Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Attention Mechanism in Machine Translation

To cite this article: Yuening Jia 2019 *J. Phys.: Conf. Ser.* **1314** 012186

View the article online for updates and enhancements.

IOP ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

This content was downloaded from IP address 168.91.115.103 on 06/11/2019 at 13:08

IOP Conf. Series: Journal of Physics: Conf. Series **1314** (2019) 012186 doi:10.1088/1742-6596/1314/1/012186

Attention Mechanism in Machine Translation

Yuening Jia

College of Science, Beijing University of Chemical Technology, Beijing, 102200, China

jyn18613@163.com

Abstract. Machine translation is an important field in machine learning. In recent years, the way to improve the effect of machine translation through attention mechanism has been well introduced and developed, and has become an important research part in machine translation. However, how to use the attention mechanism to ensure the translation speed and accuracy at the same time has also become an important problem for researchers to solve. In this article, we will introduce the development process of attention mechanism in the field of machine translation, and introduce in detail the algorithm changes of four papers on the development of Attention in Machine translation. Through the analysis of their experimental data, the development level of attention mechanism in the field of machine translation is shown.

1. Introduction

In recent years, machine translation has developed rapidly and has been widely used. For example, Yelp uses machine translation to translate comments into languages that can be understood by readers[1], and many online translation sites such as Google[2] and Microsoft Bing[3] use machine translation for faster translation.

Traditional statistical machine translation divides long sentences into small blocks[4], which leads to its higher inaccuracy. To solve such problem, neural machine translation was introduced [5] and widely adopted [2][3]. Due to the strong modeling ability of neural networks, neural machine translation has achieved some better performance than traditional machine translation algorithms[6][7][8]. However, conventional NMT based on encoder-decoder models encodes sentences into fixed vectors [9], which limited the ability of translation of long sentences[10]. Therefore, attention mechanism was introduced into neural machine translation.

Bahdanau el al. [11] firstly introduced the concept of attention into machine translation for the first time. They added a layer of attention mechanism to the existing encoder-decoder model, while learning joint alignment and translation, so that part of the input related to the original sentence is automatically selected during prediction.

Luong and Pham studied the NMT architecture based on attention mechanism, and proposed global attention and local attention to further expand the calculation[12].

Lacking of parallelization is a significant disadvantage of RNN structures. Researchers made series of studies of the combination of attention mechanism and CNN [13]. Yin el al. for instance, proposed three ways to use the attention mechanism in CNN, and realized parallel computing. And Bradbury[14] and Kalchbrenner [15] also combine attention mechanism with CNN in there application. However, none of these attempts proved to be improvements to the existing technology on the basis of large benchmark data sets [16].

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Facebook proposed a fully convoluted model combined with attention mechanism for sequence learning [16] and verified their model on large datasets. Experiments show that their algorithm performs better than attention mechanisms with RNN. Moreover, their model is easier to find the composition of sequences and is easier to optimize and accelerate.

CNN has a high computational complexity in building long - range dependencies. Google proposed an algorithm which only makes use of the attention mechanism[17]. Their model not only achieved higher BLEU[18], but also accelerated and simplified parallel training.

In this paper, we summarize a series of classical methods which improved the joint probability of attention mechanism in machine translation, analyze their advantages and disadvantages respectively, and show their performance by comparing their experimental results respectively.

2. Improvement of Attention Mechanism

In this section, we will discuss the improvement of the calculation of the attention mechanism, which makes the attention mechanism improve the joint probability in the neural machine translation model:

$$\mathrm{p}(y_1,\dots,y_{T'}|x_1,\dots,x_T)=\prod_{t=1}^{T'}\mathrm{p}(y_t|v,y_1,\dots,y_{t-1}) \tag{1}$$

Where $x_1,\dots,x_T$ are input sequences, $y_1,\dots,y_{T'}$ are corresponding output sequences,[19]

These changes in machine translation mainly go through the process of initial introduction, algorithm expansion, combination with CNN, independent use, etc. We will introduce them below.

2.1. Learning to Align and Translate

When attention was introduced, Dzmitry Bahdanau and Kyung Hyun CHO defined each conditional probability in RNN as a different context vector c_i for each target word y_i based on the structure of [10][20]. The c_i is calculated as a weighted sum of the annotation sequences h_i to which the encoder maps the input sentence

$$c_i=\sum_{j=1}^{T_X}\alpha_{ij}h_j \tag{2}$$

$$\alpha_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{T_X}\exp(e_{ik})} \tag{3}$$

Where $e_{ij}=a(s_{i-1},\ h_j)$ is an alignment model linking each word in the target language and the source language.

