DSWA: A Dilated Shift Window Attention Method for Chinese Named Entity Recognition

Xinyu Hou, Cui Zhu*, Wenjun Zhu

School of Computer Science, Faculty of Information Technology, Beijing University of Technology, Beijing, China

E-mail: houxy18@emails.bjut.edu.cn, {cuizhu, zhuwenjun}@bjut.edu.cn

Abstract

In recent times, numerous models tried to enhance the performance of Transformer on Chinese NER tasks. The model can be enhanced in two ways: one is combining it with lexicon augmentation techniques, the other is optimizing the Transformer model itself. According to research, fully connected self-attention may scatter the attention distribution, which is the reason for worse performance of the original Transformer with self-attention. In this paper, we attempt to optimize the Transformer model especially attention layer. Therefore, a novel attention mechanism, Dilated Shift Window Attention, is proposed to address this problem. By using Window Attention, this method improves the model's capacity to deal local information, meanwhile, the model can still manage long text and long-distance dependencies owing to the Window Dilatation mechanism. Experiments on various datasets also show that DSWA replacing fully connected self-attention improves the model's performance on the Chinese NER task.

Keywords— NLP, Chinese NER, Transformer, Window Attention, Self-attention

1. Introduction

Named entity recognition (NER)[1], an upstream task of natural language processing (NLP)[2], tries to extract entities from the original input text, such as names of people, places, and organizations, etc. It plays an integral role in many downstream NLP tasks.

Transformer[3] has recently excelled at other NLP tasks such as text translation, but in the domain of Chinese NER, the Transformer with self-attention produces unsatisfactory results. As a result, it is more frequently used in combination with other approaches or models. Comparatively speaking, a number of models based on CNN[4] or LSTM[5] have produced favorable results, especially BiLSTM+CRF[6] has been a proven NER solution in daily applications since its stability.

The majority of NLP tasks require researchers to broaden the model's receptive field in order to understand the semantics of sentences or paragraphs. This is also the reason why models are excel in addressing long text comprehension or long distance dependency. A good example is the self-attention mechanism, which, by calculating the attention of the entire vector sequence, enables the model to effectively perceive a wider range of data. But nevertheless, entities, the recognition targets for NER task, frequently appear in the text as a short neighboring sequence like a word or phrase. It is crucial to put more attention to local areas of various ranges and understand the semantics of sentences as well. For example in Figure 1, two entities occupy only a small part of the sentence, attention in local areas may handling them well. Meanwhile, both of them contain the same word "Zhongshan", the first one is a person name and the second one is a place, in order to discrimination, model need the ability to understand the semantics of sentences. As a result, the self-attention that acting on the entire vector sequence may good at understanding, but weak in processing such local information, which may be the reason why the Transformer model with self-attention does dissatisfiedly on Chinese NER task.

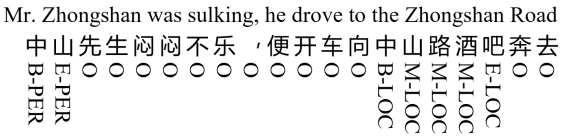


Figure 1: An Example of Chinese NER

In this paper, Dilated Shift Window Attention (DSWA), a novel attention approach, is proposed to solve the following issues. We investigated the reason why Transformer models with self-attention perform poorly on Chinese NER tasks. Additionally, we argue that the superior local information

10.18293/DMSVIVA2023-106

perception capacity of LSTM and CNN, compared to Transformer, make them more suitable for Chinese NER tasks. The new attention approach aims to improve the model's perception of local information while preserving as much of its capacity for long texts as possible. The effectiveness of this novel attention mechanism in enhancing the model's performance on the Chinese NER problem was empirically confirmed.

2. Related Work

2.1. LSTM

The Long-Short Term Memory Network (LSTM), which excels at handling serialized input and has had success in NER tasks. LSTM uses a gate mechanism to parse text as a sequence. Each cell in model processes a character while gathering data from the prior context.

Zhang proposes the lattice-LSTM[7] for the Chinese NER task, which enhances the adaptability of the LSTM model for the Chinese NER task through including lexicon information into the character-level LSTM model and adding an additional cell to encode lexicon information. However, the way the model introduces lexicon information reduces its parallelization efficiency. Hence, Liu proposed the WC-LSTM[8] in an effort to make improvements, so that the lexicon representation brought in for each character is static and stationary, improving the model's parallelization capability.

Furthermore, compared to self-attention, this method is able to process nearby contextual data more effectively. As a result, the LSTM and Transformer in conjunction can supplement the self-attention's ability to comprehend local information[9].

