Leveraging Local and Global Patterns for Self-Attention Networks

Mingzhou Xu[†] **Derek F. Wong**[†*] **Baosong Yang**[†] **Yue Zhang**[‡] **Lidia S. Chao**[†]
[†]NLP[2]CT Lab / Department of Computer and Information Science, University of Macau
[‡]School of Engineering, Westlake University
nlp2ct.{mzxu,baosong}@gmail.com, {derekfw,lidiasc}@um.edu.mo,
yue.zhang@wias.org.cn

Abstract

Self-attention networks have received increasing research attention. By default, the hidden states of each word are hierarchically calculated by attending to all words in the sentence, which assembles global information. However, several studies pointed out that taking all signals into account may lead to overlooking neighboring information (e.g. phrase pattern). To address this argument, we propose a *hybrid attention mechanism* to dynamically leverage both of the local and global information. Specifically, our approach uses a gating scalar for integrating both sources of the information, which is also convenient for quantifying their contributions. Experiments on various neural machine translation tasks demonstrate the effectiveness of the proposed method. The extensive analyses verify that the two types of contexts are complementary to each other, and our method gives highly effective improvements in their integration.

1 Introduction

Self-attention networks (SANs) (Parikh et al., 2016; Lin et al., 2017) have shown promising results for a range of NLP tasks, including machine translation (Vaswani et al., 2017), contextualized word embedding learning (Devlin et al., 2019), dependency parsing (Kitaev and Klein, 2018) and semantic role labeling (Tan et al., 2018). They learn hidden representations of a sequence by letting each word attend to all words in the sentence regardless of their distances. Such a fully connected structure endows SANs with the appealing strength of collecting the global information (Yu et al., 2018; Shen et al., 2018; Chen et al., 2018; Zhang et al., 2017a; Yang et al., 2019a).

However, some recent researches observe that a fully connected SANs may overlook the important neighboring information (Luong et al., 2015; Sperber et al., 2018; Yang et al., 2019a). They find that SANs can be empirically enhanced by restricting the attention scope to a local area. One interesting question arises: how the local and global patterns quantitatively affect the SANs. To this end, we make empirical investigations with a hybrid attention mechanism, which integrates a local and a global attentive representation via a gating scalar.

Empirical results on English-to-German and Japanese-to-English tasks demonstrate the effectiveness of using both the local and global information, which are shown complementary with each other. Our conceptually simple model consistently improves the performance over existing methods with fewer parameters. The probing tasks demonstrate that the local information is beneficial to the extraction of syntactic features, integrating with the global information further improves the performance on semantic probing tasks. The quantification analysis of gating scalar also indicates that different types of words have different requirements for the local and global information.

2 Related Works

Previous work has shown that modeling locality benefits SANs for certain tasks. Luong et al. (2015) proposed a Gaussian-based local attention with a predictable position; Sperber et al. (2018) differently applied a local method with variable window size for acoustic task; Yang et al. (2018) investigated the affect of the dynamical local Gaussian bias by combining these two approaches for the translation task. Different from these methods using a learnable local scope, Yang et al. (2019b) and Wu et al. (2019) restricted the attention area with fixed size by borrowing the concept of convolution into SANs. Although both these methods yield considerable improvements,

*Corresponding author

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3069–3075
Florence, Italy, July 28 - August 2, 2019.

they to some extent discard long-distance dependencies and the global information. On the contrary, other researchers observed that global feature fusion is one of the salient advantages of SANs. Shen et al. (2018) and Yu et al. (2018) succeeded to employ SANs on capturing global context for their downstream NLP tasks. Recent works also suggested that such the contextual information can improve word sense disambiguation (Zhang et al., 2017a), dependency parsing (Choi et al., 2017) and semantic modeling (Yang et al., 2019a). For exploring the contribution of them, our work integrates both the local and global information under a unified framework.

3 Hybrid Attention Mechanism

In order to quantify the contribution of the local and global patterns, we propose a hybrid attention mechanism. The model first generates the local and global representations (Section 3.1), which are then dynamically integrated into the final output using a gating scalar (Section 3.2).

