IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2024, accepted 19 August 2024, date of publication 26 August 2024, date of current version 3 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3449280

RESEARCH ARTICLE

A Text Summarization Approach to Enhance Global and Local Information Awareness of Transformer

ZHENG LIU[1], HENG WANG[1], CONG ZHANG[2], AND SHUAI ZHANG[1]
[1]School of Mathematics and Computer, Wuhan Polytechnic University, Wuhan 430048, China
[2]School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan 430048, China

Corresponding author: Heng Wang (wh825554@163.com)

ABSTRACT In the field of abstract text summarization, architectures based on encoder-decoder frameworks are widely applied to sequence-to-sequence generation tasks and can effectively handle sequences of unlimited length. Subsequently, the transformer model use a global attention mechanism, allowing encodings at different distances to mutually interact, greatly enhancing the model's contextual awareness. However, this context-awareness is global, requiring the model to additionally learn to extract different levels of information to increase understanding. We improve the structure of the model to introduce prior knowledge so that it can learn from the global and local information and enhance the model's understanding ability. This paper proposes global information-aware encoding and local information-aware encoding, which enhance the understanding of documents from coarse-grained and fine-grained perspectives respectively. Global encoding adds an extra feature to the encoder stage and performs attention with the document, generating a global summary encoding of the entire document to guide the generation of the summary content. Local encoding is to perform local convolution on the features extracted by the encoder, use prior knowledge to extract local features of the document and enable the model to quickly extract local detail information. Experiments show that the improved model proposed in this paper has higher rouge scores than the baseline model on the LCSTS and CSL datasets, and also has advantages over some mainstream models. The generated summaries are more accurate and informative. The code is available on github. url: https://github.com/keptupp/A-text-summarization-approach-to-enhance-global-and-local-information-awareness-of-transformer.

INDEX TERMS Text-summarization, transformer, lstm, vit, cnn.

I. INTRODUCTION

Text summarization in natural language processing aims to transform long text segments into short text summaries. Text summarization tasks can be roughly divided into extractive and abstractive [8]. Extractive text extraction mainly extracts some important sentences from the original document and combines them into a short summary, so it has high authenticity. However, when the logic is complex in the document, the extracted sentences can not express the original idea well, and the combined summary is lack of logical fluency. Abstract methods mainly use deep learning models to generate document summaries. After learning a large amount of text summary data, the deep model can understand the meaning of the document and generate smooth document summaries, which is more suitable for people's reading styles. However, the training cost of the model is high, and the performance of the model depends on the quality of the dataset [25]. When the document is beyond the scope of the model data set and the cognition of the model, the results are often unreal and easy to confuse.

Most of the existing mainstream text summarization frameworks are based on Seq2Seq [36]. On this basis, different structures are derived to improve the model and enhance the performance of the model. Recurrent neural network (RNN) and long short-term memory (LSTM [9]) are widely used in text summarization because they can process text sequences of different lengths, but they also have the disadvantages of high computational complexity and long-distance gradient disappearance or explosion [33],

The associate editor coordinating the review of this manuscript and approving it for publication was Venkateshkumar M.

[34]. The subsequent transformer [37] breaks through the limitation that data cannot be computed in parallel. The self-attention operation makes the relationship between any position not become complex with the change of position, simplifies the model structure, and improves the performance of the model.

In terms of abstract summary, to improve the understanding ability of the model, researchers proposed a deep neural network that mimics human reading, adding hierarchical perception modules to simulate different fine-grained reading, adding multi-task learning methods to simulate careful reading and grammatical error correction, and adding adversarial learning to improve the quality of summary generation [46]. In addition, the classical model cannot discard noise in the language, by adding a self-aware context selection mechanism to extract the utterance state required by the decoder, and an asynchronous bidirectional recurrent neural network to align parallel computation with sequential processing [11]. In addition, a multi-level hierarchical BART model is proposed based on the BART model. The author believes that BART ignores the interaction between sentence level and word level, and adding a hierarchical structure can capture different fine-grained features and improve the performance of the model [1]. There are also improvements to the predictive output of the model, which is a deterministic single-point distribution during training and can produce inference biases. Adding a new training paradigm with non-deterministic probability distribution can effectively avoid inference bias [23].

According to previous research results and directions, self-attentive structures have excellent performance in language processing, and the encoder obtains the full-text encoding sequence through the attentional relationship between words. To some extent, this ordering reflects the model's understanding of the document, which is based on the interactions between each word and lacks a basic description of the influence of the whole document on it [43]. In addition, when the distance between words becomes larger, the degree of interaction between words becomes smaller. Existing models usually add positional encoding to enable the model to learn positional features, so that the model can self-adjust the attention level to words in different positions. In fact, based on prior knowledge, the model's ability to perceive local words can be strengthened.

To sum up, in order to improve the performance of abstract models, the existing researches usually use feature extraction methods of different scales to obtain more abundant features. Based on this point of view, this study proposes two new methods to improve model structure based on transformer, which can significantly improve model performance.

Therefore, this paper proposes an abstract text summarization model with global summary encoding and local summary encoding. In response to the model's lack of global overview, this paper refers to the vision transformer (vit) model to introduces an additional coded fragment in the encoder so that it operates attentively with words as a global overview fragment [6], [42]. Different from vit, the generation of this encoding segment is generated by inputting the original document into a Long Short-Term memory network, which is obtained from the last time series generated by the LSTM. In this paper, a local convolution strategy is applied to the features output from the encoder to extract small-scale features, and the features are sent to the decoder for cross-attention operations, thus enhancing the local perception of the model. It has high application potential in the fields that require high accuracy and credibility of information, such as news summaries, paper summaries, and legal documents [7]. At the same time, the text summary model can also be used as a supplement to the large language model to provide more brief and accurate reference content. The main contributions of this work are as follows:

- Global summary encoding: Taking inspiration from vit, we add global encoding fragment to the model input, which enables it to perform attention operations with other word fragments to obtain overall synopsis information.
- Local summary encoding: Based on global self-attention, the output sequence of the encoder is convolved to capture local feature information.
- Text model The effectiveness of the model is verified on two abstract text datasets. The experimental results show that the rouge [22] score of the summary is improved compared with the baseline model.

