax(\psi_i)V$ (in Figure 1(a)).

Our Approach For the i-th position in target side, we propose a locally-sensitive cross-attention component for NAT to capture the neighbor signals. For simplicity, we adopt a straightforward but has been proven effective way (Luong et al., 2015; Xu et al., 2019; You et al., 2020): constricting the attention



(a) Vanilla Non-autoregressive model

(b) Localness modeling in fixed window

Figure 1: Illustration of our proposed approach, which combines (a) vanilla cross-attention and (b) localness-aware cross-attention. In (a), the word “交往” is assigned with the maximum attention weight while the adjacent word (local context) “女孩” is assigned with a low weight. In (b), we guide the model to perceive the local context.

scope to a nearby window around the aligned j-th element. In practice, we choose the source element with the highest attention weight as the aligned element, and the local range can be modeled as follows:

$$L(\psi_i) = \begin{cases} \psi_{i,j} & i - win \leqslant j \leqslant i + win \\ -\infty & otherwise \end{cases} \tag{1}$$

where $\psi_{i,j}$ denotes the attention correlation between the i- and j- elements in encoder and decoder parts, respectively. The win is the hard-coded localness modeling window. Furthermore, we design an interpolation gating mechanism to wisely combine the original and local cross-attention:

$$\text{CCAN}(\underbrace{Q_i}_{\text{Decoder}}, \underbrace{K, V}_{\text{Encoder}}) = g \cdot \textsc{Att}(\psi_i, V) + (1 - g) \cdot \textsc{Att}(L(\psi_i), V) \tag{2}$$

where $g = \sigma(WQ_i)$ is the interpolation weight conditioned on the decoder side query Q_i and $\sigma(\cdot)$ denotes the sigmoid function. Note that W is the only additional parameter to estimate the importance of original cross-attention operation, and we share it for different cross-attention heads.

4 Experiments

4.1 Setup

Data Experiments are conducted on four widely-used translation datasets, including the *small*-scale WMT16 Romanian-English (Ro-En, Gu et al. (2018)), the *medium*-scale WMT14 English-German (En-De, Vaswani et al. (2017)), the *large*-scale WMT17 Chinese-English (Zh-En, Hassan et al. (2018)), and word-order-divergent WAT17 Japanese-English (Ja-En, Morishita et al. (2017)), which consist of 0.6M, 4.5M, 20M and 2M sentence pairs, respectively. We preprocessed data via BPE (Sennrich et al., 2016) with 32K merge operations. We used BLEU (Papineni et al., 2002) as metric with statistical significance test (Collins et al., 2005).

Models We follow Gu et al. (2018) to apply sequence-level knowledge distillation (Kim and Rush, 2016) to simplify the training data. About AT Teachers, we train both BASE and BIG Transformer (Vaswani et al., 2017) models with corresponding training data. In BIG model, we adopt large batch strategy (458K tokens per batch) to optimize the performance. The main results employ Transformer-BIG for all directions except Ro-En, which is distilled by BASE. Our approach can be applied to different NAT architectures. In this paper, we mainly implement it on conditional masked language models (Ghazvininejad et al., 2019, CMLMs) and leave further investigation to future work. The model contains 6-layer encoder and 6-layer decoder, where the decoder trained with conditional mask language model fashion. The model dimension is 512 on 8 heads, with 2048 feed forward dimensions. We follow the common practices (Ghazvininejad et al., 2019; Kasai et al., 2020) to average the top three checkpoints to avoid stochasticity.

4.2 Ablation Study

In order to make best use of our proposed component for NAT, we conducted extensive ablation studies. All models are trained and validated on WMT14 En-De training and validation sets.

#	Models	BLEU	Δ
1	BASE	26.5	–
2	OURS + win 3	26.8	+ 0.3
3	+ win 5	26.8	+ 0.3
4	+ win 7	26.7	+ 0.2
5	+ win 9	26.9	+ 0.4
6	+ win 11	26.7	+ 0.2

#	Models	BLEU	Δ
1	BASE	26.5	–
2	WIN 9 + [1]	26.6	+ 0.1
3	+ [1-3]	26.7	+ 0.2
4	+ [6]	26.8	+ 0.3
5	+ [4-6]	26.8	+ 0.3
6	+ [1-6]	26.9	+ 0.4

Table 3: Effects of localness range (left) and decoder layers (right) on translation quality.

