Chinese-English Machine Translation Model Based On Transfer Learning And Self-attention

Shu Ma

Shenyang Normal University, No. 253 Huanghe North Street, Shenyang 110034, China

Corresponding author. E-mail: mashuetd@163.com

Received: Oct.17, 2023; Accepted: Nov.03, 2023

With the continuous development of machine learning and neural networks, neural machine translation (NMT) has been widely used due to its strong translation ability. Lexical information is overused in the construction of the internal nodes that make up the structure. Using phrase structure encoders can lead to over-translation problems. In addition, the number of model parameters increases with the use of grammatical structures, and the phrase nodes may not always be beneficial to the neural translation model. Therefore, we propose a novel Chinese-English machine translation model based on transfer learning and self-attention. In order to make use of the position information between words, the absolute position information of words is represented by sine-cosine position encoding in the machine translation model based on self-attention mechanism. However, while this method can reflect relative distance, it lacks direction. In this paper, a new machine translation model is proposed by combining transfer learning with self-attention mechanism. This model not only inherits the high efficiency of self-attention mechanism, but also preserves the distance information and direction information between words. The results of translation experiments show that the proposed transfer learning model is significantly better than the traditional tree model.

Keywords: Chinese-English machine translation, transfer learning, self-attention, sine-cosine position encoding

©The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202408_27(8).0015

1. Introduction

In recent years, machine translation as an important part of natural language processing has made rapid development. The original statistical, phrase-based and instance based translation methods have gradually evolved into neural machine translation methods based on neural networks and encoder models, which are even comparable to human translation in translation quality [1, 2]. At the same time, the translation mode has gradually developed from the original translation mode from one language to another language to enable simultaneous translation from one language to multiple languages, gradually showing a trend of diversification. However, there are differences in grammar rules and writing forms between Chinese and English, which makes the traditional machine translation model unable to play a good role. Therefore, designing a high-quality translation model has become an important means to improve the quality of Chinese-English translation [3].

Since its emergence, neural network machine translation has far exceeded traditional statistical machine translation and rule-based and phrase-based machine translation in terms of performance. Akeel and Mishra [4] applied artificial neural network (ANN) to Chinese-English machine translation, added CBR and translation rule base, and finally used CBR method to select Urdu translation rules for input English sentences. However, the translation rules need to be made manually, and often need to be updated and maintained, resulting in a large workload and miscellaneous. The translation and research of neural

machine translation model between large languages has attracted much attention, but the translation effect for small languages is not good [5].

The translation model based on neural network, such as recurrent neural network, is a kind of sequential structure, which contains the position information of the word in the sequence. Transformer does not include recurrent neural networks (RNN) [6, 7] and convolutional neural networks. Transformer processes every word in a sentence in parallel, has no ability to recognize the order of each token, and does not contain any location information. The inclusion of Positional Embedding (PE) [8] helps the model identify location information. In contrast, Sinusoidal position encoding uses predefined functions to calculate position coding information, which does not contain learnable parameters and is not flexible enough. The relative position encoding proposed by Hayworth et al. [9] integrated the relative position representation into transformer's self-attention mechanism, although it greatly improved the translation performance of neural machine translation system. To some extent, long-distance position dependence is sacrificed.

Neural machine translation model is an end-to-end model using an encoder-decoder framework [10]. In neural machine translation, the neural machine translation model treats the source language sentence as a sequence of words or words, ignoring the structural information inherent in the language. The first explicit introduction of syntactic knowledge into neural machine translation began with the work of Wu et al., which showed that the introduction of linguistic knowledge into neural machine translation would greatly improve translation performance [11]. However, some existing methods of fusion dependency syntax use addition or splicing dependency syntax vectors on the basis of obtaining word vectors, which will obscure the meaning expressed by word vectors to some extent.

Aiming at the limitations of traditional neural machine translation methods, a novel Chinese-English machine translation model based on transfer learning and self-attention is proposed. An example shows that the proposed model is superior to other models with fewer parameters, and has great application value in the fields of machine translation and automatic construction of domain knowledge base. This paper is organized as follows. In section 2, we give the related works. Section 3 introduces the proposed model in detail. Experiments are conducted in section 4. There is a conclusion in section 5.