II. BACKGROUND

A large number of networks with different structures emerge after abstract text abstracts are introduced into deep learning models and the time series characteristics of recurrent neural networks are widely used in abstract text abstracts. Reference [26] uses a recurrent neural network with an encoder-decoder and adds a network to capture keywords, a pointer mechanism to process low-frequency words, and a multi-level structure to capture document hierarchy. Reference [47] finds that its encoder can only take into account the representation of the read words in the process of using the cyclic network, and the author proposes the mechanism of re-reading to imitate the review behavior of human beings in the process of reading. Reference [45] Structural encoder-decoder Decouple the encoder-decoder of the recurrent neural network, and train the two parts separately to shorten the training time. In [40], an abstracted summary method is introduced, and a hybrid similarity measure is proposed by combining sentence vectors and Levenshtein distance and is integrated into the graph model for the recurrent neural network in the abstract stage. Reference [17] finds that there are often some common structures in the document, so it reflects the structure part before adding in the basic model and adopts VAEs as the generation framework to solve the inference generation problem. Reference [38] uses CNN on RNN to propose a joint attention and biased probability mechanism, merges topic information into the automatic summary model, makes it introduces context to generate more coherent and diverse summaries, and uses ROUGE to directly optimize the model for non-differentiable summary measures. Reference [29] can generate long text abstracts by combining extract and abstract abstracts and extracting some sentences from the original document for decoding in the coding part, to enhance the ability of long document abstracts.

Reference [14] adds a POClink for point-to-point information fusion using the correspondence between sentences, so as to reduce ungrammatical and meaningless output. Reference [41] adds convolution operation based on transformer structure to realize centroid attention, and compacts the number of features to reduce the calculation amount of self-attention. Reference [27] introduces a neural topic model with normalization to capture global semantic information, and uses context gating mechanism to better control global semantic expression. Reference [16] proposes the multi-scale attention mechanism, defines different language units such as sub-words, words and phrases, and builds a multi-scale transformer model based on these word boundary information and phrase level prior knowledge. In [28], it is assumed that there is a hierarchical underlying structure in the document, which captures long-term dependencies in the top coarse time dimension and preserves details in the bottom layer. Bottom-up and top-down coding strategies are introduced on the encoder of the model to capture more feature information. Reference [50] proposes a framework to solve the splitting and summarizing of long texts, which includes N coarse-grained and 1 fine-grained text. The text is segmented and paired at each stage, and the final summary output is produced after N coarse-grained stages and then a fine-grained one. Reference [44] introduces comparative learning in transformer to learn the similarities among original documents, original abstracts and generated abstracts to minimize them and obtain a better authenticity level.

Document [30] uses transformer's decoder structure to conduct unsupervised pre-training and learning on large-scale texts, so as to obtain strong language expression capabilities. After that, fine-tuning on different language tasks has achieved good performance. Reference [12] uses transformer encoder part for pre-training and mask the prediction part with mask mechanism, so that Transformer can infer the covered part based on contextual information and enhance its understanding ability. Meanwhile, additional training to distinguish whether two sentences are related is introduced to enhance its ability in question-answer reasoning. In [24], bert was further fine-tuned, removing additional NSP tasks and changing the mask mechanism of BERT from static to dynamic. Reference [15] combines the advantages of bert and gpt pre-training models, pre-training under the complete encoder decoder structure, while using a new method to reconstruct the original document destruction, and achieves good results in text generation tasks. Reference [31] uses transformer model with larger parameters, collects larger text data for pre-training, and provides a general framework for the field of pre-training models. Reference [48] proposes a new pre-training model, which learns to generate important sentences hidden in the document, and makes the model more suitable for the field of text summarization.

According to [49], when training generation tasks based on the seq2seq structure, based on maximum likelihood estimation, the model will have exposure bias in the inference process that is inconsistent with the training. A slight error of one token in the prediction process will cause the subsequent tokens to continue to stay away. Therefore, it is necessary to construct an evaluation model to reorder the candidate abstracts generated by beam search to get the final abstracts. Reference [32] also proposed the second training stage of the model based on exposure bias, and designed a robust resequencer to reorder a set of abstracts generated by the model to obtain higher-quality abstracts. Reference [13] argues that the digest system produces smooth but unreal summaries, and proposes a new decoding method, PINOCCHIO, to improve the model output by removing the last prediction token in each beam search to backtrack.

III. MODEL

In the encoder-decoder architecture, the function of the encoder is usually to extract the features of the original text, and the decoder generates the expected summary sequence according to this feature, so the quality of the feature extraction affects the quality of the generated text summary. The existing model architecture considers extracting features from different levels to enrich the feature content and improve the quality of model generation. This model also improves the transformer model from both global and local directions. The following are the two improvements of the model. Figure 1 shows the overall structure of the model. The text document is converted into a vector by the tokenizer, which is first processed by the LSTM and initialized to a global summary token. The global summary token is then combined with the document token into the encoder for self-attention operation, and the generated global summary token is separated from the document feature when the encoder outputs it. The global summary token is entered into the adaptive weighting network to generate the weights of the final summary. The extracted document features are further extracted by a convolutional network for a local range of features in addition to the cross-attention operation with the decoder. The two cross-attention operations result in two summary tokens, which are combined with the weights generated by the global summary token to the final summary token. It should be noted that in the N-layer attention operation of the encoder, N global summary tokens will be generated to correspond to the N adaptive weights in the decoder. That is to say, the tokens generated by the two cross-attention operations in each layer of the decoder will be combined before entering the next layer, not limited to the combination of the final output summary. Thus, the extracted local features and the global encoding are guaranteed to be fully utilized in the decoding process.

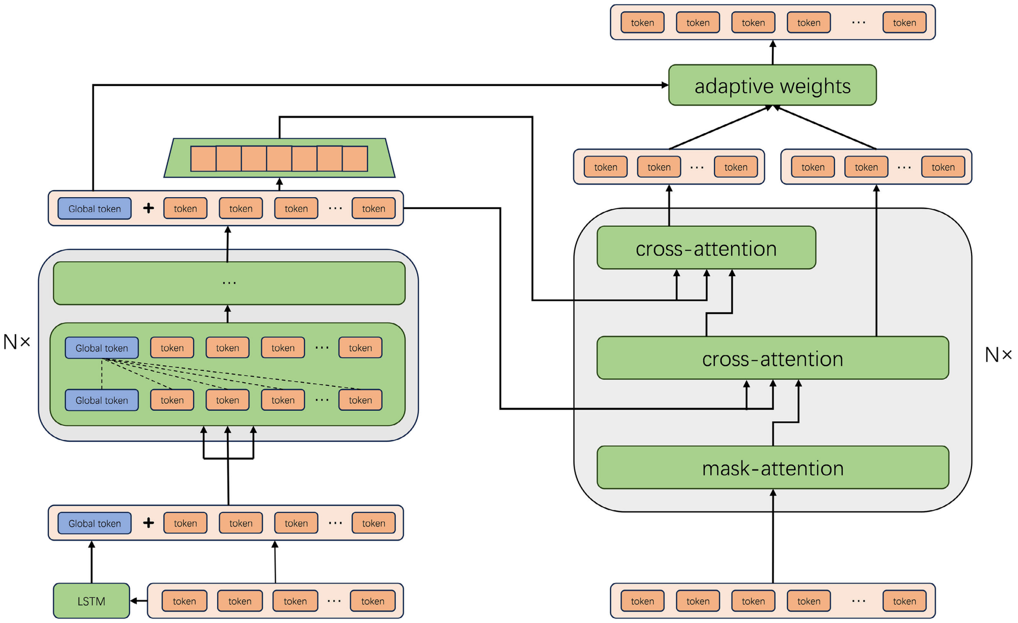

FIGURE 1. Overall model architecture for global and local information awareness.