3.1 Patterns in Attention

Our approach generates the local and global pattern from the same source. As illustrated in Figure 1, for a given input sentence $X = \{x_1, ..., x_n\}$, self-attention model first linearly projects its embedding $H \in \mathbb{R}^{n\times d}$ into queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$ and values $V \in \mathbb{R}^{n\times d}$. The i-th attention energy ξ_i is generated with a dot-product attention algorithm (Luong et al., 2015):

$$\xi_i = \frac{Q_i K^T}{\sqrt{d}} \qquad \in \mathbb{R}^n \qquad (1)$$

Then, the energy is use to produce the local and global attention distribution.

Global Pattern: One strength of SAN is capturing global knowledge by explicitly attending to all the signals. Accordingly, we immediately serve the original attention distribution as the global pattern of our approach. The global representation corresponding to the i-th element is calculated as:

$$Att(\xi_i, V) = softmax(\xi_i)V \qquad \in \mathbb{R}^d \qquad (2)$$

Local Pattern: The local attention enhances the neighbor signals via restricting the attention scope to a local part surrounding the current element. Following Yang et al. (2019b), we employ a hard bias to revise the attention energy for simplification:

$$B(\xi_i) = \begin{cases} \xi_{i,j}, & i-m \le j \le i+m, \\ -\infty, & otherwise. \end{cases} \qquad (3)$$

where $\xi_{i,j}$ denotes the energy between the i- and $j-$th elements. m is the amount of one-side adjacent signals considered in local attention.

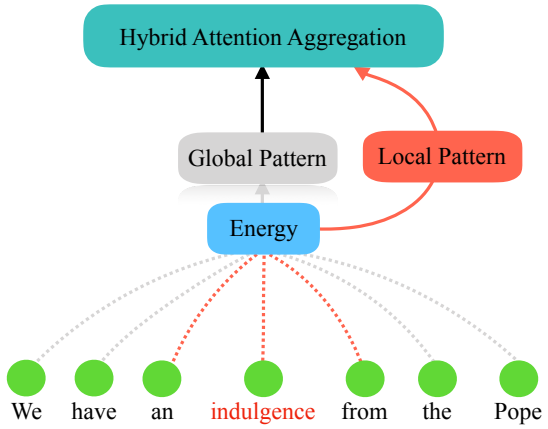

Figure 1: Illustration of hybrid attention mechanism. The global pattern attends to all signals (both grey and red) in the given sentence while the local pattern merely focuses on the neighboring information (red) surrounding the current word "indulgence" (Q_i).

3.2 Hybrid Attention Aggregation

To leverage the local and global information from the two patterns, we apply a gating scalar to dynamically integrate them to the final representation, which can be formally expressed as:

$$\begin{aligned} \hat{H}_i = &(1-g_i) * Att(\xi_i, V_i) + \\ &g_i * Att(B(\xi_i), V_i) \end{aligned} \qquad (4)$$

The gating scalar g_i conditions on H_i, namely:

$$g_i = \sigma(W H_i) \qquad \in (0, 1) \qquad (5)$$

where $\sigma(.)$ denotes the logistic sigmoid function. As seen, gating scalar offers the model a possibility to explicitly quantify the contribution of the local and global representations.

4 Experiments

We evaluate the effectiveness of the proposed approach on widely used WMT 14 English-to-German (En-De) and WAT17 Japanese-to-English (Ja-En) translation tasks. For the WAT17 benchmark, we follow (Morishita et al., 2017) to use the

first two sections of WAT17 dataset as the training data, which contains 2M sentences. The Japanese sentences are segmented by the word segmentation toolkit *KeTea* (Neubig et al., 2011). To alleviate the problem of Out-of-Vocabulary, all the data are segmented into subword units using byte-pair encoding (Sennrich et al., 2016) with 32K merge operations. We incorporate the proposed model [1] into the widely used SAN-based framework – TRANSFORMER (Vaswani et al., 2017) and following their network configuration. We refer readers to Appendix A.1 for the details of our data and experimental settings. Prior studies reveal that modeling locality in lower layers can achieve better performance (Shen et al., 2018; Yu et al., 2018; Yang et al., 2018). Therefore, we merely apply the locality model at the lowest two layers of the encoder. According to our empirical results (Section 5.2), we set the window size to 3 (i.e. $m = 1$). The 4-gram case-sensitive NIST BLEU score (Papineni et al., 2002) is used as the evaluation metric.