2. Related works

There have been some related work from different angles to try to use different methods to construct position coding information, which can be roughly divided into absolute position coding and relative position coding, and some other fancy encoding methods. Murthy and Hanumanthaiah [12] directly regarded location encoding as a trainable parameter, but this method that could learn location information from the training process could usually deal with a limited text length. Since natural languages generally rely more on relative position. Shan et al. [13] proposed relative position coding, which did not fully model the position information of each input, but was a method to consider pair relationships between input elements through improved extensions of self-attention mechanisms. Incorporate relative position representation into transformer's self-attention mechanism. Considering the relative distance between the current position and the Attention position during Attention training can greatly improve the translation performance of neural machine translation system.

Although relative location coding has no limit on text length, it sacrifices long-distance location dependence to some extent. Omote et al. [14] adopted dependency tree to represent the semantic structure of the sentence, simplified the syntactic relationship between input words, and encoded the position information according to the depth of each word in the dependency tree. Sadr et al. [15] proposed a recursive position embedding method based on word vectors to capture the sequential dependencies according to the word content in the sentence. The sequential dependence based on word content is encoded into word embeddings by learning cyclic location embeddings by recurrent neural networks. They are then integrated into an existing multi-head self-attention model either as separate heads or as part of each head. Hou and Li [16] proposed a position encoding based on continuous dynamic systems and constructed the FLOATER model. The model was learned by an ordinary differential equation solver, which could not only model the position relationship, but also was not limited by the text length, making it very flexible. However, this location coding method sacrifices parallelism to some extent, which may lead to speed bottlenecks.

In the aspect of integrating syntactic analysis into neural machine translation, Membarth et al. [17] linearized the syntax tree of source language sentences after depth-first traversal to obtain syntactic structure information. Wang et al. [18] made use of the dependency syntactic structure of the source language, took the dependency relationship, part of speech, word root and other information in the dependency structure as characteristics, respectively represented it with different vectors and spliced together with word vectors to form the input vector of each source language word.

3. Chinese-english machine translation model

3.1. Transfer learning strategy

Transfer learning generally obtains certain knowledge by training the original task and memorizes it, and then transfers the stored knowledge to a task that is similar to the original task. Transfer learning strategy allows to borrow a large amount of existing labeled data to train the network and transfer the knowledge learned to the neural network model with less labeled data, thus reducing the amount of training data for application tasks [18, 19].

Traditional machine learning methods of natural language processing first train the model for a specific language through a large number of corresponding parallel corpora, and then apply the translation model to the specific language translation task. In contrast, transfer learning no longer requires its basic conditions. First, the data used to train the machine learning model and the test data must be equally distributed and the data are independent of each other. Second, the data set used for training must ensure a certain scale in order to obtain an ideal model.

Transfer learning is one of the machine learning methods. In transfer learning, Domain is the main body of learning, which consists of data characteristics and distribution. The domain is further divided into the source domain containing the existing knowledge and the target domain to be learned. Transfer learning is the study of how to transfer the knowledge learned in the source domain to the target domain. The method can be divided into Instance-Based Transfer Learning (IBTL) according to different transfer methods, Feature-Based Transfer Learning (FBTL), Parameter-Based Transfer Learning (PBTL) and Relation-Based Transfer Learning (RBTL) [20]. In this paper, model-based migration, also known as parameter migration method, is studied in English and Chinese machine translation, that is, there are model parameters that can be shared in the source domain and the target domain. The specific method is to first train the source domain, get a model with good effect, record and transfer the model parameters to the target domain, and then continue to learn a new model according to the target domain.

The transfer learning strategy is very suitable for tasks where there is a lack of labeled data. At present, except for a few languages with abundant parallel corpus data resources (such as English and Chinese), many languages have a common problem of lack of bilingual parallel corpus resources and insufficient labeled data [21]. The introduction of transfer learning will alleviate this difficulty.