2.2. CNN

In the Chinese NER task, convolutional neural network (CNN) and its modifications also perform well. When dealing with brief text sequences like entities, CNN has a distinct advantage due to its high local information sensing. When applied to the Chinese NER task, CNN is enhanced by the external lexical information, meanwhile adjusting the weights of various words using the feedback layer; this is the LR-CNN[10] proposed by Gui. As a result, this could satisfactorily handle the issue of conflicting lexical information caused by sentence breaks.

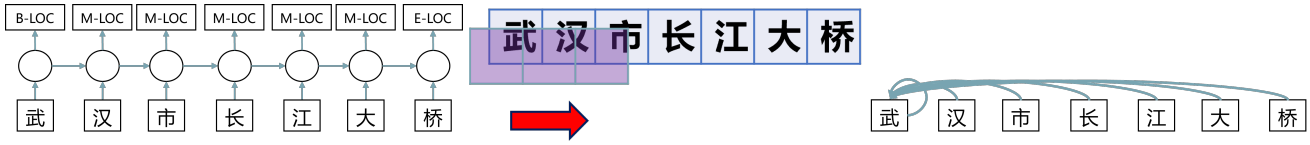
On the other hand, an effort is made to maintain the benefit of CNN's local information sensing capability while also enhancing its global information sensing capability in order to better address the issues associated with lengthy entities and long-distance dependency. With Iterative Dilated Convolution (IDCNN)[4], adding holes to the convolution kernel widens the model's receptive field while maintaining a constant computation. Additionally, IDCNN considers the iterative component. As the number of iteration rounds increases, the holes in the dilated convolutional kernel gradually enlarge and the perceived field exponentially expands. IDCNN requires less computing than traditional pooled convolution and is better at handling global data. It also gave us a fresh concept for enhancing the self-attentive mechanism.

2.3. Transformer

The traditional Transformer[3] model with self-attention performs poorly on the Chinese NER task, in contrast to other NLP domains, despite numerous improvements, which nevertheless lead to a decent score. An excellent illustration is the Simple-Lexicon model, which provides the Transformer with external lexical data. In order to adapt word embeddings to the task features of the Chinese NER, researchers have adopted a technique called "Soft-lexicon"[11] to do so. This method has produced promising results. A lexicon itself is a series of characters used in a variety of contexts, therefore the information contained inside it also serves as local information. The addition of lexicon data gives the model a new way to look at data processing and fills in the gaps left by the character-level self-attention method.

Since the majority of Chinese NER employ character-level models, this approach to applying additional lexicon, known as lexicon enhancement, has several applications and is used by many of the models mentioned above. A popular area of study in recent years has been the use of Transformer and lexicon enhancement in combination. Generally speaking, there are two technical approaches to lexicon enhancement. The first is termed dynamic architecture, and it seeks to create a dynamic architecture that is suitable with lexicon information and character embedding. The FLAT[12] model, which allows Lexicon data to be input into the model along with the character embedding, is a typical example. The Soft-lexicon discussed above is an application of the second method, which is called Adaptive Embedding, it integrates lexicon data into the character embedding without changing the input. Furthermore, because the DSWA replaces self-attention without altering the input, it works well with both two lexicon augmentation approaches. Both the FLAT model employing Dynamic Architecture and the Transformer model combined with Soft-lexicon have been enhanced after applying our DSWA, as you will see in the experimental part that follows.

Besides, other studies have attempted to enhance fully connected self-attention, Biao Hu et al. presented Adaptive Threshold Selective Self-Attention(ATSSA)[13] as such an approach. This method establishes a threshold for the at-



(a) LSTM (b) CNN (c) Self-Attention

Figure 2: Schematic for LSTM, Convolution, and Self-attention

tention score, and only the vectors that are higher than this threshold are used in the computation of attention, while the rest being discarded. In contrast to fully connected self-attention, this strategy constricts the attention distribution, but it's hard to set an appropriate threshold, may omit some information. Besides, the figure of attention visualization in this thesis demonstrates that, attention tends to be focused on the adjacent context in Chinese NER. It provides theoretical support for our approach that based on window attention.

In addition, the Bert-LSTM model is a proven industrial solution for Chinese NER tasks. The Transformer-improved Bert model has become a popular pre-training model for Chinese NER tasks. In addition to enhancing the Transformer model itself, it is also commonly used in combination with other models; a prime example of this is the LSTM-in-Trans model mentioned above. This can significantly increase the model's capacity for sequence modeling.

3. Method

As previously mentioned, the LSTM and CNN is adept in handling a limited range of input. Contrarily, the Transformer model is useful for dealing with long-range dependencies due to the self-attention mechanism used by the original Transformer, which computes attention for each vector with all vectors globally. For instance, Figure 2 is a schematic diagram of comparison between the three models. However, the entities to be identified by the NER task are frequently concentrated within the range of a few words, full-connected self-attention may not good at dealing in such a small range. Therefore, it is a challenge that how to limit the range of attention in order to deal local information on Chinese NER task.

Taking into account the mentioned considerations, we attempt to explore a strategy for Chinese NER task that better dealing with local information. It is worth noting that the window attention approach has shown effective results in other deep learning domains, effectively extracting local features from the input data. For instance, when performing object detection[14] tasks, the objects to be detected do not take up the whole image, the usage of the Window Attention method enables the neural network to concentrate on a specific area of the image to help extract reliable information. This led to the conclusion that window attention would be a useful tool for obtaining information about entities, since the process of identifying things in sentences for the NER task is quite similar to this.

However, there is a clear issue if the model using Window Attention alone: how can windows communicate with one another? If the window size is two, all words are divided into distinct windows, as in Figure 3 for example. Chinese words are made up of individual characters, thus it is absurd if characters that make up a word were split into different windows, and unable to communicate with others.

Figure 3: Words separated into different windows

The Shift Window Attention served the solution for this problem, which involves shifting the window after doing a round of Window Attention and then perform a subsequent round of attention computation in new window. Following the completion of the attention calculation inside each window, the attention calculation window is shifted by a predetermined amount, which does not exceed the window's size, and a new attention calculation is then carried out inside the relocated window. Similar to Figure 4, the attention calculation window is shown in red and moves between two rounds of attention computation. In this manner, the vectors situated in different windows can interact with others.

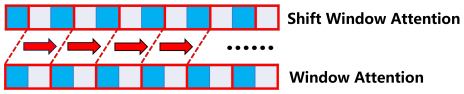


Figure 4: Shift Window Attention

In addition, if a long entity appears in the text that exceeds the length of a window, also a few of long-distance dependencies need to be solved, it cannot be handled by

simply using Window Attention. For these reason, although an approach that focus attention computing more locally is expected, the capability to obtain information on a larger scale is still necessary. Dilated Convolution may be an good idea for solution while considering about this issue, since it uses a "holey" convolution kernel to increase the receptive field of the convolution kernel. Similarly, it may allow for the expansion of the receptive field of the window when computing attention.

Each vector that occupies the same location in each attention calculation window is take out to create a number of new vector sequences, then another attention calculation is performed with the new sequences. The first vector within each window generates a new sequence, as shown in the Figure 5, and the other vectors does the same. If the window size is two, two new vector sequences are then processed similarly and iterated constantly to achieve cross-unit interaction.

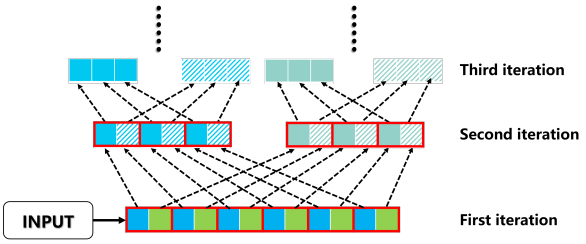


Figure 5: Window Dilitation

Due to the window attention and shifted window attention, each vector in the new sequence contains information about its nearby vectors. Thus, each vector that makes up the new sequence can be regarded as a representative of numerous vectors in the original sequence, the receptive field of one attention window in the new sequence is equivalent to several windows in the original sequence. As shown in Figure 6, several iterations are carried out in this manner, as the number of iterations rises, the receptive field of the attention window exponentially grows. This approach is named as Window Dilatation because a new window seems to be composed of original sequence with holes, similar to the Dilation Convolution.

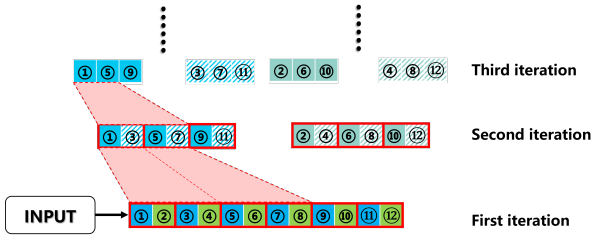


Figure 6: Receptive field changing in different iterations

The frame of our new attention module is a multi-round iteration and separated into three parts in each iteration: Window Attention, Shifted Window Attention, and window dilatation. The model can concentrate on local information by using Window Attention and Shifted Window Attention. Additionally, it enables each vector to carry the data of other vectors that come either before or after it. In this manner, the new sequence retains as much of the old sequence's information as feasible during the window dilatation. The iteration is then continued concurrently through the various sequences acquired after the window dilatation. Once a certain number of rounds have been completed or the new sequence is no longer lengthy enough to enable window dilatation, the iteration is stopped. What's more, neighboring vectors are split off into various sequences by window dilatation in the initial iterations, doing an additional Shift Window Attention at the end of the iteration, enable interaction between the original nearby vectors.

The most unique feature of DSWA is that concentrating more on extracting the local information than the self-attentive mechanism does, meanwhile information in a lager range can be fused partly. Our observation is that entities frequently exist in sentences as a string of words, therefore our NER task will be made easier by using this method to gather local information. In actuality, after experimental validation, our new attention module does boost Chinese NER's accuracy, particularly on the Weibo dataset where other approaches struggled.

4. Experiment

Weibo[15], MSRA[16], and Resume[7], three open datasets frequently utilized for Chinese NER tasks, are employed in this investigation. The Weibo dataset was gathered from the Chinese social networking site Sina Weibo; the MSRA, created by Microsoft from the news domain; and the Resume, gathered from the resumes of Chinese businesspeople. Statistics of the above datasets are shown in Table 1. Our method significantly improves the result of Weibo dataset.

Table 1: Statistics of datasets

Datasets	Type	Train	Dev	Test
Weibo	Sentence	1.4k	0.27k	0.27k
	Char	73.8k	14.5k	14.8k
Resume	Sentence	3.8k	0.46k	0.48k
	Char	124.1k	13.9k	15.1k
MSRA	Sentence	46.4k	-	4.4k
	Char	2169.9k	-	172.6k

Figure 7 is the frame of our model. In summary, it's a typical Transformer model, whereas the self-attention layer

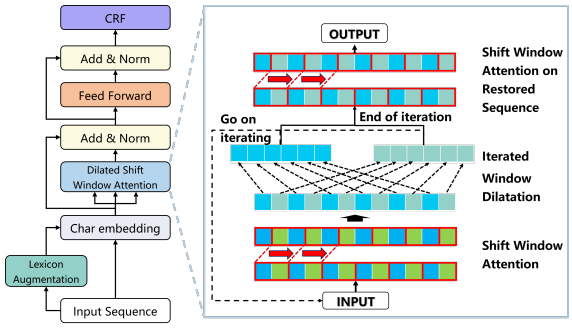


Figure 7: The Frame of DSWA

is replaced by our DSWA layer. It is worth noting that no matter which way of lexicon augmentation is used, there is a superior improvement after utilizing DSWA. Two representative models of both approaches are chosen in the experiment, FLAT is a typical example of dynamic architecture, while SoftLexicon is adaptive embedding.

SoftLexicon(Transformer) and FLAT model both is an enhanced version of the Transformer, by including external Lexicon information, the two model are enhanced considerably. Since self-attention mechanism were used by both of them, we can replace the self-attention layer with our DSWA to verify the effectiveness of our method.

Table 2: Performance on Three Datasets (F1 Score)

Model	Weibo	Resume	MSRA
Lattice-LSTM	58.79	94.46	93.18
TENER[17]	58.17	95.00	92.74
BERT+LSTM+CRF	67.33	95.51	94.83
FLAT	68.55	95.86	96.09
SoftLexicon	70.50	95.54	95.87
SoftLexicon + DSWA	**72.12**	**96.72**	**96.25**

The Table 2 shows that the DSWA-based model produced better results, particularly for the Weibo dataset where the improvement was greatest. The Weibo dataset, which is taken from the social media platform, has a strong propensity to be colloquial, meanwhile, there are a big number of short sentences and a higher occurrence of short entities according to our analysis. Therefore, DSWA, which is good at handling local information, is more suitable to handle such dataset. Both the MSRA and the Resume are plainly more stringent in their language, and more longer entities in them. Therefore, Weibo has been most significantly improved by our DSWA strategy, which places a strong emphasis on local information.

The models selected in the table are all classical models for Chinese NER task. Lattice-LSTM was the first to propose the method of external lexicon, Tener model was the first to improve Transformer, making model achieved good results on NER tasks, LSTM-CRF is a well-known baseline model for NER task, and both Flat and SoftLexicon (Transformer) are classical models that combine Lexicon Augmentation with Transformer as described above.

