Multi-Head Attention for End-to-End Neural Machine Translation

Ivan Fung, Brian Mak

The Hong Kong University of Science and Technology, Hong Kong

{hlfungaa,mak}@cse.ust.hk

Abstract

Inspired by the recent success of Google's Transformer model, works have been done on borrowing the novel idea of multi-head attention to various applications under different architectures. Albeit latest works have adopted this idea using an end-to-end recurrent model on speech recognition and voice search, making use of a similar model on machine translation has not been attempted yet. In this work, we examine multi-head attention under the attention-based recurrent encoder-decoder framework, and conduct detailed analysis on the positional response of multiple heads. Through leveraging the essence of multi-head attention, we are capable of attaining a state-of-the-art result on IWSLT'15 with 28.48 tokenized BLEU and 53.86% TER, which gives a 0.17 gain in BLEU and 0.37% reduction in TER. Similarly we achieve 25.58 tokenized BLEU and 55.03% TER on WMT'16, which provide a 0.40 gain in BLEU and 0.32% reduction in TER to the baseline model, respectively. To the best of our knowledge, this is the first work[1] that evaluates the concept of multi-head attention in an end-to-end recurrent network on machine translation tasks.

Index Terms: neural machine translation, multi-head attention, end-to-end deep learning

1. Introduction

Despite of the long-term dominance of traditional phrase-based statistical machine translation (SMT) model [1, 2] a decade ago, neural methods [3, 4] have now taken over and become the leading trend and direction in the research community owing to their superior qualities in translation results. Recurrent encoder-decoder neural network with attention mechanism [5, 6] has been the most well-known and promising framework prior to the emergence of the recently proposed non-recurrent Transformer model [7]. Among the multiple new strategies used in that model, multi-head attention has great potential of being used in other neural architectures as well. Whilst recurrent multi-head attention-based model has shown a promising gain in the area of speech recognition, whether it is beneficial to machine translation remains unknown. In this work, we apply the multi-head extension to the global soft-attention through concatenation, and carry out investigations into its effect on end-to-end machine translation.

2. Literature Review

Since the idea of multi-head attention was proposed [7], it has been widely adopted in disparate areas under a broad spectrum of architectures. With minor and minimal modification to the original Transformer, it has been proved practical in speech recognition [8], video captioning [9], and even multi-task learning [10] in image captioning, object recognition, speech recognition and machine translation. In view of its effectiveness, it is also natural to incorporate the multi-head extension to other models where self-attention is common, and enhance the performance in tasks such as constituency parsing [11], semantic role labeling [12, 13], question answering [14] and even graph learning [15]. Apart from the abovementioned non-recurrent approaches, works resorting to multi-head attentional recurrent model or its variants have also been observed in speech recognition [16, 17] and voice search [18]. Nevertheless, as far as we know, end-to-end recurrent network with multi-head attention has not been explored and thoroughly studied in machine translation.

[1] We notice a concurrent work along with ours, using a significantly different network architecture: https://arxiv.org/pdf/1804.09849.pdf

3. Attention Mechanism

The majority of attention mechanisms can be categorized into two groups, namely Bahdanau's additive [6] and Luong's multiplicative [5] styles. In this work, we adopt the latter, which is the default and recommended setting in many popular toolkits such as OpenNMT [19] and TensorFlow [20]:

$$score(\mathbf{h}_t, \hat{\mathbf{h}}_s) = \mathbf{h}_t^\top \mathbf{W}_{sc} \hat{\mathbf{h}}_s \quad (1)$$

where $\mathbf{W}_{sc} \in \mathbb{R}^{d_h \times d_h}$, and $\mathbf{h}_t, \hat{\mathbf{h}}_s \in \mathbb{R}^{d_h}$ are the t^{th} and s^{th} hidden state of decoder and encoder with hidden size of d_h, respectively.