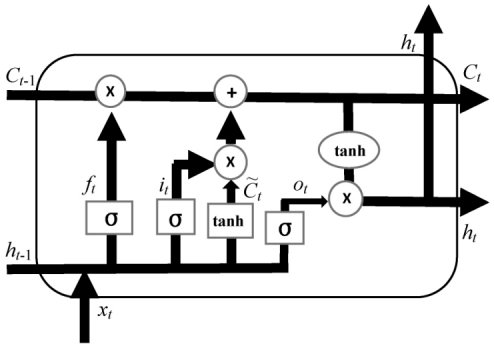

Fig. 1. LSTM structure diagram

3.2. Long Short-term Memory (LSTM)

1. One-way LSTM. Natural language processing is a typical sequence-to-sequence machine learning task, for which Recurrent Neural Networks (RNNS) are most commonly used. RNNS are able to memorize a sequence of data input to the network, and from this information RNNS can make some predictions. Long Short-Term Memory (LSTM) is a type of RNN that improves on the basic RNN. LSTM has a longer memory capacity, effectively overcoming the problem of long-distance dependence in machine translation, and significantly improving the fluency and readability of machine translation.

   LSTM adds a memory unit to determine whether the information is useful on the basis of ordinary RNN, which is called a block. The block is mainly controlled by gate mechanism. Three gate control units and a memory cell are placed in a block, including a forgetting gate control unit, an input gate control unit and an output gate control unit. The LSTM memory unit structure is shown in Fig. 1.

   In Fig. 1 h_t represents the output of the block at time t.x_t is the new information currently entered. In the forgetting gate, f_t is obtained from h_{t-1} and x_t, and it is used to calculate the degree of forgetting of information in C_{t-1}. An upper bound of 1 means complete memory, and a lower bound of 0 means complete forgetting, which is calculated as follows:

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right) \tag{1}$$

   In the input gate, i_t is used to control the updating degree of the new input status information, which is calculated by h_{t-1} and x_t. The new status information is $\tilde{C}$, and its calculation formula is as follows:

$$i_t = \sigma\left(W_i \times [h_{t-1}, x_t] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \times [h_{t-1}, x_t] + b_C\right) \quad (3)$$

The current new information C_t can be calculated by the above state, that is, forgetting some information and adding new information to be remembered. The calculation formula is as follows:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

In the output gate, o_t is used to control what information needs to be output. The calculation formula is as follows:

$$o_t = \left(W_o \times [h_{t-1}, x_t] + b_o\right) \quad (5)$$

$$h_t = o_t \times \tanh\left(C_t\right) \quad (6)$$

Where W is the weight matrix. b is the bias matrix.

In general, RNN may have the situation of gradient disappearance in training, and the perception of the later time node to the previous time node decreases, so that the connection between two nodes that are far away is weakened. LSTM can remember information that fits algorithms or rules, and other information is forgotten. When the information in the neural network enters the hidden node, the memory unit controls the propagation of information through the three gate mechanism. Only the information that conforms to the gate control will be remembered and transmitted backward, otherwise the forgetting gate will discard the information, and the retained information will continue to propagate backward or directly output the result according to the demand.

2. Two-way LSTM. One problem with unidirectional LSTM modeling sentences is that it cannot encode information from back to front. However, in machine translation, the translation of a word often needs context-dependent information, so bidirectional LSTM can better capture bidirectional semantic dependencies.

   Bidirectional LSTM is a combination of forward LSTM and backward LSTM. The forward calculation is performed from time 1 to time i, and the output of the forward hidden layer is obtained and saved at each time. The calculation is also reversed along time i to time 1, obtaining and saving the output of the backward hidden layer at each time. Finally, the final output is obtained by combining the output results of the corresponding time at each moment. The bidirectional LSTM structure diagram is shown in Fig. 2.

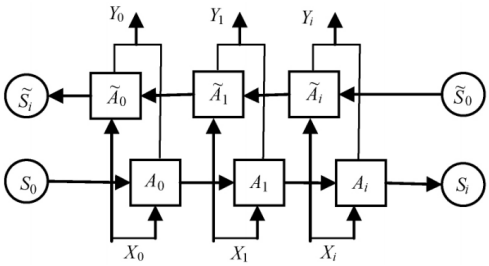

Fig. 2. Bidirectional LSTM structure diagram

In Fig. 2, X_i is the input at time i, and bidirectional LSTM will retain two values, the hidden layer state A_i in the forward calculation process and the hidden layer state $\tilde{A}_i$ in the reverse calculation process, and the final output value Y_i depends on the above two values. That is, in forward computation, hidden layer S_i is related to S_{i-1}, and in reverse computation, S_i is related to S_{i+1}.