Table 3: Comparison of Self-Attention and DSWA (F1 Score)

Model	Weibo	Resume	MSRA
FLAT	68.55	95.86	96.09
FLAT + DSWA	69.63	96.21	96.03
SoftLexicon	70.50	95.54	95.87
SoftLexicon + DSWA	**72.12**	**96.72**	**96.25**

As shown in Table 3, comparing two Transformer-based models, the DSWA version have improved over the self-attention version; once more, the improvement is particularly noticeable on the Weibo dataset. It has been proven that using the DSWA method increases the model's accuracy and offers it a greater local information gathering capabilities. Therefore, we believe that DSWA can be used to replace the self-attention layer on more self-attention-based Transformer models to improve the model on Chinese NER task.

Concerning the problem of computation size, even though our method computes attention in multiple rounds, the computation size of the model actually decreases. When one attention computation is viewed as a unit, each vector of self-attention performs an attention calculation with every other vector. A total of n^2 attention computations are needed, where n is the length of the sequence produced by the sentence translation into vectors. However, Window Attention only needs to calculate attention within the window, and each window needs to perform W^2 attention calculations, where W is the length of the window, there are n/W windows in the sentence, so a sentence only needs to perform $W^2 \times n/W$, i.e. $n \times W$ attention calculations. Cause the window length W must be shorter than the sentence length n, the computation's complexity stays the same or even less than for self-attention. Even after r rounds of iteration, $r \times n \times W$ is smaller than n^2. In practice, r is generally set to a maximum of 4, because in the 4th iteration, the window's receptive field expands to 16, a size that essentially exceeds the length of all entity.

Table 6 demonstrates that the DSWA approach is marginally more efficient than doing a single round of self-attention computation, even numerous rounds of attention computation were carried out.

Compared with ASSTA, another method that modifies attention mechanism, in comparison, DSWA gets a similar f1 score, but a significant improvement in computational complexity, as shown in Table 4 and Table 5.

Table 4: Performance of ATSSA and DSWA

model	Weibo	Resume	MSRA
FLAT+ATSSA	72.53	96.73	96.45
SoftLexicon + DSWA	72.12	96.72	96.25

Table 5: Complexity of different methods. r is the rounds of iteration, W is the length of the window, r, W are constants.

Model	Complexity
Self-attention	O(n^2)
ATSSA	O(n^2 + 2n + nlogn)
DSWA	O($r \times W \times n$)

5. Conclusion

Dilatation Shifted Window Attention, which we suggest as a new attention calculation approach in this paper, is intended to improve the Transformer model's perception of local information while still preserving some perception of global information. It can better receive local information thanks to the use of the Window Attention mechanism, yet it can still perceive global information according to the iterative window dilatation approach, which enlarges each window's receptive field.

In brief, DSWA is a way of attention calculation that may be employed in place of self-attention and has a better ability to gather local information than the self-attention approach. It may also do well on other tasks that need deal data more locally. The experiment's findings match what we had anticipated.

Acknowledgement

The work is supported by the National Natural Science Foundation of China under grant 62276011.

Table 6: Efficiency of Two Attention Method

Datasets	Average Seconds per epoch	
	DSWA	Self-attention
Weibo	44	64
MSRA	2020	2729
Resume	105	131

References

[1] Jason P.C. Chiu and Eric Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 07 2016.

[2] Detmar Meurers. Natural language processing and language learning. *Encyclopedia of applied linguistics*, pages 4193–4205, 2012.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[4] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2670–2680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[6] Ying An, Xianyun Xia, Xianlai Chen, Fang-Xiang Wu, and Jianxin Wang. Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf. *Artificial Intelligence in Medicine*, 127:102282, 2022.

[7] Yue Zhang and Jie Yang. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*, 2018.

[8] Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. An encoding strategy based word-character lstm for chinese ner. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[9] Jun Yin and Cui Zhu. Embedding lexicon and direction information in chinese ner. In *International Conference on Computer Information Science and Artificial Intelligence*, 2021.

[10] Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. Cnn-based chinese ner with lexicon rethinking. In *International Joint Conference on Artificial Intelligence*, 2019.

[11] Minlong Peng, Ruotian Ma, Qi Zhang, and Xuanjing Huang. Simplify the usage of lexicon in chinese ner. *ArXiv*, abs/1908.05969, 2019.

[12] Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 6836–6842, Online, July 2020. Association for Computational Linguistics.

[13] Biao Hu, Zhen Huang, Minghao Hu, Ziwen Zhang, and Yong Dou. Adaptive threshold selective self-attention for Chinese NER. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1823–1833, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[14] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, pages 1–20, 2023.

[15] Nanyun Peng and Mark Dredze. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Gina-Anne Levow. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia, July 2006. Association for Computational Linguistics.

[17] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: Adapting transformer encoder for named entity recognition. *ArXiv*, abs/1911.04474, 2019.