3.1. Single-Head Attention

The usual single-head attention can be characterized by the following simple formulas:

$$\alpha_{ts} = \frac{\exp[score(\mathbf{h}_t, \hat{\mathbf{h}}_s)]}{\sum_{s'=1}^{|S|} \exp[score(\mathbf{h}_t, \hat{\mathbf{h}}_{s'})]} \quad (2)$$

$$\mathbf{c}_t = \sum_{s=1}^{|S|} \alpha_{ts} \hat{\mathbf{h}}_s \quad (3)$$

$$\tilde{\mathbf{a}}_t = \tanh(\mathbf{W}_{sm}[\mathbf{c}_t; \mathbf{h}_t]) \quad (4)$$

$$\bar{\mathbf{a}}_t = \tanh(\mathbf{W}_{fd}[\mathbf{c}_t; \mathbf{h}_t]) \quad (5)$$

where $\alpha_{ts} \in \mathbb{R}$ is the attentional weight of the t^{th} decoder hidden state to the s^{th} encoder hidden state; $|S| \in \mathbb{R}$ is the source sentence length; $\mathbf{c}_t \in \mathbb{R}^{d_h}$ is the context vector of the decoder; $\tilde{\mathbf{a}}_t \in \mathbb{R}^{d_{sm}}$ and $\bar{\mathbf{a}}_t \in \mathbb{R}^{d_{fd}}$ are the input to the softmax layer and input to the next decoder step from the t^{th} decoder hidden state, respectively.

Note that $\mathbf{W}_{sm} \in \mathbb{R}^{d_{sm} \times 2d_h}$ and $\mathbf{W}_{fd} \in \mathbb{R}^{d_{fd} \times 2d_h}$ are two independent linear projections, which are different from the original ones as implemented in OpenNMT and TensorFlow.

3.2. Multi-Head Attention

Through replicating the above single-head operations, we can create multiple instances of attention with different parameters,

which can potentially learn disparate representations and information between the same hidden states of encoder and decoder. Following the same idea of concatenation in Transformer, we have:

$$\tilde{\mathbf{a}}'_t = concat([\tilde{\mathbf{a}}^1_t, \dots, \tilde{\mathbf{a}}^h_t]) \quad (6)$$

$$\bar{\mathbf{a}}'_t = concat([\bar{\mathbf{a}}^1_t, \dots, \bar{\mathbf{a}}^h_t]) \quad (7)$$

where $\tilde{\mathbf{a}}^i_t \in \mathbb{R}^{d'_{sm}}$, $\bar{\mathbf{a}}^j_t \in \mathbb{R}^{d'_{fd}}$, $d'_{sm} = d_{sm}/h$, $d'_{fd} = d_{fd}/h$, and h is the number of heads.

Note that dividing by h has the effect of conserving the dimension of $\tilde{\mathbf{a}}'_t$ and $\bar{\mathbf{a}}'_t$.

4. Network Architecture and Training Setup

We employ the standard end-to-end attentional model as adopted in [6], which consists of three parts, namely encoder, decoder and the aforementioned attention mechanism. For both the encoder and decoder, we use the same number of recurrent layers, denoted as N, of long-short term memory cells (LSTM), where the layers in the encoder are bidirectional and those in the decoder are unidirectional. A dropout rate of 0.5 is used. The initial learning rate is 0.1, which is reduced by half starting from the epoch with no improvement on the validation perplexity. Training is stopped when the validation accuracy becomes stable.

5. Experiments

We implement and verify the idea using the OpenNMT [19] toolkit due to its elegance and simplicity, and carry out experiments on the small IWSLT'15 English-to-Vietnamese and large WMT'16 English-to-German corpora without the use of any external monolingual data to strengthen the language model, as described in details below. Results are reported in two popular evaluation matrices, tokenized BLEU and TER.

5.1. Training and Testing Options

In order to allow reproducibility, we provide the key parameters for training each corpus here[2]. Default options are used unless specified otherwise.

5.1.1. IWSLT'15 English-to-Vietnamese

Table 1: *Training options for IWSLT'15 En-to-Vi*

parameter	value
word_vec_size	1024
encoder_type	brnn
enc_layers	2
dec_layers	2
rnn_size	1024
batch_size	32
epochs	28
dropout	0.5

[2]Since the toolkit written in PyTorch does not have any stable release, we used the code downloaded from the official site near the end of March 2018.