3.3. Self-attention (SA)

The SA mechanism has attracted wide Attention due to its parallel computing capability and modeling flexibility, while the Multi-Head Attention (MHA) mechanism in the SA mechanism enables the model to focus on the corresponding information from different sub-spaces [22]. Since the SA mechanism ignores the position factor of the word in the sentence, it can explicitly capture the semantic relationship between the current word and all the words in the sentence, while the MHA mechanism maps the input sequence to different subspaces, which use the SA mechanism to further enhance the performance of the machine translation model. Compared with the traditional RNN model, the SA mechanism has the advantages of fewer parameters, faster speed and better effect.

As shown in Fig. 3, when SA mechanism is used to process each word (that is, each element in the input sequence), for example, when x_i is calculated, SA mechanism can associate it with all the words in the sequence and calculate the semantic similarity between them. The advantage of this mechanism is that it can help to mine the semantic relationship between all the words in the sequence. To encode words more accurately.

Each attention head has an input sequence $X = (x_1, x_2, \cdots, x_n)$, $x_i \in R^{d_x}$ for a set of n-tuples, and then calculate the output sequence $Z = (z_1, z_2, \cdots, z_n)$, $z_i \in R^{d_z}$ of a set of n-tuples.

The element z_i in the output sequence Z is derived from

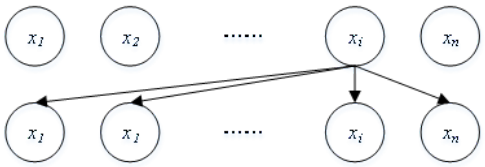

Fig. 3. SA structure

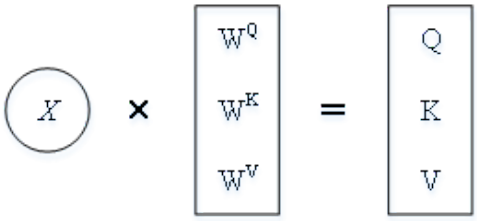

Fig. 4. Calculation process of Q, K, and V

the input element x_i, which is linearly transformed and its weighted sum computed to obtain:

$$z_i = \sum_{j=1}^{n} \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) V_j \tag{7}$$

In the softmax function, a linear transformation of the input elements enhances the expressibility. The softmax function is calculated as follows:

$$\text{softmax}\left(a_{ij}\right) = \frac{\text{expa}_{ij}}{\sum_{k=1}^{n} \text{expa}_{ik}} \tag{8}$$

Q, K, and V stand for query, key, and value respectively. They are useful abstract representations for calculating attention scores. d_k is the dimension of key. $\sqrt{d_k}$ is the scaling dot product, so that the gradient is more stable, Q, K, and V are calculated as follows:

$$Q_i = x_i W^Q \tag{9}$$

$$K_j = x_j W^K \tag{10}$$

$$V_j = x_j W^V \tag{11}$$

As shown in Fig. 4, W^Q, W^K, and W^V are the matrices learned in the training process. These are the weights of Q, K, and V. Each attention head has its own unique weight matrix.

The SA mechanism uses l attention heads, the output z_h of all attention heads is merged, and then a linear transformation is performed to get the output of each sub-layer. The multi-head attention mechanism extends the model's ability to focus on different locations. The output result of the multi-head attention mechanism is calculated as follows:

$$z^O = \text{Concat}\left(z_{\text{head 1}}, \cdots, z_{\text{headl}}\right) W^O \tag{12}$$

Z_{headi} represents the output vector of the i-th attention head. The function of Concat() is to combine the output vectors of all the attention heads. W^O is the weight matrix generated during model training. The multi-head attention mechanism combines the output of the individual attention heads and then performs a linear transformation to produce the final output.

3.4. Parameter transfer in TensorFlow

In this paper, the English-Chinese machine translation model based on parameter transfer is built in TensorFlow environment, using the memory object inside TensorFlow to store the model. The transfer of parameter weights is realized by importing the pre-trained model. The trained English-Chinese neural network parameter weights are transferred into the translation model proposed, that is, the node parameters of the network are no longer initialized randomly during English-Chinese neural machine translation, but the trained English-Chinese model parameters are imported into the proposed model for initialization. In TensorFlow deep learning framework, the calculation graph and related parameters are stored separately, so importing the pre-trained model requires two steps: first, the neural network model graph needs to be constructed; Then it loads the weight parameters.