A. GLOBAL SUMMARY ENCODING

In the process of text encoding, the self-attention operation obtains the encoded content generated at that encoding location by calculating the attention between different words. The information at each time series is computed with the full-text information. This computation is reasonable and the network captures the understanding of the whole document during the learning process, but the role of the encoder is to extract the hierarchically rich semantic information, and if each encoding structure of the encoder tends to combine the whole to make the output, it will make learning semantically rich feature extraction difficult. From this, a single coded fragment can be designed to operate attentively with the whole document to obtain a global summary encoding, allowing the encoder to focus more on hierarchically rich semantic extraction. This idea comes from the vit model, the vit model adds a classification encoding segment to the classification operation and takes the segment separately for classification operation when the model performs classification output. For the generation of this encoding, we don't take the random initialization of the learnable parameters of vit, but set an LSTM layer to input the original document, and take the last output sequence of the LSTM time series as the initial global summary encoding. The advantage of this method is that the global summary encoding fragment has roughly extracted the entire document before performing attention operations, which improves the accuracy of attention operations. After the attention operation, the global summary encoding does not input into the decoder with the extracted features for cross-attention, but generates the weights for different summaries by adaptive learnable weights.

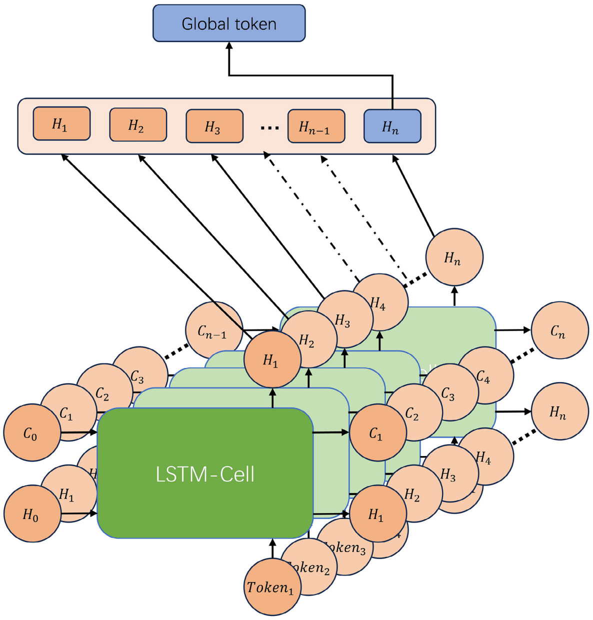

FIGURE 2. Global summary encoding structure.

Figure 2 shows the schematic diagram of generating global summary encoding by the LSTM model. The input and output at different moments are overlapped in the time dimension to obtain the global feature that needs to be extracted in this paper, and this global feature is the output H_n of the model at the last moment. The image shows only one layer of LSTM structure, while in the actual model, multiple layers of LSTM structure are set up and stacked in the same way.

Assume that the input of the encoding document is $(X_1, X_2, X_3, \ldots, X_m)$, then the global summary encoding is first generated by LSTM, and the output at each time is $(h_1, h_2, h_3, \ldots, h_m)$. The encoding used to maintain the long-term memory is $(C_1, C_2, C_3, \ldots, C_m)$, so the output at the final moment is expressed as follows:

$$o_m = sigmoid\left(\vec{W_o} \times [\vec{h_{m-1}}, X_m] + b_o\right) \quad (1)$$

$$h_m = o_m \times tanh(C_m) \quad (2)$$

where W_o is the learnable parameter, b_o is the inductive bias term, sigmoid is the activation function $sigmoid\ (x) = \frac{1}{1+e^{-x}}$, and Tanh as the activation function $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The output h_m at the last moment is used as the global summary encoding for the entire document, which is combined with the document encoding input $(X_1, X_2, X_3, \ldots, X_m)$ as an extra fragment X_0. Together with the document encoding, the segment encoding is fed into the transformer structure for global attention calculation to extract the information of the entire document. The computation for X_0 is as follows:

$$Q_0, Q_1, Q_2, \ldots, Q_m = MLP_Q\ (X_0, X_1, X_2, \ldots, X_m) \quad (3)$$

$$K_0, K_1, K_2, \ldots, K_m = MLP_K\ (X_0, X_1, X_2, \ldots, X_m) \quad (4)$$

$$V_0, V_1, V_2, \ldots, V_m = MLP_V\ (X_0, X_1, X_2, \ldots, X_m) \quad (5)$$

where Q_t, K_t and V_t are the query code, the key code and the value code at their respective time respectively. The MLP is a multilayer perceptual machine that converts the time series codes into their respective Q, K and V codes. The perceptual machines for Q, K, and V encoding at different times are parameter-shared. Then, the attention of different moments is obtained through the attention operation of Q_0 of the global summary encoding and the Q encoding of other K encoding. According to the attention size, the V encoding at different moments is extracted to obtain the output A_{out} of the final global summary encoding.

$$A_0, A_1, A_2, \ldots, A_m = \vec{Q_0} \times \left(\vec{K_0}, \vec{K_1}, \vec{K_2}, \ldots, \vec{K_m}\right) \quad (6)$$

$$\begin{aligned} A_{0\_out} &= softmax\ (A_0, A_1, A_2, \ldots, A_m) \\ &\quad \times (V_0, V_1, V_2, \ldots, V_m) \end{aligned} \quad (7)$$

where $A_0, A_1, A_2, \ldots, A_m$ represent the attention size of the global summary encoding X_0 to other encodations, $Softmax\ (x) = \frac{e^{xi}}{\sum e^{xi}}$, which is used to convert the attention size to the probability of sum to 1, and the attention size is used to convert to the probability of sum to 1. The output of the global profile A_0_out is summed over the products of the probabilities of other encoded attention and their respective values. An adaptive weight module is introduced to generate the weight for the decoder to decode the summary by global summary encoding.

$$S = \sum (T_1, T_2) \times awm\left(A_{0\_out}\right) \quad (8)$$

where S is the final output summary encoding, (T_1, T_2) is the two pre-selected summary codes generated by the decoder, and Awm is the adaptive weight module, whose input is the global summary encoding consistent with the model dimension, and the output is the weight of the two pre-selected summary encoding. The structure of the module is shown in Figure 3, which is composed of two layers of fully connected neural networks. The first layer further extracts the dimension of the summary encoding to be reduced by a factor of 8, and the second layer is transformed into a weight that adds the output of two cross-attention operations of the decoder. The two cross-attention in the decoder are described in Section III-B.