5.1.2. WMT'16 English-to-German

Table 2: *Training options for WMT'16 En-to-De*

parameter	value
word_vec_size	1024
encoder_type	brnn
enc_layers	4
dec_layers	4
rnn_size	1024
batch_size	128
epochs	18
dropout	0.5

We provide the parameters for testing both of the corpora here. In particular, we would like to emphasize that the application of length penalty by averaging in beam search is essential and crucial for attaining good results, which reduces the bias in favoring shorter sentences during decoding.

Table 3: *Testing options for both of the datasets*

parameter	value
replace_unk	-
verbose	-
report_bleu	-
length_penalty	avg
beam_size	50

5.2. IWSLT'15 English-to-Vietnamese

This dataset is comprised of 122K of En-to-Vi sentence pairs with a vocabulary size of 17K and 7.7K in the source and target sentences respectively. In order to have a fair comparison with the benchmark in TensorFlow, we have not pre-processed the dataset using other known effective techniques prior to training, and use newstest2012 and newstest2013 as our validation and test set respectively.

Table 4: *Preliminary single-head experiments on the choice of key parameters on IWSLT'15 En-to-Vi; setups leading to inconsistent results in multiple runs are denoted as 'unstable'*

N	d_{sm}	**BLEU**	**TER (%)**
3	1024	27.60	54.59
2	2048	unstable	
2	1024	28.21	54.31
2	512	**28.31**	**54.23**
2	256	unstable	
1	512	26.03	58.70
1 (TensorFlow)	512	26.1	n/a

We have performed multiple experiments on this task, of which the first endeavor is to explore the impact of different parameter combinations on the quality of the translation results, and to come up with a competitive baseline model to start with, prior to utilization of the multi-head attention extension. As shown in Table 4, we experiment on typical parameters including N and d_{sm} while keeping others unchanged, and find that the setup with 2 layers in both encoder and decoder with an input dimension of 512 to the softmax layer performs the best. It attains a BLEU score of 28.31 and TER of 54.23%. We do not

find manipulating d_{fd} helpful and use $d_{fd} = 1024$ throughout all the setups. Note that we are able to obtain results with similar settings that are close to the benchmark documented by TensorFlow.

Table 5: *Effect of the number of heads on IWSLT'15 En-to-Vi [$N = 2$, $d_{sm} = 512$]*

h	**BLEU**	**TER (%)**
1 (Baseline)	28.31	54.23
2	28.24	53.90
4	**28.48**	**53.86**
6	28.17	54.15
8	28.26	54.13
10	28.08	54.68

Our second experiment focuses on the effect of multiple heads on the original attention mechanism. We begin with the previously found setup and denote it as baseline, and increase the number of heads while keeping the dimension of the final concatenated attentional vector unchanged as described above. As exhibited in Table 5, we are able to obtain a state-of-the-art result in the 4-head setup with a BLEU score of 28.48 and TER of 53.86% that represents a gain of 0.17 in BLEU and a reduction of 0.37% in TER when compared to the baseline. On the other hand, increasing the number of heads further does not show additional benefits but only degrades the performance.

Table 6: *Size (number of parameters) of various models on IWSLT'15 En-to-Vi*

N	h	d_{sm}	**size**
3	1	1024	147M
2	1	512	122M
2	4	512	125M

To avoid unfair comparisons due to discrepancy in model size, we list out the parameter size information (excluding the linear layer connecting to softmax) in Table 6 for the selected setups. It confirms that a 3-layer encoder-and-decoder model performs worse even though it has an addition of 22M parameters, and different head combinations do not change the model size significantly.

5.3. WMT'16 English-to-German

This corpus is composed of 4.5M En-to-De sentence pairs with a shared vocabulary size of 37K in the source and target sentences. We follow the same pre-processing procedure as in the TensorFlow recipe, which includes a corpus cleaning and shared byte-pair encoding (BPE) learning stage that involves a total number of 32K iterations, and use newstest2013 and newstest2015 as our validation and test set, respectively.

Table 7: *Selected experiments on WMT'16 En-to-De [$N = 4$, $d_{sm} = 1024$]*

h	**BLEU**	**TER (%)**	**size**
1 (Baseline)	25.18	55.35	205M
4	**25.58**	**55.03**	210M

Since each 4-head setup roughly takes 11 days to run, we do not manage to carry out the experiments with parameter settings as extensively as we did in the previous smaller corpus. Yet we have conducted key experiments on the benefits brought by multi-head attention. As illustrated in Table 7, we are capable of achieving a 25.58 BLEU and 55.03% TER with multi-head attention, which give an absolute gain of 0.40 in BLEU and reduction of 0.32% in TER when compared to the baseline. Notice that our baseline has got a similar tokenized BLEU to [21], which is 25.23 BLEU.

6. Analysis

Irrespective of the success of incorporating multi-head attention in BLEU and TER improvement as unveiled in the previous section, it remains unclear how each head affects and influences the model behind the scene. Consequently, we propose a new way of visualizing the behavior of each head through calculating the sum of attentional weights by their relative positions in the source sentence separately, as a measurement of the relative positional response of each head. As depicted in Figure 1, it can be seen that the worse model on the left tends to have similar response characteristics across all the 4 heads, while the better model on the right does show more diversities and deviations. This gives a strong evidence that multi-head attention can learn different sorts of information, resulting in superior performance; yet practical and reliable ways of enforcing the heads to learn different information from the source sentence will be explored in the future work.

We have also selected representative samples to give insights on the difference in the translation results among models with different number of heads for WMT'16 En-to-De in Table 8 for reference.

7. Multi-Level Attention Extension

We have investigated the use of multi-level attention in the recurrent model, as motivated by works in computer vision that have shown promising results with multi-layer attention [22]. We have carried out two experiments. The first one is the idea of attending the past context vectors with the current word embedding as the query, and feeding the resulted context vector to the next decoder step. This requires a modification to equation (2) as follows:

$$\alpha'_{tt'} = \frac{\exp[score(\mathbf{e}_t, \mathbf{c}_{t'})]}{\sum_{t''=1}^{|T|-1} \exp[score(\mathbf{e}_t, \mathbf{c}_{t''})]} \quad (8)$$

where $\mathbf{e}_t$ and $|T|$ are the current word embedding and number of decoder steps, and equations (3) – (5) can be modified similarly and accordingly.

The second experiment utilizes self-attention immediately after our proposed multi-head attention mechanism with an idea similar to [23]. We combine the context vectors produced by the multiple heads as the matrix H described in that work. Some preliminary experimental results are shown in Table 9. However, we do not find these approaches leading to additional gains, but only harmful to the original architecture instead.

8. Conclusions

We have successfully demonstrated the applicability of the multi-head attention idea under end-to-end attentional model for machine translation, in which we are able to attain a beneficial gain in both IWSLT'15 English-to-Vietnamese and WMT'16 English-to-German corpora in tokenized BLEU. In the future, we are going to examine the effectiveness of the