After the pre-trained model is trained, the Checkpoint file (corresponding extension .ckpt) stores the values of relevant variables such as weights, bias, and gradients of all nodes in the neural network. save all network parameters by calling the save() method of the Saver object in TensorFlow. Then in the new neural network, the parameters of the pre-trained model are loaded by the restore() method, and the training is continued on this basis. Figure 5 shows the process of TensorFlow parameter migration.

4. Experiments and results

This paper conducts an experiment on Chinese-English language pairs. 50,000 sentence pairs randomly selected from News Commentary v12 in the WMT17 corpus are used as the training set in the translation task. The data came from the news field using newsdev2017 as the validation set and newstest2017 as the test set. Unused words. UNK > Indicates. This article uses the StanfordCoreNLP toolkit to obtain the dependency matrix for each sentence. The scale of experimental data is shown in Table 1.

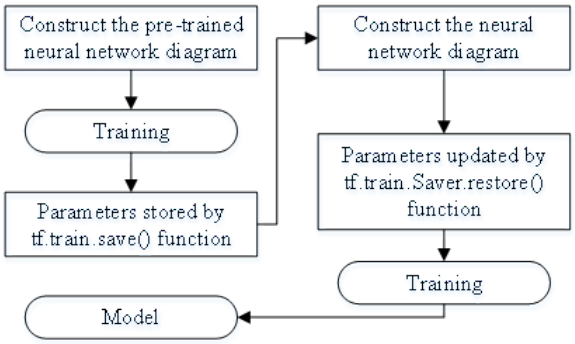

Fig. 5. Parameters transfer

Table 1. Experimental data scale statistics table

Pairs	Training sets	Validation set	Test set
Chines-English	50000	2005	2002

Experimental environment: Operating system Windows11, memory 32 GB, disk space: 1 T, CPU: Intel Core i7, graphics card: NVIDIA GeForce RTX 2060, python3.6, 64-bit, using the deep learning framework tensorflow development.

Parameter setting. The basic structure of the model proposed in this paper is based on Transformer. The encoder and decoder of the model in this paper are stacked with 6 encoder and 6 decoder sub-layer respectively. The dimension of the word vector and the size of the hidden layer at source and destination are set to 512 . The number of attention heads is 8 , the dimension of the feed-forward neural network is 2048. Dropout is set to 0.1, and the initial learning rate is set to 0.5 . The model parameters are updated using the Adam algorithm, and the optimizer parameters are set to $\beta 1 = 0.9, \beta 2 = 0.98, \varepsilon = 10^{-9}$. Only sentences with no more than 100 sentences in the source language are used, and the CBOW model context window size c is 2 .

In order to verify the effectiveness of this model, the experiment is mainly improved on the basis of the baseline model Transformer. In order to more intuitively understand the importance of different improvement modules of this model for system performance, the proposed model is decomposed during the experiment, and the complete model proposed in this paper is compared with the model of a shielded module. Understand and analyze the importance of each sub-module to the translation performance of the model. The experiment in this paper mainly analyzes the importance of the two sub-models added after the improvement of transformer based on the original baseline model, that is, the dependency matrix is used to add semantic information to the original model and the LSTM model is used to obtain the location coding of language sequences. For the module using dependency tree, only the semantic relations of the source language sentences are represented by dependency matrix, and the word vector is integrated in such a way that the two words with the highest degree of closeness of each target word are added to the predicted sequence. For the LSTM model, the input sequence obtained by the dependency tree module is input into the LSTM model for training, and the output of each time step is obtained. Because of the powerful memory function of LSTM, the output of each time step is regarded as the position information of each word in the sentence in this paper. In this paper, BLEU value [23]is used as the evaluation criterion of the model. The experimental results are shown in Table 2.

Table 2. Comparison results

Model	BLEU
TNN [24]	26.99
BIAM [25]	27.12
SSM [26]	27.44
Proposed-SA	29.65
Proposed	30.78

As can be seen from Table 2, the translation model proposed in this paper performs better than the benchmark model in Chinese-English translation. This paper takes newsdev2017 as the verification set and newstest2017 as the test set to start the experiment. The average BLEU value of the model proposed in this paper can reach 30.78, which is 3.79 BLEU points higher than the TNN model.