This new context vector c_i, unlike the fixed context vector in the ordinary seq2seq, only needs to consider the more important part of the current decoding. This improvement makes the model not only not need to encode the sentence into a fixed-length sentence, but also focus on the information related to the next target word, thus greatly improving the processing ability of the model for long sentences.

2.2. Local attention and Globe attention

Minh - Thang Luong et al have proposed two models of NMT architecture based on attention mechanism. Their main difference lies in the range of positions contained in the context vector c_i.

The approach in Globe is the same as the idea of Attention proposed in the previous paper, which deals with all words in the source language. But when calculating the value of Attention matrix, they proposes several simple extended calculations.

In Local model, the author first predicts the position PT of the source language end to be aligned at the current decoding according to a prediction function.

$$p_t=t \tag{4}$$

$$p_t=S\cdot sigmoid\left(v_p^T\tanh\left(W_ph_t\right)\right) \tag{5}$$

Then, only the words in the window $[p_t\text{-}D,p_t+D]$ are considered.

When calculating the weight, the author multiplies a Gaussian distribution to reduce the influence of the distant position.

$$\alpha_t = align(h_t, \overline{h_s}) \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right) \quad (6)$$

The paper provides different attention calculations for different situations and improves the effect of attention application.

2.3. ConvS2S

Wen Peng Yin, Hin Rich Schutzee and others proposed three methods of using attentiveness in CNN, respectively trying to add attentiveness mechanism to the two methods of pre - convolution, pooling and merging.They provides us with the idea of applying the Attention mechanism to CNN.

Facebook proposes a fully convoluted sequence-to-sequence model.At each layer of the decoder, attention mechanism is used and the result is input to the next convolution layer. The principle is similar to that of the traditional attention, the attention weight is determined by the output part before the current output of the decoder, the output of the encoder is weighted by the weight, and the obtained ci and hi are added to form a new hi.

The success of this application proves the effect of Attention on CNN and shows the development potential of the combination of Attention and CNN.

2.4. Transformer model

The transformer model proposed by Noam Sha Zeer and Niki Parmar et al. has the following structure:

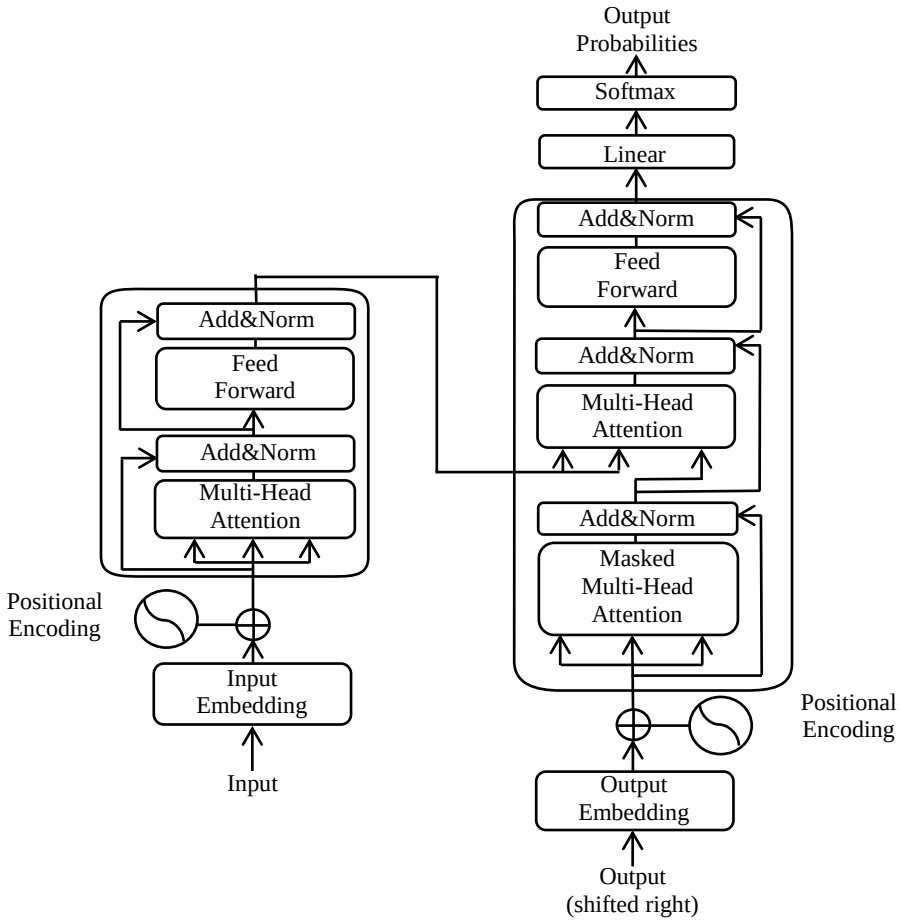

Figure 1. The Transformer-model architecture

They proposed and used multi-head attention:

$$\text{MultiHead}(\text{Q},\text{K},\text{V}) = \text{Concat}(head_1, \dots, head_h)W^o \quad (7)$$

$$\text{where } head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (8)$$

and apply a multi - head attention model of self-interest between that beginning of the encode and each layer of the decoder. The model utilizes the sequence order by position coding.

This model presents the first completely attention-based sequence transformation model, showing the feasibility of attention mechanism in the future.

3. Experimental results

In this section, we compared these different experimental results and analyzed the reasons of them.

3.1. BLEU and Perplexity

In these experiments, we choose the evaluation results of BLEU and Perplexity(PPL) as the comparison index to analyze the above algorithm results.

BLEU is an efficient, low-cost and language-independent translation evaluation method proposed by Kishore Papineni et al[21]. It has a great correlation with the human evaluation, which makes bleu more meaningful for reference. PPL is an index derived from information theory and used to evaluate the quality of a probability distribution prediction result. It can be used to evaluate language probability model in machine translation [22].

3.2. Evaluation results

Table1. Performances of different algorithms on WMT 2014, in terms of BLEU and PPL

MODEL		**BLEU**		**PPL**
		EN-FR	EN-GE	
LEARNING TO ALIGN AND TRANSLATE	RNNencdec-30	24.19		
	RNNsearch-30	31.44		
	RNNencdec-50	26.71		
	RNNsearch-50	34.16		
LOCAL ATTENTION	Base+reverse+dropout		14.0	8.1
	Base+reverse+dropout +local		19.0	5.9
	Base+reverse+dropout+local+feed		20.9	
GLOBE ATTENTION	Base+reverse+dropout+globe		16.8	7.3
	Base+reverse+dropout+globe+feed		18.1	6.4
CONVS2S		41.4	26.43	
TRANSFORMER MODEL	(base mdel)	38.1	27.3	
	(big model)	41.0	28.4	

Table2. Performances of different algorithms on the WMT 15 and WMT 16, in terms of BLEU and PPL

MODEL		**ON THE WMT 15**			**ON THE WMT 16**	
		BLUE		PPL	BLUE	PPL
		EN-FR	GE-EN		EN-RO	
GLOBE ATTENTION	Base (reverse)		16.9	14,3		
	+global (location)		19.1	12.7		
	+global (location)+feed		20.1	10.9		
	+global (dot)+drop+feed		22.8	9.7		
	+global (dot)+drop+feed +unk		24.9			
CONVS2S	Word 80K				29.45	
	BPE 40K				30.02	

In the table 1 and 2, RNNencdec is RNN Encoder-Decoder[20], and RNNsearch is a model proposed by Bahdanau et al. Where 30 and 50 are the largest numbers of words in the translated sentence. Table 1 shows that RNNsearch has a higher BLEU score, especially in dealing with long sentences. Because it is no longer limited by fixed vectors, its translation results are smoother and more accurate.

In the table 1 for local and global attention, Base+reverse+dropout refers to the basic model added by the two attention mechanisms, where Base refers to a vanilla Sequence to Sequence Learning without an attention mechanism. Feed is an inputfeeding approach to take the past alignment information into account. Although these two methods are not significantly improved in bleu and ppl compared with other attention mechanisms, we think they still have great significance because they tell us how to expand the calculation method of attention and the local attention method.

ConvS2S has a significantly higher bleu score than RNN. This is because the multi-hop attention it uses has an attention mechanism for each convolution layer, so that when the model gets the next attention, it can also take into account the words that have already been noticed before. It is the key to improve the effect of the model.

For Transformer Model, we have chosen two models. Big model is obtained by base model by changing the number of decoder layers, dimension and drop speed. From the indicator display, the translation effect of Transformer Model is almost as good as or even better than convolution architecture. This confirms the broader potential of the attention mechanism.

3.3. Sample translations

In this section, we have selected two sample translations from two papers. Through this intuitive way, we showed the attention mechanism to improve the accuracy and smoothness of machine translation, as well as its advantages in long sentence translation.

In Table 3, best model refers to the model in [12] that performs best among various global attention and local attention, and base refers to the basic model without Attention mechanism.

Table3. English-French[11] form Learning to Align and Translate model

SOURCE	In a press conference on Thursday, Mr Blair stated that there was nothing in this video that might constitute a ”reasonable motive” that could lead to criminal charges being brought against the mayor.
REFERENCE	En conférence de presse, jeudi, M.Blair a affirme qu’il n’y avait rien dans cette vidéo qui puisse constituer des” motifs raisonnables” pouvant mener au dépôt d’une accusation criminelle contre le maire.
RNNENC-50	Lors de la conférence de presse de jeudi, M. Blair a dit qu’il n’y avait rien dans cette vidéo qui pourrait constituer une ”motivation raisonnable” pouvant entraıner des accusations criminelles portées contre le maire.
RNNSEARCH-50	Lors d’une conférence de presse jeudi, M.Blair a déclaré qu’il n’y avait rien dans cette vidéo qui pourrait constituer un ”motif raisonnable” qui pourrait conduire à des accusations criminelles contre le maire.

Table4. English-German [12] the best and base model from globe attention

SOURCE	“ We’re pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security” , said Roger Dow , CEO of the U.S. Travel Association .
REFERENCE	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht im Widerspruch zur Sicherheit steht” , sagte Roger Dow , CEO der U.S. Travel Association .
BEST MODEL	“ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit unvereinbar ist” , sagte Roger Dow , CEO der US - die .
BASE	“ Wir freuen uns über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit” , sagte Roger Cameron , CEO der US - <unk> .

4. Conclusion

Nowadays, in an era when artificial intelligence is more and more powerful, machine translation is one of the important sub-fields. How to improve the translation speed while keeping the accuracy of translation is the basic problem that the researchers of machine translation mainly consider.

In this paper, we introduce the development process of attention mechanism in the field of machine translation, and introduce four kinds of research in detail, and compare their experimental results to show the development trend of attention mechanism in the field of machine translation. In addition, reasonably expanding the -machine translation and how to optimize the calculation method of the weight in the attention mechanism have become an important direction to tap the potential of the attention mechanism.

References

[1] https://www.yelp.com/
[2] https://translate.google.cn/
[3] http://cn.bing.com/translator
[4] Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gau-vain. (2006) Continuous space language models for statistical machine translation.In ACL. Sydney. pp. 723–730.
[5] Nal Kalchbrenner, Phil Blunsom. (2013) Recurrent Continuous Translation Models. In EMNL. Seattle, Washington, USA. pp. 1700–1709.
[6] Marcin Junczys-Dowmunt, Tomasz Dwojak, Hieu Hoang. (2016) Is Neural Machine Translation Ready for Deployment?A Case Study on 30 Translation Directions. 13th International Workshop on Spoken Language Translation, Seattle
[7] Jie Zhou, Ying Cao, Xuguang Wang, Peng Li, Wei Xu. (2016) Deep recurrent models with fast-forward connections for neural machine translation. Transactions of the Association for Computational Linguistics, vol. 4, pp. 371–383.
[8] Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, Marcello Federico. (2016) Neural versus Phrase-Based Machine Translation Quality: a Case Study. In EMNL. Austin, Texas. pages 257–267.
[9] Cho, Kyunghyun, Bar van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In EMNLP. Doha, Qatar pp. 1724–1734
[10] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio. (2016) On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In EMNLP. Doha, Qatar. pp. 103–111.
[11] Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio. Neural Machine Tanslation by Jointly Learning to Align and Translate, ICLR, 2015.
[12] Minh-Thang Luong, Hieu Pham and ChristopherD.Manning. (2015) Effective Approaches to Attention-based Neural Machine Translation. In EMNLP. Lisbon, Portugal. pp. 1412-1421.
[13] Wenpeng Yin, Hinrich Schutze, Bing Xiang, Bowen Zhou. (2014) ABCNN: Attention-Based Convolutional Neural Networkfor Modeling Sentence Pairs. Transactions of the Association for Computational Linguistics, vol. 4, pp. 259–27
[14] James Bradbury, Stephen Merity, Caiming Xiong, Richard Socher. (2017) Quasi-Recurrent Neural Networks. In ICLR.
[15] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, Koray Kavukcuoglu. (2016) Neural Machine Translation in Linear Time.
[16] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats and Yann N. (2017) Dauphin. Convolutional Sequence to Sequence Learning. Proceedings of the34thInternational Conference on Machine Learning,Sydney, PMLR 70. Australia.
[17] Ashish Vaswani, Noam Shazeer, Niki Parmar and Jakob Uszkoreit. (2017) Attention is all you need. In NIPS. Long Beach, CA, USA.

[18] Bogdan Babych，Anthony Hartley (2009). Automated error analysis for multiword expressions: using bleu-type scores for automatic discovery of potential translation errors. *Linguistica Antverpiensia, 8*, 81-104.
[19] Ilya Sutskever, Oriol Vinyals and Quoc V. Le. (2014) Sequence to Sequence Learning with Neural Networks. In NIPS.
[20] Kyunghyun Cho, Bart van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In EMNLP. Doha, Qatar. pp. 1724–1734
[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL. Philadelphia. pp. 311-318.
[22] Yao Kaisheng, Zweig G, Peng Baolin. (2015) Attention with intention for a neural network conversation model. In NLPS.