B. LOCAL SUMMARY ENCODING

Similarly, in the final features extracted by the encoder, the features on each time series contain all the information of the words. However, according to prior knowledge, most of the attention of the words should be focused on the vicinity of the words, which requires the model to learn the perception of different locations. We can actively perform local feature extraction after the model feature extraction so that the model directly obtains this part of the a priori knowledge, and let the self-attention mechanism pay more attention to the feature extraction of the relationship between distant words. The features output by the encoder are convoluted, and the range of locally extracted features is controlled by changing the size of the convolution kernel. At this time, the global summary encoding does not participate in the convolution operation, and the extracted features of this part are also transmitted to the decoder in the way of cross-attention. Then, the model adds output based on the cross-attention after convolution operation to the decoder structure and retains the original cross-attention and the output after convolution. In the process of the model's previous propagation, the two encodings are separately separated and weighted summed to obtain the output of three encodings.

After the attention operation of the encoder is finished, the output features are expressed as $A_1, A_2, \ldots, A_m$. The output features are entered into a layer of convolutional neural network for local feature extraction to enhance the ability of local perception of the model.

$$\begin{aligned} &F_1, F_2, \ldots, F_n \\ &\quad = cnn\ ((A_1, A_2, \ldots, A_m)\ , cnn\_kernel\_size, stride) \end{aligned} \quad (9)$$

where $F_1, F_2, \ldots, F_n$ are represented as the feature output after convolution extraction, cnn is a convolutional neural network that encodes the input with a convolutional kernel size of con_kernel_size and a step size of $stride$. This is a one-dimensional convolution operation, assuming that the inputs $A_1, A_2, \ldots, A_m$ are of size (B, M, D), where B is the batch size, M is the number of features (corresponding to the input text length), and D is the dimension size of each feature set by the model. Here, the convolution is aimed at the number of features. For the second dimension convolution operation, the same dimension corresponding to different features is parameter sharing to ensure the consistency of feature operation in the local synopsis. If the number of

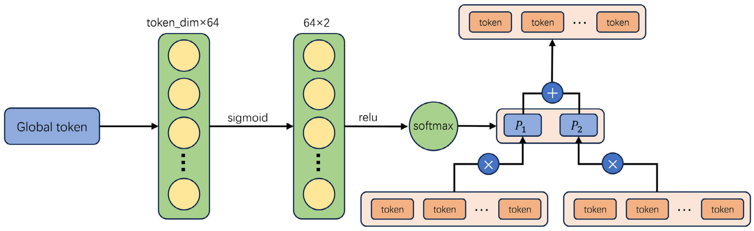

FIGURE 3. Adaptive weight module.

features after convolution is N, the following formula is satisfied:

$$N = \frac{M - con\_kernel\_size}{stride} + 1 \tag{10}$$

After the final convolution, the size of feature dimensions is not changed, only the number of features is changed, and the size of output features $F_1, F_2, \ldots, F_n$ is (B, N, D). The one-dimensional convolution operation of feature encoding is shown in Figure 4, which shows a convolution operation with a convolution kernel size of 3. After removing the global synopsis encoding, the remaining tokens are convoluted to keep the dimension of each token unchanged.

Suppose that the summary encoding input of the text is $(S_1, S_2, S_3, \ldots, S_i)$, and i is the length of the summary. The mask operation of the original summary is performed by Mask-Attention. The specific operation is to set 0 where the mask is needed in the attention matrix obtained by calculation and eliminate the attention to the mask. The other parts are consistent with the formula at the global profile encoding (refer to equations 3, 4, 5, 6, 7).

$$\begin{aligned}(S_{m1}, S_{m2}, S_{m3}, \ldots, S_{mi})\\ = mask\_attention\,(S_1, S_2, S_3, \ldots, S_i)\end{aligned} \tag{11}$$

The masked encoding $(S_{m1}, S_{m2}, S_{m3}, \ldots, S_{mi})$ is continuously input into the cross-attention to extract the output features of the original document. First cross attention for the original encoder output characteristics, $A_1, A_2, \ldots, A_m$ and the output of the decoder after the mask $S_{m1}, S_{m2}, S_{m3}, \ldots, S_{mi}$, (refer to equations 3, 4, 5, 6, 7), The generated k_a and v_a come from encoder features $A_1, A_2, \ldots, A_m, q_{sm}$ come from decoder $S_{m1}, S_{m2}, S_{m3}, \ldots, S_{mi}$, and the output of this cross-attention is the original text summarization output.

$$S_{o1}, S_{o2}, S_{o3}, \ldots, S_{oi} = cross\_attention(q_{sm}, k_a, v_a) \tag{12}$$

The inputs of the second cross-attention operation are the encoder's features $F_1, F_2, \ldots, F_n$ after local convolution and the decoder's outputs of the first cross-attention $S_{o1}, S_{o2}, S_{o3}, \ldots, S_{oi}$, refer to equations 9, 11, where the generated k_f and v_f are from $F_1, F_2, \ldots, F_n$, q_{so} from $S_{o1}, S_{o2}, S_{o3}, \ldots, S_{oi}$, and the output of this cross-attention is a summary after local information extraction.

$$S_{c1}, S_{c2}, S_{c3}, \ldots, S_{ci} = cross\_attention(q_{so}, k_f, v_f) \tag{13}$$

$S_{o1}, S_{o2}, S_{o3}, \ldots, S_{oi}$ and $S_{c1}, S_{c2}, S_{c3}, \ldots, S_{ci}$ are the two final output summarization encodings, which are summed up with the weight obtained from the global summary encoding to be the text summarization output of the inference process.

C. COMPARISON OF MODEL DIFFERENCES

In general, on the basis of transformer, we enhance the model's ability to extract global and local information, which is achieved by adding structure. Compared with transformer model, we add LSTM structure in text initialization to browse the entire document in advance. We add an additional code to calculate with the document to get the global information. At the end of text encoder feature extraction, we add convolutional network to further extract local information. In the text decoder, we introduce a second cross-attention structure to fuse the extracted local information. In the final output of the decoder, we add an adaptive weight structure to combine the decoding features to get the final output.

In addition, we also compare with the recently proposed model structure. Yuanyuan et al. [18] also encodes documents from a global and local perspective, encoding source text twice to obtain more feature information. This paper uses the attention mechanism to obtain global information at the information extraction stage, and constructs a local encode after extracting features. Yang and Roben [4] adopted a two-stage coding strategy and used extended convolution and gated convolution to extract important information. In this paper, one-dimensional convolution is used to further extract local features while retaining original features to ensure that no other information is lost.

D. MODEL TRAINING LOSS

A total of four cross-entropy loss functions are used in the model, which is the original output summary, the output summary after convolution, the summary of the output after summation of the original output and the weights after convolution, and adaptive weight learning. It's cross-entropy loss is expressed as follows:

$$loss = -\sum_{i=1}^{c} p_i \log\left(q_i\right) \tag{14}$$

where c represents the size of the bag of words at the time of prediction, p_i is the probability of predicting the ith word in the bag of words, and q_i is the distribution of the true value

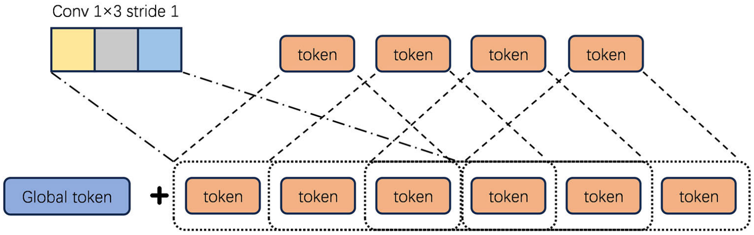

FIGURE 4. Local summary encoding convolution structure.

at the *i* th word in the bag of words. During model training, we know that the original output summary has the ability to output better summary sentences. Therefore, after adding the local summary cross-attention, the original summary is still retained and its loss is calculated, so that the network has the output ability before the local summary is further obtained, and the summary sentence is further optimized after the local summary is obtained.

The above two summaries are not the final output summaries of the model. During the experiment, we learned that although the summary after local summarization crossover is generally better than the original one, there are still some original summaries that have higher scores than the one after local summarization crossover. This is a common phenomenon caused by the loss function of the model and the evaluation index not completely corresponding. Therefore, this paper chooses to sum the two summaries according to a certain weight as the final summary output and introduces a third loss to further optimize the summary quality of the final output. Meanwhile, the fourth loss is generated by the adaptive learning module for learning to generate the magnitude of weights between the two summaries. The weight true value is obtained by calculating the score based on the above two summaries and the true summary.

IV. EXPERIMENT

To prove the effectiveness and rationality of the improvement of transformer structure in this study, two different Chinese text summarization datasets are selected for experiments. The baseline model and other models are used to compare with the model proposed in this paper. At the same time, each module proposed in the text is separated for ablation experiments to study the effect of each module on the performance of the model.

A. DATASET AND EVALUATION METRICS

In the experiment, two Chinese datasets are selected to verify the impact on the model improvement. LCSTS [10] is a large-scale Chinese short text summary database, whose data is from short text content published by high-quality users of Weibo, and the data sets are divided according to the original paper division rules. CSL [19] is Chinese scientific literature dataset contains 396,209 Chinese core journal papers, which are randomly divided according to the ratio of training set, validation set, and test set of the original papers to 8:1:1. The evaluation metrics are chosen to be rouge-N and rouge-L in rouge, which represent the quality of generation at the word and sentence level in the generated summaries, respectively. In addition, Bleu [3] and Meteor [2] evaluation indicators are used to make a more comprehensive evaluation of the proposed model. Bleu is often used in machine translation to calculate scores based on n-gram comparisons. Meteor metrics complement Bleu, taking into account accuracy and recall rates as well as sentence fluency.

B. EXPERIMENTAL SETUP

The baseline model uses the standard transformer structure and differs only with model improvements. The word segmentation model uses pre-trained bert-base-chinese [5], the bag of words size is 21128, the longest document sequence supported is 512, and the rest is truncated. The model uses 6 layers of encoder and decoder, the number of hidden units is 512, the number of multi-head attention is 8, and the dropout rate between network layers is 0.1. The maximum length of the summary output is 100. The LSTM network used to generate the global synopsis encoding is set to five layers and the number of hidden units is again 512. The CNN network for local synopsis encoding is set with 1 layer, convolution kernel size is 5, and step size is 1. Using the Adamw optimizer, the learning rate is fixed at 0.0001 and the batch size is 32. In the loss calculation of the adaptive learnable weight, the summary score based on the local synopsis crossover has a high probability of being better than the original summary, so the loss calculation ratio is set to 0.3 to 0.7. Locally crossed summaries were used for training the LCSTS dataset, and both were used for training the CSL dataset.

C. ANALYSIS OF RESULTS

The mainstream text summary models all use rouge evaluation index to measure the quality of the summary, which can better measure the coverage of the generated summary, but there are some problems such as ignoring the accuracy rate and being insensitive to word order. For a more comprehensive analysis of the performance of the model proposed in this paper, we also compared Bleu and Meteor indicators for additional reference. The results are shown in the following Table 1.

TABLE 1. Model performance under different indicators.

Dataset	Method	Rouge-L	Bleu	Meteor
LCSTS	Transformer	36.4	16.38	35.59
	OurModel	**37.35**	**17.39**	**36.19**
CSL	Transformer	55.27	34.50	57.02
	OurModel	**55.82**	**35.00**	**57.40**

The experimental results show that our improved model has improved on the three indexes, indicating that a more accurate and smooth summary is generated. In addition, we also calculated the difference in reasoning speed of the improved model, which was 355 tokens/s in the original model and 295 tokens/s in the improved model. We believe that the phenomenon of decreasing inference speed is reasonable, because the improved model adds complex structures to increase the model performance, among which the lstm structure is serial inference, which affects the inference speed of the model to some extent.

Table 2 shows the scoring performance of the improved model and other advanced models on the LCSTS dataset. TD-NHG [20] uses the decoder strategies of top-k, top-p, and repeated penalty mechanisms to improve the accuracy and diversity of the summary. WeLM [35] is the result of fine-tuning a large pre-trained model introduced by the WeChat team on text summarization. TI-C-NHG [21] extracts words with topic information from the text to improve the accuracy and readability of the model. GP_Step_0.3 is equipped with a stepwise gradient penalty mechanism of similarity to reduce training time and improve accuracy. Transformer is the attention model of encoder-decoder architecture, which is also the baseline model of this paper. The data show that the proposed model is better than other models, and has an improvement of nearly 1 point on its baseline model, indicating that the structural improvement proposed in this paper is effective.

TABLE 2. Performance on the LCSTS dataset.

Models	ROUGE-1	ROUGE-2	ROUGE-L
TD-NHG [20]	31.28	12.68	28.31
WeLM [35]	32.23	-	-
TI-C-NHG [21]	34.26	16.74	32.03
GP_Step_0.3	36.24	22.56	34.36
Transformer	40.3	27.0	36.4
OurModel	**41.0**	**28.1**	**37.3**

Table 3 shows the scoring performance of the improved model and other advanced models on the CSL dataset. Original T5 250 [39] is the performance of the pre-trained large model after fine-tuning with 250 samples, PEGASUS is the size pre-trained model proposed by the Google team for text summarization. BART is a pre-trained model with encoder-decoder that combines the characteristics of BERT and GPT pre-trained models. CSL-T5 is the performance of the CSL dataset team after fine-tuning using the T5 model combined with domain adaptation. LSTM-seq2seq is the result of a long short-term memory network trained on the dataset under the encoder-decoder architecture. From the results, the model with the improved structure proposed in this paper is better than other models selected, including the addition of pre-trained models such as T5 and PEGASUS, indicating that through the fine improvement of the model, the small model can even compete with some large models on a single task.

TABLE 3. Performance on the CSL dataset.

Models	ROUGE-1	ROUGE-2	ROUGE-L
Original T5 250 [39]	56.45	45.01	53.96
PEGASUS [48]	-	-	55.2
BART [15]	-	-	49.9
CSL-T5 [19]	-	-	52.10
LSTM-seq2seq [39]	46.48	30.48	41.8
Transformer	60.43	46.57	55.27
OurModel	**60.94**	**47.16**	**55.82**

D. ABLATION EXPERIMENT

Further discussing the effect of different parts of the proposed improved model on the model performance, the global summary encoding and local summary encoding structures of the model are eliminated one by one, the hyperparameters of the different structures are adjusted, and the performance performance of the model is observed using the CSL dataset training.

In the global summary encoding, the LSTM network is used to generate the global summary encoding in advance to accelerate the performance of the model in the attention operation. As shown in Table 4 below, the LSTM network is eliminated and the learnable random encoding is directly used instead. In the local summary encoding, this paper uses a layer of convolutional network to extract local information, as shown in Table 4 below. The convolution kernel size of the convolutional network is limited to 1 to eliminate the enhanced perception of local information.

TABLE 4. Different modules are added to the baseline model.

Models	Parmas	ROUGE-1	ROUGE-2	ROUGE-L
Transformer	76.61M	60.43	46.57	55.27
+global	76.66M	60.44	46.66	55.38
+local	86.37M	60.36	46.63	55.19
OurModel	96.88M	**60.94**	**47.16**	**55.82**

The experimental results show that, on the basis of the baseline model, adding global and local perception encoding respectively has different effects on the overall performance of the model. In general, the model with global encoding will have a certain performance improvement, and rouge-l has the largest improvement. However, the model performance on rouge-1 and rouge-L decreases instead of improving after adding local encoding. The most likely explanation

TABLE 5. Comparison of models on weibo news summaries.

Source text	Abstractive summary
号称“最大”的CSDN.NET数据库遭窃事件波及范围之广，持续时间之长让众多网民和互联网厂商心有余悸。而在黑客江 湖中，CSDN事件可能只是尘埃，或许，黑客在冷眼观看人们 的反应，享受着成功的快感。速途网独家对话一位匿名黑客，讲述黑客江湖的暗色人生。	S：一个黑客的自白“黑吃黑”的暗色江湖 B：黑客江湖的暗色人生 M：黑客江湖的暗色人生
在市场分化的同时，不同房企之间也在分化、转型。数以万计 的中小房企正经历一个大浪淘沙的过程。万科总裁郁亮认为，房地产行业已经过了青春期，不再是长“个头”，应该是长“力量”的时候，应该换个标准衡量已过青春期的房地产。	S：楼市已过青春期：房地产应长力量而不是长个头 B：万科郁亮：房地产行业已过青春期不在长个头 M：房地产已过青春期：长力量不再是长个头
跨年夜湖北武汉江汉路有市民扎堆放飞气球，气球触达电线瞬 间发生爆燃。一名现场目击者称看见气球缠到电线杆上面，然 后导致电线杆短路起火，多家店铺突然断电，警务人员对爆炸 事故路段进行了封锁。据步行街综合服务中心工作人员称，官 方并未组织跨年活动，气球是市民自行购买自行放飞。且事故 并非发生在步行街上，早在20年前，步行街经历两次改造，所 有架空电线已入地。此次事故发生地疑似夜市区域。	B：武汉跨年夜爆燃 M：武汉市民放飞气球爆炸
璀璨灯光、绚丽舞台、花样节目......诸多精彩汇聚一堂。日迈 月征，朝暮轮转，岁末尾声我们将这场视听盛宴，送给2023努 力奋斗的你！锁定12月31日20点档，齐聚《启航2024——中 央广播电视总台跨年晚会》，共迎美好明天！	B：《启航2024》跨年晚会 M：#启航2024#晚会精彩汇聚一堂

TABLE 6. The model generates a comparison of abstracts in Chinese scientific literature.