Therefore, the method proposed in this paper can improve the translation performance of transformer model. In order to verify the validity of the two sub-modules proposed in this paper, an ablation experiment is designed. As can be seen from Table 2, the overall model proposed in this paper improves by 1.13 BLEU points compared with the model without SA, which indicates that adding semantic information to the model can better help the model understand the sentence structure information, thus improving the translation performance. A translation example is shown in Table 3.

The BLEU value of Chinese to English translation on the validation set increases with the number of training steps. Compared with the baseline model, the method proposed in this paper has a better translation effect. It shows that the method proposed in this paper has a clearer understanding of the semantic results of sentences after adding linguistic knowledge to the baseline model Transformer, and the improved accuracy of position coding information obtained by using LSTM makes the proposed model more accurate

Table 3. Translation example.

Sentence	国际中文教育资源的数字化为全世界的汉语爱好者提供了海量优质教育资源
Reference Version	The digitization of international Chinese education resources has provided a large number of high-quality educational resources for Chinese lovers all over the world
TNN	The digitization of international Chinese education resources offers a large number of high-quality educational resources for Chinese lovers in the world
Proposed	The digitization of international Chinese education resources has provided a large number of high-quality educational resources for Chinese lovers in the world

in translating source language sentences and generating more accurate translation results.

5. Conclusions

In this paper, LSTM and SA mechanism are combined to improve the performance of machine translation. The experimental results show that the proposed model can more accurately represent the position relationship between long-distance words, and can achieve better scores in data sets with more long sentences. This is because LSTM convergence is slow, and the concept of long distance is blurred in a progressive way, which can more accurately capture the difference in the position relationship between long-distance words. However, the experimental results of the proposed model in the data set with more short sentences are not satisfactory, because the accuracy of the position relationship between short distance words is different from the actual situation when logarithmic subscript is used. The next step will continue to study how to effectively combine syntactic analysis and SA mechanism to enhance the ability of the model to model the syntactic structure. The effective combination of convolutional network and SA mechanism to enhance the ability of the model to obtain local information is also worth further exploring.

References

[1] H. L. Trieu, D.-V. Tran, and M. Le Nguyen. "Investigating phrase-based and neural-based machine translation on low-resource settings". In: *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*. 2017, 384–391.

[2] B. Marie and A. Fujita. "A smorgasbord of features to combine phrase-based and neural machine translation". In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. 2018, 111–124.

[3] Y. Zhang, Y. He, and L. Zhang, (2023) *"Recognition method of abnormal driving behavior using the bidirectional gated recurrent unit and convolutional neural network"* **Physica A: Statistical Mechanics and its Applications *609***: 128317. DOI: 10.1016/j.physa.2022.128317.

[4] M. Akeel and R. Mishra, (2014) *"ANN and rule based method for english to arabic machine translation."* **Int. Arab J. Inf. Technol. *11***(4): 396–405.

[5] J. Zheng, Z. Zhao, M. Chen, J. Chen, C. Wu, Y. Chen, X. Shi, Y. Tong, et al., (2020) *"An improved sign language translation model with explainable adaptations for processing long sign sentences"* **Computational Intelli-**

gence and Neuroscience *2020*: DOI: 10.1155/2020/8816125.

[6] A. Jisi, S. Yin, et al., (2021) *"A new feature fusion network for student behavior recognition in education"* **Journal of Applied Science and Engineering *24***(2): 133–140. DOI: 10.6180/jase.202104_24(2).0002.

[7] S. Yin and H. Li, (2020) *"Hot region selection based on selective search and modified fuzzy C-means in remote sensing images"* **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing *13***: 5862–5871. DOI: 10.1109/JSTARS.2020.3025582.

[8] K. Chen, R. Wang, M. Utiyama, and E. Sumita. "Recurrent positional embedding for neural machine translation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, 1361–1367.

[9] K. J. Hayworth, M. D. Lescroart, and I. Biederman, (2011) *"Neural encoding of relative position."* **Journal of Experimental Psychology: Human Perception and Performance *37***(4): 1032. DOI: 10.1037/a0022338.