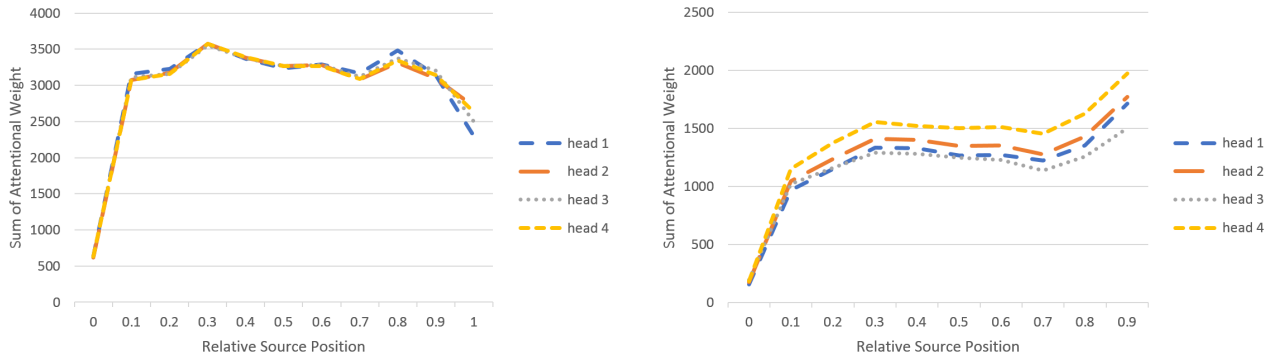

Figure 1: *Response of different heads in the worst and best run of the 4-head model on IWSLT'15 En-to-Vi during testing [Left: 28.22 BLEU; Right: 28.73 BLEU]*

Table 8: *Representative samples generated by various models on WMT'16 En-to-De*

1 head	4 heads	ground truth
Hats off ! ‘ , Stephanie Jenß schrieb über Amula .	Hats off ! ‘ , schrieb Stephanie Jenß über Amula .	Hut ab ! ‘ , schreibt Stephanie Jenß zum Amula .
‘ Sorry , aber für zwölf Euro , habe ich wirklich erwartet mehr ‘ , Melanie Meier kommentiert auf der Seite der Veranstaltung .	‘ Sorry , aber für zwölf Euro , habe ich wirklich mehr erwartet ‘ , melanie Meier kommentiert auf der Veranstaltungsseite .	‘ Sorry , aber für zwölf Euro hat man einfach mehr erwartet ‘ , äußert sich Melanie Meier auf der Veranstaltungsseite .
Wir freuen uns , diese Forderung beantworten zu können , indem wir den ersten Luxus-Service des Vereinigten Königreichs für die Studenten von heute starten .	Wir freuen uns , diese Nachfrage beantworten zu können , indem wir den ersten luxuriösen Reisedienst des Vereinigten Königreichs für die Studenten von heute starten .	Wir freuen uns , auf diese Nachfrage reagieren zu können und den den ersten luxuriösen Reisedienst Großbritanniens für Studenten von heute anzubieten .

Table 9: *Preliminary experiments on multi-level attention on top of our 4-head setup on IWSLT'15 En-to-Vi*

Type	**BLEU**	**TER (%)**
Global	27.02	55.40
Self	28.03	54.55

multi-head attention in other applications, and explore for new ideas to bring further improvement to the attention mechanism.

9. References

[1] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, vol. 1. Association for Computational Linguistics, 2003, pp. 160–167.

[2] P. Koehn, *Statistical machine translation*. Cambridge University Press, 2009.

[3] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1700–1709.

[4] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[6] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[8] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[9] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," *arXiv preprint arXiv:1804.00819*, 2018.

[10] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit, "One model to learn them all," *arXiv preprint arXiv:1706.05137*, 2017.

[11] N. Kitaev and D. Klein, "Constituency parsing with a self-attentive encoder," *arXiv preprint arXiv:1805.01052*, 2018.

[12] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, "Linguistically-informed self-attention for semantic role labeling," *arXiv preprint arXiv:1804.08199*, 2018.

[13] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *AAAI Conference on Artificial Intelligence*, 2018.

[14] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "QANet: Combining local convolution with global self-attention for reading comprehension," in *International Conference on Learning Representations*, 2018.

[15] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[16] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[17] T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Multi-head decoder for end-to-end speech recognition," *arXiv preprint arXiv:1804.08050*, 2018.

[18] T. N. Sainath, C.-C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, and Z. Chen, "Improving the performance of online neural transducer models," *arXiv preprint arXiv:1712.01807*, 2017.

[19] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proc. ACL*, 2017. [Online]. Available: https://doi.org/10.18653/v1/P17-4012.

[20] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[21] D. Britz, A. Goldie, M.-T. Luong, and Q. V. Le, "Massive exploration of neural machine translation architectures," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[22] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[23] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *International Conference on Learning Representations*, 2017.