4.1 Results

In this section, we give the ablation study of the proposed model and compare several existing works upon the same architecture.

Effectiveness of Hybrid Attention Mechanism To make the evaluation convincing, we reproduced the reported results in Vaswani et al. (2017) on the same data as the baseline. We first investigate the effect of the local pattern without the global information. As shown in Table 1, restricting the attention scope to a local part is able to improve the performance of translation task, showing the effectiveness of localness modeling. By integrating with the global information, the hybrid models progressively improves the translation quality, confirming that the local and global information are complementary to each other. Specifically, we investigate two combination methods: one uses gating scalar, the other simply concatenates the two sources of information. Obviously, dynamically combining two types of representations using gating scalar outperforms its fixed counterpart (concatenation). It is worth noting that the additional projection layer used in the concatenation method brings additional parameters over the method which using the gating scalar.

[1] Our codes are released at: https://github.com/scewiner/Leveraging

Model	Param.	BLEU
TRANSFORMER	88.0M	27.67
+ NEIGHBOR	+0.4M	27.90
+ LOCAL_H	+0.4M	28.03
+ LOCAL_S	+0.8M	28.11
+ LOCAL_PATTERN	+0.0M	28.13
+ HYBRID (Concate)	+0.3M	28.15
+ HYBRID (Gate)	+0.0M	28.31

Table 1: Results of the re-implemented approaches and our method on En-De translation task. NEIGHBOR (Sperber et al., 2018) and LOCAL_H (Luong et al., 2015) apply Gaussian biases to regularize the conventional attention distribution with a learnable window size and a predicable central position, respectively. LOCAL_S (Yang et al., 2018) is the combination of these two approaches. "Param." denotes the model size.

Model	En-De	Ja-En
TRANSFORMER	27.67	28.10
+ LOCAL_PATTERN	28.13	28.23
+ HYBRID (Gate)	28.31$^{\Uparrow}$	28.66$^{\Uparrow}$

Table 2: Experimental results on WMT17 En⇒De and WAT17 Ja⇒En test sets. "⇑": significant over the vanilla self-attention counterpart ($p < 0.05$), tested by bootstrap resampling (Koehn, 2004).

Comparison to Existing Approaches We re-implement and compare several existing methods (Sperber et al., 2018; Luong et al., 2015; Yang et al., 2018, 2019b) upon TRANSFORMER. Table 1 reports the results on the En-De test set. Clearly, all the models improve translation quality, reconfirming the necessity of modeling locality for SANs. By leveraging the local and global properties, our models outperform all the related works with fewer additional parameters.

Performance across Languages We further conduct experiments on WAT17 Ja-En task, which is a distant language pair (Isozaki et al., 2010). As concluded in Table 2, the proposed hybrid attention mechanism consistently improves translation performance over strong TRANSFORMER baselines across language pairs, which demonstrates the universality of the proposed approach.

5 Analysis

We further investigate how the local and global patterns matter SANs. In this section, we try to answer two questions: 1) which linguistic properties are exactly improved by the proposed method;

Model	Surf.	Sync.	Semc.
TRANSFORMER	76.75	64.67	74.88
+ LOCAL_PATTERN	**77.15**	**66.00**	74.74
+ HYBRID (Gate)	76.25	65.60	**75.14**

Table 3: Classification accuracy on 10 probing tasks of evaluating the linguistic properties. We category 10 probing tasks into three groups ("Surf.": surface, "Sync.": syntax and "Semc.": semantics) following the setting in Conneau et al. (2018). For simplistic, we merely reported the average score on each group.

and 2) how different representations learn the locality and globality.

5.1 Linguistic Properties

Although the proposed model improves the translation performance dramatically, we still lack of understanding on which linguistic perspectives are exactly improved by the two sources of information. To this end, we follow Conneau et al. (2018) and Li et al. (2019) to conduct 10 classification tasks to study what linguistic properties are enhanced by our model.

Experiment Setting These tasks are divided into three categories (Conneau et al., 2018): tasks in "**Surf.**" focus on the surface properties learned in the sentence embedding; "**Sync.**" are the tasks which designed to evaluate the capabilities of the encoder on capturing the syntactic information; and "**Semc.**" tasks assess the ability of a model to understanding the denotation of a sentence.

For the model setting, we replace the decoder of our translation model to a MLP classifier and keep the encoder with the configuration shown in Section 4. The mean of the last encoding layer is passed to the classifier as the sentence representation. We train and examine all the model of each task on the dataset provided by Conneau et al. (2018), which contains 100k sentences for training, 10k sentences for validating and testing, respectively. To quantify the linguistic properties of the pre-trained encoders, the parameters of the encoders are fixed, while merely update those in the output layer. We set the hyper parameters of these tasks following the configuration of Conneau et al. (2018). The mini-batch size is 1k samples. The training of each model early-stops with the accuracy on the validation set. More details of the evaluation setting and accuracy in finer-grained level can be found in Appendix B.

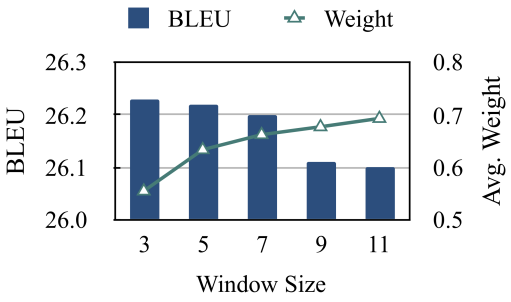

Figure 2: The BLEU scores of the model with different window sizes and their associated weights for the local context. The axis of the histogram (BLEU) is shown in the left, and the right is the axis of the curve (weight). Obviously, the window size being 3 results in the best performance on validation set and the contribution of the local context increases along with the window size.

Results of Probing Tasks As reported in Table 3, our methods outperform baseline model on both "**Sync.**" and "**Semc.**" tasks. Specifically, the local information is obviously more conducive to the "**Sync.**" tasks, which indicates that enhancing the local information in the lower layer could improve the ability to learn the syntactic properties (FitzGerald et al., 2015). Nevertheless, further integrating with the global information benefits to the capturing of the semantic information (Yang et al., 2019a). Moreover, the hybrid model underperforms baseline model on "**Surf.**' tasks, the reason is that a model tends to forget these superficial features for capturing deeper linguistic properties (Conneau et al., 2018; Hao et al., 2019).

5.2 Analysis on Different Representations

We further investigate how the local and global patterns harmonically work with different representations via reporting the average weight output by the gating scalar (Equation 5).

Investigation of Window size Figure 2 depicts the results of our investigations with the different window sizes on the En-De validation set. In order to measure the reliability of the evaluation, we assess each setting via averaging the best 5 models in different training steps. As seen, the model with the window size of 3 (i.e $m = 1$) gets a slight improvement over the others. This is inconsistent with the previous findings (Luong et al., 2015; Yang et al., 2019b) which show that the window size being 11 leads to the best performance. One possible reason is that their models will discard the global information when assigns a small local scope. On the contrary, our hybrid model not

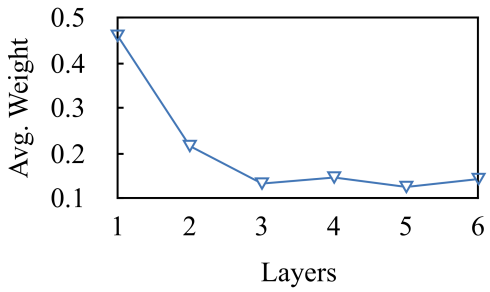

Figure 3: Visualization of the importance of the local information on different layers. The importance is assessed by averaging the scalar factors in Equation 5 over the validation set.

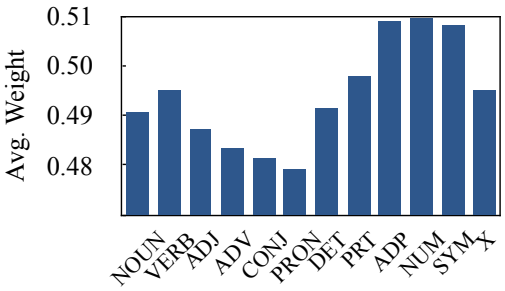

Figure 4: The weights of the local information corresponding to different POS. Obviously, different types of representations need different requirements of the local and global information.

only utilizes the local context but also exploits the global information. Accordingly, the local pattern can attend to a smaller scope without the loss of global context. The hypothesis can be confirmed by the curve regarding to weights of the local pattern. As seen, the requirement of the local information increases with the window size.

Gating Scalar across Layers As visualized in Figure 3, the requirements of the local information are reduced with the stacking of layers. This is consistent with the prior findings that the lower layers tend to learn more word- and phrase-level properties than the higher layers, while the top layers of SANs seek more global information (Peters et al., 2018; Yang et al., 2018; Devlin et al., 2019). Moreover, the local information is less than the global information even in the first layer, verifying our hypothesis that both the local and global patterns are necessary for SANs.

Gating Scalar across POS We further explore how different types of words learn the local information. In response to this problem, we categorize different words in validation set using the Universal Part-of-Speech tagset.[2] Figure 4 shows the averaged factors learned for different types of words at the first layer. As seen, contrary to the content words (e.g, "NOUN","VERB","ADJ"), the function words (e.g, "CONJ" and "PRON"), which have little substantive meaning, seek to more global information in the source sentence. However, we also find that other function words (e.g, "ADP", "NUM","SYM") pay more attention on neighboring signals. We attribute this to the fact that these function words need more local context to determine their syntactic and semantic roles in the sentence. Both these results show that different words indeed have distinct requirements of the local and global information. Therefore, modeling locality and globality in a flexible fashion is necessary for SANs on sentence modeling.

[2]Including: "SYM"-symbols, "DET"- determiner, "CONJ"-conjuntion, "PRT"-partical, "PRON"-pronoun, "ADP"-adposition, "NOUN"-noun, "VERB"-verb, "ADV"-adverb, "NUM"-number, "ADJ"-adjective, and "X"-others.

6 Conclusion

In this study, we propose to integrate the local and global information for enhancing the performance of SANs. Experimental results on various machine translation tasks demonstrate the effectiveness of the proposed model. We further empirically compare the two kinds of contextual information for different types of representations and probing tasks. The extensive analyses verify that: 1) fully leveraging both of the local and global information is beneficial to generate a meaningful representation; and 2) different types of representations indeed have distinct requirements with respect to the local and global information. The proposed method gives highly effective improvements in their integration.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (Grant No. 61672555), the Joint Project of Macao Science and Technology Development Fund and National Natural Science Foundation of China (Grant No. 045/2017/AFJ) and the Multi-Year Research Grant from the University of Macau (Grant No. MYRG2017-00087-FST). Yue Zhang is supported by the startup grant at Westlake University. We would like to thank the anonymous reviewers for their insightful comments.

References

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In *ACL*.

Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. 2017. Context-Dependent Word Representation for Neural Machine Translation. *COMPUT SPEECH LANG*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What You Can Cram into A Single $\&!\#*$ Vector: Probing Sentence Embeddings for Linguistic Properties. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic Role Labeling with Neural Network Factors. In *EMNLP*.

Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019. Modeling Recurrence for Transformer. In *NAACL*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *ICLR*.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with A Self-Attentive Encoder. In *ACL*.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*.

Jian Li, Baosong Yang, Zi-Yi Dou, Xing Wang, Michael R. Lyu, and Zhaopeng Tu. 2019. Information Aggregation for Multi-Head Attention with Routing-by-Agreement. In *NAACL*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Aattentive Sentence Embedding. In *ICLR*.

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 *WAT*.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *EMNLP*.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting Contextual Word Embeddings: Architecture and Representation. In *EMNLP*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. 2018. Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling. In *ICLR*.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. 2018. Self-Attentional Acoustic Models. In *Interspeech*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep Semantic Role Labeling with Self-attention. In *AAAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NISP*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay Less Attention with Lightweight and Dynamic Convolutions. In *ICLR*.

Baosong Yang, Jian Li, Derek Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019a. Context-Aware Self-Attention Networks. In *AAAI*.

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, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019b. Convolutional Self-Attention Networks. In *NAACL*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *ICLR*.

Biao Zhang, Deyi Xiong, Jinsong Su, and Hong Duan. 2017a. A Context-Aware Recurrent Encoder for Neural Machine Translation. *TASLP*.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huanbo Luan, and Yang Liu. 2017b. THUMT: An Open Source Toolkit for Neural Machine Translation. *arXiv:1706.06415*.

Model	Surf.		Sync.			Semc.				
	SeLn	WC	TDep	ToCo	BShif	Tense	SubN	ObjN	SoMo	CoIn
TRANSFORMER	90.1	**63.4**	43.9	78.5	71.6	88.7	85.8	85.0	51.7	**63.2**
LOCAL_PATTERN	**91.3**	63.0	44.8	**78.8**	**74.4**	**88.8**	86.1	84.7	51.8	62.3
HYBRID	89.9	62.6	**44.9**	78.4	73.5	88.5	**87.0**	**85.4**	**52.1**	62.8

Table 4: The classification accuracy of 10 probing tasks. We pass the representations from the last encoding layer to the classifier.

A Machine Translation

A.1 Experimental Setting

We evaluate our method on the advanced TRANSFORMER architecture (Vaswani et al., 2017) that was reproduced by the toolkit *THUMT* (Zhang et al., 2017b). We use the same configuration as Vaswani et al. (2017), in which the hidden size is 512, the number of encoder and decoder layer is 6, the number of head is 8 and the label smoothing is 0.1. Different to Vaswani et al. (2017), we set the L2 regularization to $\lambda = 10^{-7}$. The training of each model was early-stopped to maximize the BLEU score on the development set. The training set is shuffled after each epoch. We use Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate linearly warms up over the first 4,000 steps, and decreases thereafter proportionally to the inverse square root of the step number. We use a dropout rate of 0.1 on all layers. All the models are trained with each batch containing approximately 25000 source tokens and 25000 target tokens.

B Probing Tasks

We conduct 10 classification tasks (Conneau et al., 2018) to study what linguistic properties are enhanced by the proposed model.

B.1 Tasks Description

As seen in Table 4, "SeLn" is to predict the length of a given sentence. "WC" tests whether it is possible to recover information about the word from the sentence embedding. "TDep" checks whether an encoder infers the hierarchical structure of input sentences. In "ToCo" task, sentences should be classified in terms of the sequence of top constituents. "BShif" tests whether two consecutive tokens within the sentence have been inverted. "Tense" is a task for evaluating the tense of the main-clause verb. "SubN" focuses on finding out the number of the subject of the main clause. "ObjN" tests the number of the direct object of the main clause. In "SoMo", a noun or verb of the sentence are replaced with another noun or verb and the classifier should tell whether a sentence has been modified or not. "CoIn" divides a sentence into two coordinate clauses. Half of the sentences are inverted the order of the clauses and the task is to tell whether a sentence is intact or modified.

B.2 Results in Detail

We investigate the performance of the proposed model on probing tasks and list the result in Table 4. As seen, TRANSFORMER which seeks more global information outperforms other models in both the "WC" and "CoIn" tasks. On the contrary, modeling locality is beneficial to "SeLn", "ToCo", "BShif" and "Tense" tasks. By combining these two sources of information, the model with hybrid attention aggregation gets better performance in 3 of the five "**Semc.**" tasks, which demonstrates that leveraging both the local and global information is able to raise the ability of SANs to learn semantic properties. Moreover, HYBRID underperforms the others in "**Surf.**" tasks, which means that this model is more suitable for capturing deeper linguistic properties.