[10] F. Stahlberg, (2020) *"Neural Machine Translation: A Review and Survey"* **Journal of Artificial Intelligence Research *69***: 343–418.

[11] S. Wu, D. Zhang, Z. Zhang, N. Yang, M. Li, and M. Zhou, (2018) *"Dependency-to-dependency neural machine translation"* **IEEE/ACM Transactions on Audio, Speech, and Language Processing *26***(11): 2132–2141. DOI: 10.1109/TASLP.2018.2855968.

[12] P. N. Murthy and S. K. Y. Hanumanthaiah, (2022) *"A simplified and novel technique to retrieve color images from hand-drawn sketch by human."* **International Journal of Electrical & Computer Engineering (2088-8708) *12***(6): DOI: 10.11591/ijece.v12i6.pp6140-6148.

[13] W. Shan, H. Lu, S. Wang, X. Zhang, and W. Gao. "Improving robustness and accuracy via relative information encoding in 3d human pose estimation". In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, 3446–3454. DOI: 10.1145/3474085.3475504.

[14] Y. Omote, A. Tamura, and T. Ninomiya. "Dependency-based relative positional encoding for transformer NMT". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019, 854–861. DOI: 10.26615/978-954-452-056-4_099.

[15] H. Sadr, M. M. Pedram, and M. Teshnehlab, (2019) *"A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks"* **Neural processing letters *50***: 2745–2761. DOI: 10.1007/s11063-019-10049-1.

[16] S. Hou and Y. Li, (2010) *"Detecting nonlinearity from a continuous dynamic system based on the delay vector variance method and its application to gear fault identification"* **Nonlinear Dynamics *60***: 141–148. DOI: 10.1007/s11071-009-9586-9.

[17] R. Membarth, O. Reiche, F. Hannig, J. Teich, M. Körner, and W. Eckert, (2015) *"Hipa cc: A domain-specific language and compiler for image processing"* **IEEE Transactions on Parallel and Distributed Systems *27***(1): 210–224.

[18] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, and K. Gryllias, (2022) *"A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges"* **Mechanical Systems and Signal Processing *167***: 108487–. DOI: 10.1016/j.ymssp.2021.108487.

[19] X. Yu, J. Wang, Q.-Q. Hong, R. Teku, S.-H. Wang, and Y.-D. Zhang, (2022) *"Transfer learning for medical images analyses: A survey"* **Neurocomputing *489***: 230–254. DOI: 10.1016/j.neucom.2021.08.159.

[20] L. Zhang, L. Guo, H. Gao, D. Dong, G. Fu, and X. Hong, (2020) *"Instance-based ensemble deep transfer learning network: A new intelligent degradation recognition method and its application on ball screw"* **Mechanical Systems and Signal Processing *140***: 106681. DOI: 10.1016/j.ymssp.2020.106681.

[21] J. Li, L. Zhang, X. Shu, Y. Teng, and J. Xu, (2023) *"Multi-instance learning based on spatial continuous category representation for case-level meningioma grading in MRI images"* **Applied Intelligence *53***(12): 16015–16028. DOI: 10.1007/s10489-022-04114-x.

[22] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, (2023) *"Uniformer: Unifying convolution and self-attention for visual recognition. arXiv 2022"* **IEEE Transactions on Pattern Analysis and Machine Intelligence *45***(10): 12581–12600.

[23] A. K. Yadav, A. Singh, M. Dhiman, Vineet, R. Kaundal, A. Verma, and D. Yadav, (2022) *"Extractive text summarization using deep learning approach"* **International Journal of Information Technology *14***(5): 2407–2415. DOI: 10.1007/s41870-022-00863-7.

[24] J. Liang, M. Du, et al., (2022) *"Two-way neural network chinese-english machine translation model fused with attention mechanism"* **Scientific Programming *2022***: DOI: 10.1155/2022/1270700.

[25] L. Yonglan, H. Wenjia, et al., (2022) *"English-Chinese machine translation model based on bidirectional neural network with attention mechanism"* **Journal of Sensors *2022***: DOI: 10.1155/2022/5199248.

[26] Z. Wang, X. Liu, and M. Zhang, (2022) *"Breaking the representation bottleneck of chinese characters: Neural machine translation with stroke sequence modeling"* **arXiv preprint arXiv:2211.12781**: