Mixed Multi-Head Self-Attention for Neural Machine Translation

Hongyi Cui[1], **Shohei Iida**[1], **Po-Hsuan Hung**[1], **Takehito Utsuro**[1], **Masaaki Nagata**[2]

[1]Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan

Abstract

Recently, the Transformer becomes a state-of-the-art architecture in the filed of neural machine translation (NMT). A key point of its high-performance is the multi-head self-attention which is supposed to allow the model to independently attend to information from different representation subspaces. However, there is no explicit mechanism to ensure that different attention heads indeed capture different features, and in practice, redundancy has occurred in multiple heads. In this paper, we argue that using the same global attention in multiple heads limits multi-head self-attention's capacity for learning distinct features. In order to improve the expressiveness of multi-head self-attention, we propose a novel Mixed Multi-Head Self-Attention (MMA) which models not only global and local attention but also forward and backward attention in different attention heads. This enables the model to learn distinct representations explicitly among multiple heads. In our experiments on both WAT17 English-Japanese as well as IWSLT14 German-English translation task, we show that, without increasing the number of parameters, our models yield consistent and significant improvements (0.9 BLEU scores on average) over the strong Transformer baseline. [1]

1 Introduction

Neural machine translation (NMT) has made promising progress in recent years with different architectures, ranging from recurrent neural networks (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), convolutional networks (Gehring et al., 2017) and most recently, self-attention networks (Transformer) (Vaswani et al., 2017).

Among the different architectures, the Transformer (Vaswani et al., 2017) has recently attracted most attention in neural machine translation, due to its high parallelization in computation and improvements in quality. A key point of its high-performance is the multi-head self-attention which allows the model to jointly attend to information from different representation subspaces at different positions. There is a huge gap (around 1 BLEU score) between the performance of the Transformer with only one head and eight heads (Vaswani et al., 2017; Chen et al., 2018).

However, all encoder self-attention heads fully take global information into account, there is no explicit mechanism to ensure that different attention heads indeed capture different features (Li et al., 2018). Concerning the results presented by some latest researches, the majority of the encoder self-attention heads, can even be pruned away without substantially hurting model's performance (Voita et al., 2019; Michel et al., 2019). Moreover, the ability of multi-head self-attention, in which lacking capacity to capture local information (Luong et al., 2015; Yang et al., 2018; Wu et al., 2019) and sequential information (Shaw et al., 2018; Dehghani et al., 2019), has recently come into question (Tang et al., 2018).

Motivated by above findings, we attribute the redundancy arising in encoder self-attention heads to the using of same global self-attention among all attention heads. Additionally, it is because of the redundancy, multi-head self-attention is unable to leverage its full capacity for learning distinct features in different heads. In response, in this paper, we propose a novel Mixed Multi-Head Self-Attention (MMA) which can capture distinct features in different heads explicitly by different attention function. Concretely, MMA is composed of four attention functions: Global Atten-

[1]Our code is available at: https://github.com/yokusama/transformer-mma

Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019), pages 206–214
Hong Kong, China, November 4, 2019.
www.aclweb.org/anthology/D19-56%2d

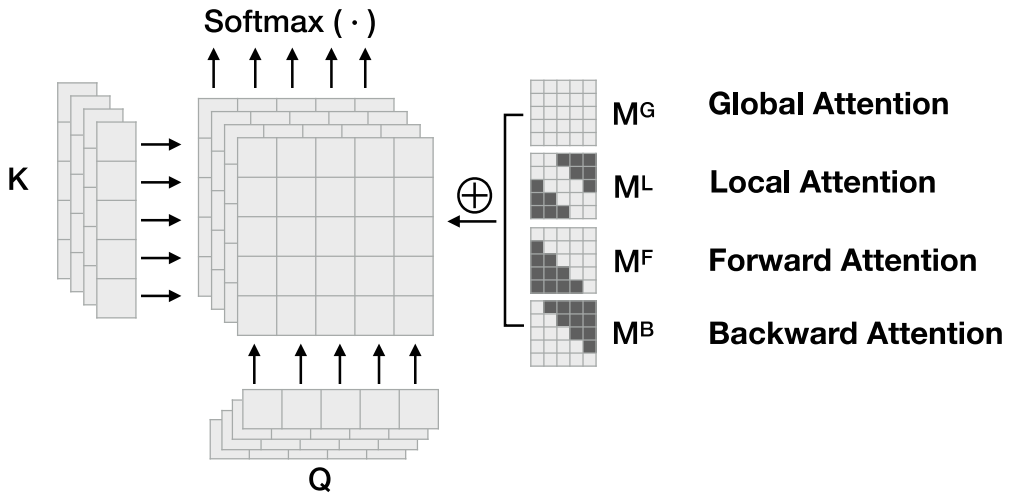

Figure 1: The architecture of Transformer with Mixed Multi-Head Self-Attention

tion which models dependency of arbitrary words directly. Local Attention, where attention scope is restricted for exploring local information. Forward and Backward Attention which attends to words from the future and from the past respectively, serving as a function to model sequence order. MMA enables the model to learn distinct representations explicitly in different heads and improves the expressive capacity of multi-head self-attention. Besides, our method is achieved simply by adding hard masks before calculating attention weights, the rest is the same as the original Transformer. Hence our method does not introduce additional parameters and does not affect the training efficiency.

The primary contributions of this work can be summarized as follows:

- We propose a novel Mixed Multi-Head Self-Attention (MMA) that extracts different aspects of features in different attention heads.
- Experimental results on two language pairs demonstrate that the proposed model consistently outperforms the vanilla Transformer in BLEU scores. Qualitative analysis shows our MMA can make better use of word order information and the improvement in translating relatively long sentence is especially significant.

2 Transformer Architecture

In this section, we briefly describe the Transformer architecture (Vaswani et al., 2017) which includes an encoder and a decoder. The Transformer aims to model a source sentence x to a target sentence y by minimizing the negative log likelihood of the target words.

The encoder consists of N identical layers, each layers has two sublayers with residual connection (He et al., 2016). The first is a multi-head self-attention layer and the second is a position wise fully connected feed-forward network layer:

$$\tilde{H}^l = \mathrm{LN}(H^{l-1} + \mathrm{MA}(Q^{l-1}, K^{l-1}, V^{l-1})) \quad (1)$$

$$H^l = \mathrm{LN}(\tilde{H}^l + \mathrm{FFN}(\tilde{H}^l)) \quad (2)$$

where Q^{l-1}, K^{l-1}, V^{l-1} come from the output of the previous encoder layer H^{l-1}. $\mathrm{LN}(\cdot)$ and $\mathrm{FFN}(\cdot)$ represent layer normalization (Ba et al., 2016) and feed-forward networks.

The multi-head attention $\mathrm{MA}(\cdot)$ linearly project the queries, keys and values h times for different representation of Q, K, V, and computes scaled dot-product attention (Luong et al., 2015) $\mathrm{ATT}(\cdot)$ for each representation. Then these are concatenated and once again projected, the final attentional context is calculated as follows:

$$head_h = \mathrm{ATT}(QW_h^Q, KW_h^K, VW_h^V) \quad (3)$$

$$\mathrm{MA} = \mathrm{Concat}(head_h)W^O \quad (4)$$

where W_h^Q, W_h^K and W_h^V are parameter matrices to transform hidden state into different representation subspaces and W^O is output projection.

ATT$(\cdot)$ is computed by:

$$e_i = \frac{Q_i K^\top}{\sqrt{d}} \tag{5}$$

$$\mathrm{ATT}(Q, K, V) = \mathrm{Softmax}(e_i)V \tag{6}$$

where e_i is the i-th energy and d is the dimension of hidden state.

The decoder is also composed of N identical layers and it contains a third sublayer, which performs attention over the output of the encoder between the self-attention sublayer and feed-forward network sublayer.

3 Proposed Architecture

Our proposed approach is mainly motivated by the fact that redundancy has occurred in multi-heads (Voita et al., 2019; Michel et al., 2019), which limits the capacity of multi-head self-attention. As each self-attention layer has a same global receptive field, this can not guarantee that every head has learned useful features in different subspaces through the same attention function.

To tackle the problem mentioned above, besides global information, we also model local and sequential information for multi-head self-attention by applying local attention, forward attention and backward attention respectively. We refer to it as Mixed Multi-head Self-Attention (MMA), as shown in Figure 1. This is achieved by adding hard mask to each attention head. In this way, Eq.(3) is redefined as:

$$\mathrm{ATT}(Q, K, V) = \mathrm{Softmax}(e_i + M_i)V \tag{7}$$

Since attention weights are calculated by the softmax function, for i-th word, if a mask $M_{i,j} = -\infty$ is added to the j-th position, it means that $Softmax(e_{i,j} + M_{i,j}) = 0$ and there is no attention of Q_i to K_j. On the contrary, if a mask $M_{i,j} = 0$, it means no change in attention function and Q_i attends to and captures relevant information from K_j.

3.1 Global and Local Attention

Global attention and local attention differ in terms of whether the attention is placed on all positions or only a few positions. Global attention is the original attention function in Transformer (Vaswani et al., 2017), and it has a global receptive field which is used to connect with arbitrary words directly. Under our framework, we define the hard mask for global attention as follows:

$$M^G_{i,j} = 0 \tag{8}$$

But global attention may be less powerful and can potentially render it impractical for longer sequences (Luong et al., 2015). On the other hand, self-attention can be enhanced by local attention which focuses more on restricted scope rather than the entire context (Wu et al., 2019; Xu et al., 2019). Based on the above findings, we also define a local attention which simply employs a hard mask to restrict the attention scope by:

$$M^L_{i,j} = \begin{cases} 0, & i - w \le j \le i + w \\ -\infty, & otherwise \end{cases} \tag{9}$$

where w is the attention scope which means, for a given i-th word, it can only attends to the set of words within the window size $[i - w, i + w]$.

We aim to combine the strengths both of global attention and local attention. Towards this goal, we apply global attention and local attention to two distinct attention heads.

3.2 Forward and Backward Attention

As for RNN-based NMT, bidirectional recurrent encoder (Schuster and Paliwal, 1997) is the most commonly used encoder (Bahdanau et al., 2015). It consists of forward and backward recurrent encoding that receive information from both past and future words. However, the Transformer foregoes recurrence and completely relies on predefined position embedding to represent position information. Therefore, it has considerable difficulties in considering relative word order (Shaw et al., 2018).

In order to enhance the ability of position-awareness in self-attention, we present an straightforward way of modeling sequentiality in the self-attention by a forward attention which only attends to words from the future, and a backward attention which inversely only attends to words from the past. The masks in forward and backward attention can be formally defined as:

$$M^F_{i,j} = \begin{cases} 0, & i \le j \\ -\infty, & otherwise \end{cases} \tag{10}$$

$$M^B_{i,j} = \begin{cases} 0, & i \ge j \\ -\infty, & otherwise \end{cases} \tag{11}$$

	En-Ja			Ja-En		
Model	**#Params**	**BLEU**	**Δ**	**#Params**	**BLEU**	**Δ**
Transformer	71M	33.58	–	71M	23.24	–
Transformer MMA	+ 0	34.39††	+ 0.81	+ 0	24.16††	+ 0.92

Table 1: Evaluation results on WAT17 English⇔Japanese translation task. #Params denotes the number of parameters and Δ denotes relative improvement over the Transformer baseline. † † ($p < 0.01$) indicates statistical significance different from the Transformer baseline.

Model	**De-En**
Variational Attention (Deng et al., 2018)	33.30
Pervasive Attention (Elbayad et al., 2018)	34.18
Multi-Hop Attention (Iida et al., 2019)	35.13
Dynamic Convolution (Wu et al., 2019)	35.20
RNMT Fine-tuned (Sennrich and Zhang, 2019)	35.27
Transformer (Vaswani et al., 2017)	34.46
Transformer MMA	35.41††

Table 2: Evaluation results on IWSLT14 De-En. Δ denotes relative improvement over the Transformer baseline. † † ($p < 0.01$) indicates statistical significance different from the Transformer baseline.

With the help of forward and backward attention, we assume that the Transformer can can make better use of word order information.

3.3 Mixed Multi-Head Self-Attention

With different heads applied different attention function and different receptive field, the model is able to learn different aspects of features. To fully utilize the different features, we concatenate all mixed attention heads as in Eq.(4):

$$\mathrm{MA} = \mathrm{Concat}(head_G, head_L, head_F, head_B)W^O$$

where $head_G$, $head_L$, $head_F$, $head_B$ represent head with global attention, local attention, forward attention and backward attention respectively.

Our method only adds hard masks before softmax function, the rest is the same as the original model. Hence our method brings increase the parameters of the Transformer and does not affect the training efficiency.

4 Experiments

4.1 Datasets

To test the proposed approach, we perform experiments on WAT17 English-Japanese and IWSLT14 German-English translation task with different amounts of training data.

WAT17 English-Japanese: We use the data from WAT17 English-Japanese translation task which created from ASPEC (Nakazawa et al., 2017). Training, validation and test sets comprise 2M, 1.8K, 1.8K sentence pairs respectively. We adopt the official 16K vocabularies preprocessed by sentencepiece.[2]

IWSLT14 German-English: We use the TED data from the IWSLT14 German-English shared translation task (Cettolo et al., 2014) which contains 160K training sentences and 7K validation sentences randomly sampled from the training data. We test on the concatenation of tst2010, tst2011, tst2012, tst2013 and dev2010. For this benchmark, data is lowercased and tokenized with byte pair encoding (BPE) (Sennrich et al., 2016).

4.2 Setup

Our implementation is built upon open-source toolkit fairseq[3] (Ott et al., 2019). For WAT17 dataset and IWSLT14 dataset, we use the configurations of the Transformer *base* and *small* model respectively. Both of them consist of a 6-layer encoder and 6-layer decoder, the size of hidden state and word embedding are set to 512. The dimensionality of inner feed-forward layer is 2048 for *base* and 1024 for *small* model. The dropout probability is 0.1 and 0.3 for *base* and *small* model. Models are optimized with Adam (Kingma and Ba, 2014). We use the same warmup and decay strategy for learning rate as Vaswani et al. (2017) with 4000 warmup steps.

[2] https://github.com/google/sentencepiece
[3] https://github.com/pytorch/fairseq

Model	De-En	Δ	Ja-En	Δ
Transformer	34.46	–	23.24	–
- Position Embedding	16.55	–	12.83	–
Transformer MMA	35.41	+ 0.95	24.16	+ 0.92
- Position Embedding	34.66	+ 18.11	23.80	+10.97

Table 3: Results on IWSLT14 De-En and WAT17 Ja-En for effectiveness of learning word order. "- Position Embedding" indicates removing positional embedding from Transformer encoder or Transformer MMA encoder. Δ denotes relative improvement over the counterpart of the Transformer baseline.

During training, we employ label smoothing of value 0.1 (Szegedy et al., 2016). All models are trained on a single NVIDIA RTX2080Ti with a batch size of around 4096 tokens. The $base$ model are trained for 20 epochs, the $small$ model are trained for 45 epochs.

The number of heads are 8 for $base$ model and 4 for $small$ model. We replace multi-head self-attention in the encoder layers by our mixed multi-head self-attention. For a fair comparison, we apply each attention function twice in $base$ model. By doing this, our Transformer MMA have the same number of parameters as the original Transformer.

For evaluation, we use a beam size of 5 for beam search, translation quality is reported via BLEU (Papineni et al., 2002) and statistical significance test is conducted by paired bootstrap resampling method (Koehn, 2004).

4.3 Results

In Table 1 and Table 2, we present the experiment results measured by BLEU on WAT17 and IWSLT14.

On WAT17 English⇒Japanese (En-Ja) and Japanese⇒English (Ja-En) translation task, without increasing the number of parameters, our Transformer MMA outperforms the corresponding baseline 0.81 BLEU score on En-Ja and 0.92 BLEU score on En-Ja.

On IWSLT14 German⇒English (De-En) translation task, our model achieves 35.41 in terms of BLEU score, with 0.95 improvement over the strong Transformer baseline. In order to compare with existing models, we list out some latest and related work and our model also achieves considerable improvements over these results.

Overall, our evaluation results show the introduction of MMA consistently improves the translation quality over the vanilla Transformer, and the proposed approach is stable across different languages pairs.

5 Analysis

5.1 Effectiveness of MMA

Neural machine translation must consider the correlated ordering of words, where order has a lot of influence on the meaning of a sentence (Khayrallah and Koehn, 2018). In vanilla Transformer, the position embedding is a deterministic function of position and it allows the model to be aware of the order of the sequence (Yang et al., 2019). As shown in Table 3, Transformer without position embedding fails on translation task, resulting in a decrease of 17.91 BLEU score. With the help of proposed MMA, the performance is only reduced by 0.75 BLEU score without position embedding, and 18.11 points higher than the Transformer baseline. The same result holds true for a distant language pair Japanese-English where word oder is completely different. When removing position embedding, the Transformer baseline drops to 12.83 BLEU score. However, our model still achieves 23.80 in terms of BLEU score, with 10.97 points improvement over the Transformer counterpart.

From the cognitive perspective, due to the character of local attention which only focuses on restricted scope, the local attention head's dependence on word order information is reduced. In the forward and backward head, directional information is explicitly learned by our forward and backward attention. The above experimental results confirm our hypothesis that, other than global information, Transformer MMA takes local and sequential information into account when performing self-attention function, revealing its effectiveness on utilizing word order information.

5.2 Effect on Sentence Length

Following Bahdanau et al. (2015), we group source sentences of similar lengths to evaluate the performance of the proposed Transformer MMA and vanilla Transformer. We divide our test set

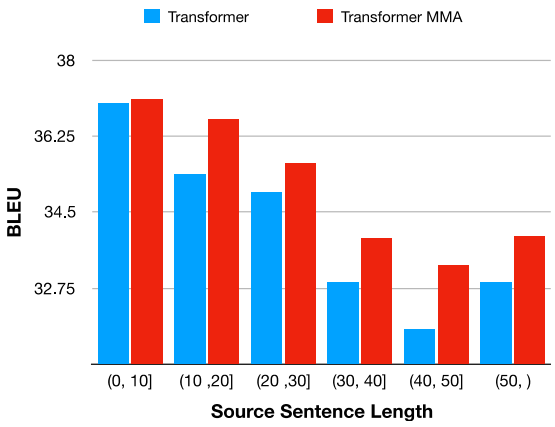

Figure 2: Translation results on test sets relative to source sentence length for IWSLT14 De-En.

into six disjoint groups shown in Figure 2. The numbers on the X-axis represent source sentences that are not longer than the corresponding length, e.g., "(0, 10]" indicates that the length of source sentences is between 1 and 10.

In all length intervals, Transformer MMA consistently outperforms the Transformer baseline. Specifically, as the length of the source sentence increases, so does the increase in the improvement brought by MMA. One explanation is that when the length of the sentence is very short, four different attention functions are similar to each other. But as the length of the sentence increases, more distinct characteristics can be learned and the performance gap is becoming larger.

Moreover, encoding long sentences usually requires more long-range dependency. Concerning the ability to connect with distant words directly, global self-attention was speculated that it is better suited to capture long-range dependency. However, as noted in (Tang et al., 2018), aforesaid hypothesis is not empirically correct and self-attention does have trouble handling long sentences. In case of our Transformer MMA, with the exist of other attention functions served as auxiliary feature extractors, we think that the Transformer has more capacity for modeling longer sentences.

5.3 Ablation Study

For ablation study, the primary question is whether the Transformer benefits from the integration of different attention equally. To do evaluate the impact of various attention functions, we keep global self-attention head unchanged, and next we replace other heads with different attention function.

Model	De-En	Δ
Transformer	34.46	–
+ Local Attention	35.05	+ 0.59
+ Forward Attention	34.83	+ 0.37
+ Backward Attention	35.13	+ 0.67
+ MMA	**35.41**	**+ 0.95**

Table 4: Results of ablation experiments on IWSLT14 De-En. Δ denotes relative improvement over baseline.

Model	De-En	Δ
Transformer	34.46	–
+ MMA (w = 1)	**35.41**	**+0.95**
+ MMA (w = 2)	35.31	+ 0.85
+ MMA (w = 3)	35.35	+ 0.89
+ MMA (w = 4)	35.22	+ 0.76

Table 5: Results of different attention scope on IWSLT14 De-En. Δ denotes relative improvement over baseline.

The results are listed in Table 4. Compared with the Transformer baseline, all integration methods that incorporate other attention function improve the performance of translation, from 0.37 to 0.67 BLEU score. And we can see that Transformer MMA performs best across all variants with the improvement of 0.95 BLEU score.

Furthermore, we investigate the effect of attention scope in our Transformer MMA, as illustrated in Table 5. As the number of attention scope progressively increases, there is no absolute trend in performance. However it is worth noting that when the attention scope is relatively small, the overall performance is better. Specifically, when the size of attention scope is 1, our Transformer MMA achieves the best result. One possible reason is that, in the case where there are already global features captured by global attention, the smaller the attention scope, the more local features can be learned by local attention.

5.4 Attention Visualization

To further explore the behavior of our Transformer MMA, we observe the distribution of encoder attention weights in our models and show an example of Japanese sentence as plotted in Figure 3.

The first discovery is that we find the word overlooks itself on the first layer in the global attention head. This contrasts with the results from Raganato and Tiedemann (2018). They find that, on the first layer of original Transformer, more en-

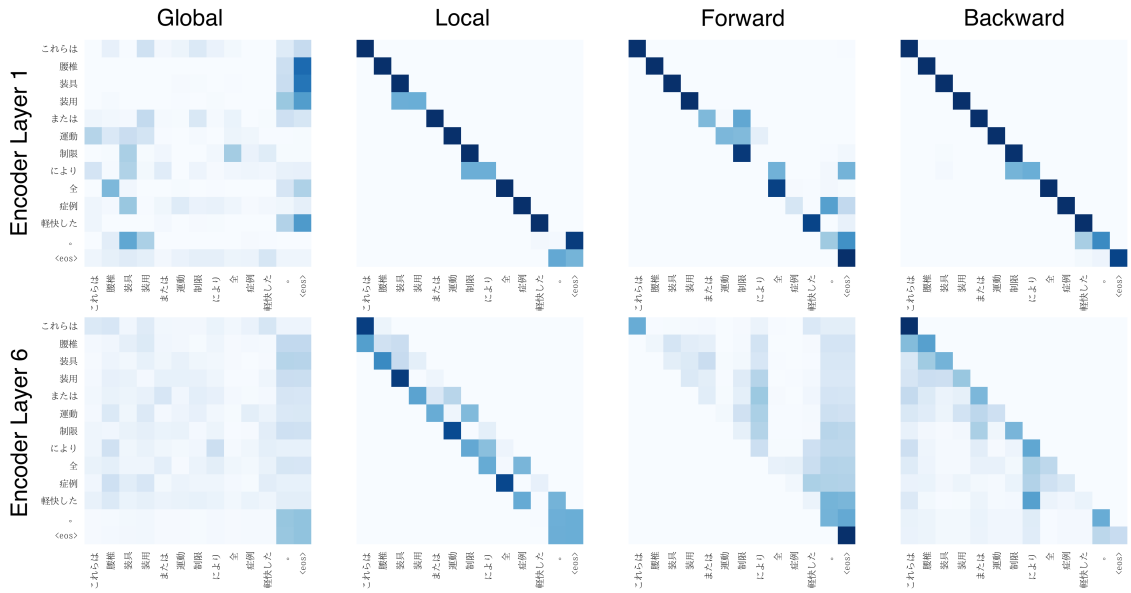

Figure 3: Visualization of the attention weights of Japanese sentence "これらは 腰椎 装具 装用 または 運動 制限 により 全 症例 軽快した 。" (meaning "These persons were improved in all cases by wearing lumbar braces or limiting exercises"). The deeper blue color refers to larger attention weights.

coder self-attention heads focus on the word itself. This change is in line with our assumption that, due to the existence of other attention heads, global attention head can focus more on capturing global information.

The second discovery is that, on the upper layers, forward and backward attention heads move the attention more on distant words. This suggests forward and backward attention is able to serve as a complement to capturing long-range dependency.

6 Related Work

In the field of neural machine translation, the two most used attention mechanisms are additive attention (Bahdanau et al., 2015) and dot attention (Luong et al., 2015). Based on the latter, Vaswani et al. (2017) proposed a multi-head self-attention, that is not only highly parallelizable but also with better performance.

However, self-attention, which employs neither recurrence nor convolution, has great difficulty in incorporating position information (Vaswani et al., 2017). To tackle this problem, Shaw et al. (2018) presented an extension that can be used to incorporate relative position information for sequence. And Shen et al. (2018) tried to encode the temporal order and introduced a directional self-attention which only composes of directional order. On the other hand, although with a global receptive field, the ability of self-attention recently came into question (Tang et al., 2018). And modeling localness, either restricting context sizes (Yang et al., 2018; Wu et al., 2019; Child et al., 2019) or balancing the contribution of local and global information (Xu et al., 2019), has been shown to be able to improve the expressiveness of self-attention. In contrast to these studies, we aim to improve the self-attention in a systematic and multifaceted perspective, rather than just paying attention to one specific characteristic.

Compared to a conventional NMT model with only a single head, multi-head is assumed to have a stronger ability to extract different features in different subspaces. However, there are no explicit mechanism that make them distinct (Voita et al., 2019; Michel et al., 2019). Li et al. (2018) had shown that using a disagreement regularization to encourage different attention heads to have different behaviors can improve the performance of multi-head attention. Iida et al. (2019) proposed a multi-hop attention where the second-hop serves as a head gate function to normalize the attentional context of each head. Not only limited in the field of neural machine translation, Strubell et al. (2018) combined multi-head self-attention with multi-task learning, this led to a promising result for semantic role labeling. Similar to the above studies, we also attempt to model diversity for multi-head attention. In this work, we apply dif-

ferent attention function to capture different aspects of features in multiple heads directly, which is more intuitive and explicit.

7 Conclusion

In this work, we improve the self-attention networks by modeling multi-head attention to learn different aspects of feature through different attention function. Experimental results on WAT17 English-Japanese and IWSLT14 German-English translation tasks demonstrate that our proposed model outperforms the Transformer baseline as well as some latest and related models. Our analysis further shows our Transformer MMA can make better use of word order information and the improvement in translating longer sentences is especially significant. Moreover, we perform ablation study to compare different architectures. To explore the behavior of our proposed model, we visualize the attention distribution and confirm the diversity among multiple heads in MMA.

In the future, we plan to apply our method on other sequence to sequence learning tasks, such as text summarization.

References

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*.

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

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *IWSLT*.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *ACL*, pages 76–86.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *ArXiv*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *NIPS*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction. In *CoNLL*, pages 97–107.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *ICML*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *CVPR*.

Shohei Iida, Ryuichiro Kimura, Hongyi Cui, Po-Hsuan Hung, Takehito Utsuro, and Masaaki Nagata. 2019. Attention over heads: A multi-hop attention for neural machine translation. In *ACL: Student Research Workshop*, pages 217–222.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *ACL: Workshop on Neural Machine Translation and Generation*, pages 74–83.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *EMNLP*, pages 2897–2903.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *ArXiv*.

Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, Hideya Mino, Isao Goto, Hideto Kazawa, Yusuke Oda, Graham Neubig, and Sadao Kurohashi. 2017. Overview of the 4th workshop on Asian translation. In *WAT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *ACL*, pages 211–221.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *NAACL*, pages 464–468.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *EMNLP*, pages 5027–5038.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. In *EMNLP*, pages 4263–4272.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, pages 5797–5808.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *ICLR*.

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*, pages 3069–3075.

Baosong Yang, Zhaopeng Tu, Derek F. Wong, Fandong Meng, Lidia S. Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *EMNLP*, pages 4449–4458.

Baosong Yang, Longyue Wang, Derek F. Wong, Lidia S. Chao, and Zhaopeng Tu. 2019. Assessing the ability of self-attention networks to learn word order. In *ACL*, pages 3635–3644.