#	Models	Iteration	Ro-En	En-De	Zh-En	Ja-En
	Autoregressive					
1	Transformer-BASE	n/a	34.1	27.3	24.4	29.2
2	Transformer-BIG	n/a	n/a	29.2	25.3	29.8
	Non-Autoregressive					
3	NAT (Gu et al., 2018)	1	31.4	19.2	n/a	n/a
4	Iterative NAT (Lee et al., 2018)	10	30.2	21.6	n/a	n/a
5	DisCo (Kasai et al., 2020)	4.8	33.3	26.8	n/a	n/a
6	Levenshtein (Gu et al., 2019)	2.5	33.3	27.3	n/a	n/a
7	CMLMs (Ghazvininejad et al., 2019)	10	33.3	27.0	23.2	n/a
	Our Implementation					
8	CMLMs	10	33.3	27.0	24.0	28.9
9	+CCAN		33.7	27.5†	24.6†	29.4†

Table 4: Results of proposed method and comparison with previous work on WMT16 Ro-En, WMT14 En-De and WMT17 Zh-En datasets. “†” indicates statistically significant difference ($p < 0.05$) from the CMLM model.

Effects of Localness Range We investigate the localness window size within [3,5,7,9,11] and report the translation performance in Table 3 (left). As seen, our context-aware cross-attention with the window size of 9 achieves the best BLEU, which is therefore used as the default setting.

Effects of Decoder Layers As shown in Table 3 (right), deploying CCAN on the top-layer slightly outperforms deploying on the bottom-layer (“[6]”>“[1]”). In NAT, multiple decoding layers can be cast as the refiner, and the source central word chosen by the bottom-layer cross-attention is not as accurate as of the top-layer one. Our method, highly conditioned on the predicted central words, thus can gain a better effect on the top-layer compared to the bottom layer. In the end, modelling all layers (“[1-6]”) achieves the best performance and we thus use this setting in the following experiments.

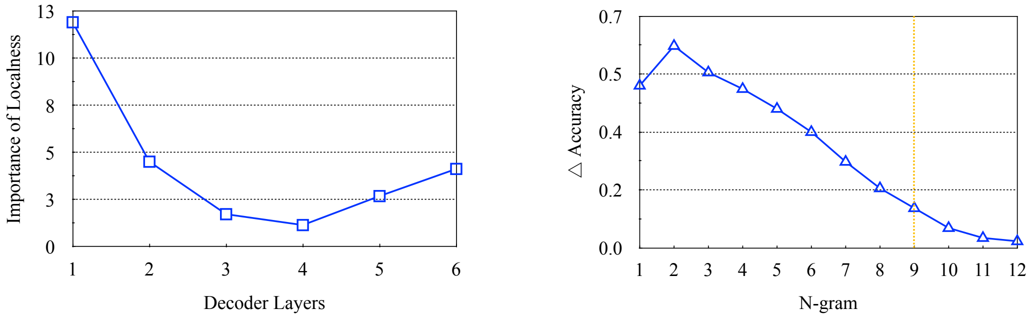
4.3 Main Results

Table 4 lists main results and comparison with previous NAT models on WMT16 Ro-En, WMT14 En-De, WMT17 Zh-En and WAT17 Ja-En datasets. We mainly implemented our approach on top of the advanced CMLMs model. As seen, our approach (Row 9) consistently improves translation performance (BLEU↑) over CMLMs on four language pairs. Note that our approaches only modify the cross-attention module and introduce fewer extra parameters, leading to negligible loss on latency. Encouragingly, our approach even slightly outperforms its AT teachers (Transformer-BASE) on three tasks.

5 Analysis

In this section, we conduct extensive analyses on WMT14 En-De to better understand how our method contribute to performance gains.

Importance of Localness The importance of localness should be different over layers. We explore it through gating (in Equation 2) analyzing. Specifically, we cast the weighting scalar of local cross-attention



(a) Importance of localness on different layers.

(b) Improvement on phrase translation.

Figure 2: Analyses on localness and phrasal patterns.

Model	Surface		Syntactic			Semantic				
	SeLen	WC	TrDep	ToCo	BShif	Tense	SubNm	ObjNm	SoMo	CoIn
CMLMs	**93.2**	**79.4**	45.8	79.4	73.5	**88.9**	87.5	85.4	52.7	**63.1**
+CCAN	92.8	78.1	**46.5**	**79.7**	**74.1**	88.3	**88.1**	**85.7**	**52.9**	62.5

Table 5: Results of probing tasks. We evaluate linguistic properties learned by sentence encoder.

as its importance degree and calculate the importance of localness for each decoder layer. As shown in Figure 2(a), during information flow evolving from bottom to top layers, the importance of localness continues to decline till the penultimate layer, and then increases. The possible reason for the increase in the last two layers is that the top layer followed by softmax, requiring more source-side context to choose lexicons.

Phrasal Patterns Our approach is expected to pay more attention to the most relevant source token and its neighbours, such that the phrasal translation can be improved. To evaluate the accuracy of phrase translations, we calculate the improvement on n-gram tokens in Figure 2(b), where the golden dashed line indicates that the window size is 9. As seen, CCAN consistently outperforms the baseline (ΔAccuracy>0), indicating that our method can enhance the ability of NAT model on capturing the phrasal information, which is similar with Yang et al. (2018)'s findings.

Linguistic Properties Intuitively, our proposed cross-attention component brings context-aware representation, may affecting the linguistic properties learned by the encoder. We quantitatively investigate it from linguistic perspectives with probing tasks (Conneau et al., 2018). These tasks can be categorized into three types: "**Surface**" focuses on the simple surface properties learned from the sentence embedding; "**Syntactic**" quantifies the syntactic reservation ability; and "**Semantic**" assesses the deeper semantic representation ability. To evaluate the representation ability of CCAN equipped NAT model, we compare the pre-trained vanilla NAT and CCAN equipped NAT encoders, followed by a MLP classifier. Specifically, the mean of the top encoding layer, as sentence representation, will be passed to the classifier. We can see from Table 5, the CCAN equipped NAT encoder preserves rich syntactic and semantic information.

6 Conclusion and Future Work

We reveal a localness perception problem in NAT. To alleviate it, we propose the context-aware approach to make the cross-attention pay more attention to source-side local words, which in turn improves the translation performance over several benchmarks. In future work, we will investigate selectively choosing the context (Geng et al., 2020; Yang et al., 2019) rather than the fixed window size. Besides, it is interesting to enhance NAT model with extra signals, such as cross-lingual position embedding (Ding et al., 2020), larger context (Wang et al., 2017) and pre-trained initialization (Liu et al., 2020).

Acknowledgements

This work was supported by Australian Research Council Projects under grants FL-170100117, DP-180103424, and IC-190100031. We are grateful to the anonymous reviewers and the area chair for their insightful comments and suggestions.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*.

Alexis Conneau, German Kruszewski, Guillaume Lample, et al. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *ACL*.

Liang Ding, Longyue Wang, and Dacheng Tao. 2020. Self-attention with cross-lingual position representation. In *ACL*.

Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu. 2020. How does selective mechanism improve self-attention networks? In *ACL*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *ICLR*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *NeurIPS*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. In *arXiv*.

Jungo Kasai, James Cross, Marjan Ghazvininejad, and Jiatao Gu. 2020. Parallel machine translation with disentangled context transformer. In *ICML*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *EMNLP*.

Zhuohan Li, Zi Lin, Di He, Fei Tian, Qin Tao, Wang Liwei, and Tie-Yan Liu. 2019. Hint-based training for non-autoregressive machine translation. In *EMNLP*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. In *arXiv*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2017. Ntt neural machine translation systems at wat 2017. In *IJCNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2019. Guiding non-autoregressive neural machine translation decoding with reordering information. In *arXiv*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Raphael Shu, Jason Lee, Hideki Nakayama, and Kyunghyun Cho. 2020. Latent-variable non-autoregressive neural machine translation with deterministic inference using a delta posterior. In *AAAI*.

Zhiqing Sun, Zhuohan Li, Haoqing Wang, Di He, Zi Lin, and Zhihong Deng. 2019. Fast structured decoding for sequence models. In *NeurIPS*.

Zhaopeng Tu, Zhendong Su, and Premkumar Devanbu. 2014. On the localness of software. In *FSE*.

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 *NeurIPS*.

Wenxuan Wang and Zhaopeng Tu. 2020. Rethinking the value of transformer components. In *COLING*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Mingzhou Xu, Derek F. Wong, Baosong Yang, Yue Zhang, and Lidia S. Chao. 2019. Leveraging local and global patterns for self-attention networks. In *ACL*.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *EMNLP*.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *AAAI*.

Yilin Yang, Longyue Wang, Shuming Shi, Prasad Tadepalli, Stefan Lee, and Zhaopeng Tu. 2020. On the sub-layer functionalities of transformer decoder. In *EMNLP*.

Weiqiu You, Simeng Sun, and Mohit Iyyer. 2020. Hard-coded gaussian attention for neural machine translation. In *ACL*.