Source text	Abstractive summary
针对卫星通信中传统载波频偏估计算法精度低的问题,提出一种 新的频偏估计方法.该方法首先利用传统方法在解调前端进行初 始频偏估计,然后充分利用译码信息对残留频偏进行进一步估计 和补偿.仿真结果表明,该方法可有效抑制传统频偏估计方法带 来的残留频偏的波动,减少残留频偏对系统性能的影响,提高了 系统译码性能,降低了解调门限,在不占用系统频带的同时,提高 了系统功率利用率.	S：一种新的卫星通信系统中的载波频率同步技术 B：基于传统方法的卫星通信中传统载波频偏估计 M：基于传统载波频偏估计的频偏估计方法 M-g：一种新的卫星通信载波频偏估计方法 M-l：卫星通信中传统载波频偏估计方法研究
鉴于电子表格软件Excel提供的"规划求解"(Solver)加载宏可 以求解线性、非线性以及混合整数规划等单目标优化的问题的 特点,文章开发了排放口最优规划问题求解的应用程序(DPOP MS).程序包括了多河段BOD-DO耦合矩阵模型的计算、电子 表格模型的自动创建、调用"规划求解"加载宏进行规划求解以 及规划求解结果的输出等主要功能模块,并将程序应用于汾河太 原城区段的污染控制规划,所得的规划结果与实际情况较为吻合 .运用该程序进行排放口最优规划,避免了复杂的模型求解过程的算法程序编制,实现较为方便.	S：Excel在排放口最优规划中的应用 B：基于城市污染控制规划研究 M：基于加载宏的排放口最优规划问题求解方法 M-g：藕合矩阵模型在汾河太原城区段污染物控制规划中的应用 M-l：电子表格软件“规划求解”的应用程序

for this phenomenon is caused by the difference in model structure. Global summary encoding and text encoding for attention computation to obtain, and its resulting encoding on the decoder to produce a summary of some guidance, the role of the more obvious. The local encoding is generated directly into the decoder for cross-attention computation, which produces a larger granularity of encoding, which is also responsible for the fact that only rouge-2 is improved. When the two encodings are combined, the global summary encoder uses adaptive learnable weights to weight the initial summary and the large granularity summary, so that the resulting summary is reflected in different scales, thereby improving the performance of the model to generate the summary as a whole.

E. ABSTRACT ANALYSIS

The improvement of the model on rouge score reflects the better performance of the model to a certain extent, but the lack of measurement of the semantic information and the completeness of the information in the summary may generate summaries that are not smooth or untrue. It is necessary to compare and analyze the summary of the output of the model on the test set, so as to evaluate the quality of the summary generated by different models from a human perspective. In this section, two news statements are selected from the test set of the LCSTS data set to compare the quality of the answers given by the models. At the same time, two real-time news paragraphs are selected from Weibo to compare the baseline model and the improved model. On the CSL dataset, two texts are also selected on the test set. In addition to the comparison with the baseline model, the summary generated by the model with global summary encoding and local summary encoding in turn in the ablation experiment are compared. Where summary S represents the original summary of the dataset, B represents the summary generated by the baseline model, M represents the summary generated by the improved model, M-G represents the summary generated by the model with only global summary encoding, and M-L represents the summary generated by the model with only local summary encoding.

On the LCSTS dataset in Table 5, the performance of the first paragraph of news on the baseline and the improved model is consistent, although the model has improved in the rouge score, indicating that simple news is likely to generate accurate and consistent summary content on different performance models. In the second news paragraph, the improved model reflects the overall understanding ability of the model. Compared with the baseline model, the summary of the improved model accurately understands the previous and current situation of the real estate, while the baseline model only understands the current situation of the real estate. The next two pieces of text are selected from the current hot news on Weibo. The third piece of the news baseline model understands less key information, and the generated content is easy to be ambiguous. In the fourth news segment, the improved model not only understands the 2024 New Year's Eve party but will also depict the hilarious scene when the party takes place.

Table 6 shows the comparison of the summaries generated by the model on the CSL dataset, first the analysis of the summaries generated by the baseline and the improved model. The baseline model produces a summary that deviates from the facts in the first document, and interprets a new estimation method based on the traditional method as a traditional estimation method, while the improved model correctly states an estimation method based on the traditional method. In the second document, the baseline model only abstracts the study of urban pollutants, while the improved model understands more detailed research methods and means. The summary generated by global summary encoding and local summary encoding of the improved model is analyzed. After adding global summary encoding to the first document, the model's control of the overall direction of the document is improved, and it is understood that it is a new estimation method and the summary generation is more accurate, while the addition of local summary encoding has no effect compared with the baseline model. In the second document, more entities related to pollutants were extracted after adding global summary encoding, but the tool for using the coupling matrix was not found due to the lack of understanding of the details of the use method. The model with local summary encoding accurately found the main method used by Excel software for the document, but the lack of global summary encoding caused the deviation of the document theme and generated summaries that were not realistic.

V. CONCLUSION AND PROSPECT

This paper proposes an abstract text summarization model based on global and local summary encoding, which uses two different scales of methods to extract features of text information, and can more effectively understand the document to generate better text summaries. The improved model has higher Rouge indicators on LCSTS and CSL datasets. In addition, this paper also does ablation experiments for the effects of the two scales of coding respectively and analyzes the reasons for the advantages and disadvantages of the summaries generated at different scales.

Global summary coding can also be used in some fields where overall information needs to be extracted, such as sentiment analysis in natural language processing, document classification, and image classification in image processing. Local summary coding can be used in areas where local details need to be extracted, such as named entity recognition in natural language, part-of-speech tagging, semantic segmentation in image processing, and image hypersegmentation. In addition, the text length of the model is usually limited, and the global coding structure proposed in this paper may be a new way of text compression, using a fixed length of the code instead of the historical text, so that the model can process the fixed length of the code and improve the limitation of the text length of the model. The local coding structure also provides a new feature extraction method for transformer-cnn architecture. Large models evolve rapidly and perform well in the field of text summaries, but in some specialized areas, small expert models produce summaries that are more accurate and fast. In our subsequent summary research, we may explore ways to combine large models and small models to improve the performance of large models in the field of text summarization.

ACKNOWLEDGMENT

(Zheng Liu and Heng Wang contributed equally to this work.)

REFERENCES

[1] K. Akiyama, A. Tamura, and T. Ninomiya, "Hie-BART: Document summarization with hierarchical BART," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, student Res. Workshop*, 2021, pp. 159–165.

[2] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[3] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in *Proc. 11th Conf. Eur. Chapter Assoc. Comput. linguistics*, 2006, pp. 249–256.

[4] Y. Chen and R. A. Juanatas, "Two-stage text summary model," *IEEE Access*, early access, Jul. 12, 2024, doi: 10.1109/ACCESS.2024.3427390.

[5] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3504–3514, 2021, doi: 10.1109/TASLP.2021.3124365.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[7] P. Du, Y. Gao, L. Li, and X. Li, "SGAMF: Sparse gated attention-based multimodal fusion method for fake news detection," *IEEE Trans. Big Data*, early access, Jun. 13, 2024, doi: 10.1109/TBDATA.2024.3414341.

[8] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, Mar. 2021, Art. no. 113679.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[10] B. Hu, Q. Chen, and F. Zhu, "LCSTS: A large scale Chinese short text summarization dataset," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, L. Màrquez, C. Callison-Burch, and J. Su, Eds., 2015, pp. 1967–1972.

[11] L. Huang, W. Chen, Y. Liu, S. Hou, and H. Qu, "Summarization with self-aware context selecting mechanism," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5828–5841, Jul. 2022.

[12] J. Devlin, M.-W. C. Kenton, and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.

[13] D. King, Z. Shen, N. Subramani, D. S. Weld, I. Beltagy, and D. Downey, "Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search," in *Proc. 2nd Workshop Natural Lang. Gener., Eval., Metrics (GEM)*, Abu Dhabi, United Arab Emirates, A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, and L. Perez-Beltrachini, Eds., Dec. 2022, pp. 555–571.

[14] L. Lebanoff, F. Dernoncourt, D. S. Kim, L. Wang, and W. Chang, "Learning to fuse sentences with transformers for summarization," U.S. Patent 11 620 457, Apr. 4, 2023.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Jul. 2020, pp. 7871–7880.

[16] B. Li, T. Zheng, Y. Jing, C. Jiao, T. Xiao, and J. Zhu, "Learning multiscale transformer models for sequence generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 13225–13241.

[17] P. Li, W. Lam, L. Bing, and Z. Wang, "Deep recurrent generative decoder for abstractive text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Copenhagen, Denmark, Dec. 2017, pp. 2091–2100.

[18] Y. Li, Y. Huang, W. Huang, and W. Wang, "A global and local information extraction model incorporating selection mechanism for abstractive text summarization," *Multimedia Tools Appl.*, vol. 83, no. 2, pp. 4859–4886, Jan. 2024.

[19] Y. Li, Y. Zhang, Z. Zhao, L. Shen, W. Liu, W. Mao, and H. Zhang, "CSL: A large-scale Chinese scientific literature dataset," in *Proc. 29th Int. Conf. Comput. Linguistics*, Gyeongju, South Korea, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds., Oct. 2022, pp. 3917–3923.

[20] Z. Li, J. Wu, J. Miao, and X. Yu, "News headline generation based on improved decoder from transformer," *Sci. Rep.*, vol. 12, no. 1, p. 11648, Jul. 2022.

[21] Z. Li, J. Wu, J. Miao, X. Yu, and S. Li, "A topic inference Chinese news headline generation method integrating copy mechanism," *Neural Process. Lett.*, vol. 55, no. 2, pp. 1337–1353, Apr. 2023.

[22] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[23] Y. Liu, P. Liu, D. Radev, and G. Neubig, "BRIO: Bringing order to abstractive summarization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, S. Muresan, P. Nakov, and A. Villavicencio, Eds., 2022, pp. 2890–2903.

[24] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A robustly optimized BERT pre-training approach with post-training," in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Cham, Switzerland: Springer, 2021, pp. 471–484.

[25] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156043–156070, 2021.

[26] R. Nallapati, B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn.*, Berlin, Germany, 2016, pp. 280–290.

[27] T. Nguyen, A. T. Luu, T. Lu, and T. Quan, "Enriching and controlling global semantics for text summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9443–9456.

[28] B. Pang, E. Nijkamp, W. Kryscinski, S. Savarese, Y. Zhou, and C. Xiong, "Long document summarization with top-down and bottom-up inference," in *Proc. Findings Assoc. Comput. Linguistics, EACL*, C. Dubrovnik, A. Vlachos, and I. Augenstein, Eds., 2023, pp. 1267–1284.

[29] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 9308–9319.

[30] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[32] M. Ravaut, S. Joty, and N. Chen, "SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 4504–4524.

[33] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[34] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306.

[35] H. Su, X. Zhou, H. Yu, X. Shen, Y. Chen, Z. Zhu, Y. Yu, and J. Zhou, "WeLM: A well-read pre-trained language model for Chinese," 2022, *arXiv:2209.10372*.

[36] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014, *arXiv:1409.3215*.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. 1Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu, and Q. Du, "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4453–4460.

[39] M. Wang, P. Xie, Y. Du, and X. Hu, "T5-based model for abstractive summarization: A semi-supervised learning approach with consistency loss functions," *Appl. Sci.*, vol. 13, no. 12, p. 7111, Jun. 2023.

[40] S. Wang, X. Zhao, B. Li, B. Ge, and D. Tang, "Integrating extractive and abstractive models for long text summarization," in *Proc. IEEE Int. Congr. Big Data*, Sep. 2017, pp. 305–312.

[41] L. Wu, X. Liu, and Q. Liu, "Centroid transformers: Learning to abstract with attention," 2021, *arXiv:2102.08606*.

[42] L. Xu, Q. Cui, R. Hong, W. Xu, E. Chen, X. Yuan, C. Li, and Y. Tang, "Group multi-view transformer for 3D shape analysis with spatial encoding," *IEEE Trans. Multimedia*, early access, Apr. 29, 2024, doi: 10.1109/TMM.2024.3394731.

[43] L. Xu, Z. Wang, S. Zhang, X. Yuan, M. Wang, and E. Chen, "Modeling student performance using feature crosses information for knowledge tracing," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1390–1403, 2024, doi: 10.1109/TLT.2024.3381045.

[44] S. Xu, X. Zhang, Y. Wu, and F. Wei, "Sequence level contrastive learning for text summarization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, Jun. 2022, pp. 11556–11565.

[45] Y. Xu, J. H. Lau, T. Baldwin, and T. Cohn, "Decoupling encoder and decoder networks for abstractive document summarization," in *Proc. MultiLing Workshop Summarization Summary Eval. Across Source Types Genres*, 2017, pp. 7–11.

[46] M. Yang, C. Li, Y. Shen, Q. Wu, Z. Zhao, and X. Chen, "Hierarchical human-like deep neural networks for abstractive text summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2744–2757, Jun. 2021.

[47] W. Zeng, W. Luo, S. Fidler, and R. Urtasun, "Efficient summarization with read-again and copy mechanism," 2016, *arXiv:1611.03382*.

[48] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11328–11339.

[49] X. Zhang, Y. Liu, X. Wang, P. He, Y. Yu, S.-Q. Chen, W. Xiong, and F. Wei, "Momentum calibration for text generation," 2022, *arXiv:2212.04257*.

[50] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. Awadallah, D. Radev, and R. Zhang, "SUMM^N: A multi-stage summarization framework for long input dialogues and documents," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, S. Muresan, P. Nakov, and A. Villavicencio, Eds., May 2022, pp. 1592–1604.

ZHENG LIU received the B.E. degree from Wuhan Polytechnic University, Wuhan, China, in 2022, where he is currently pursuing the M.S. degree in software engineering. His research interests include music information retrieval, artificial intelligence technology, and its application.

HENG WANG received the B.E. degree from Huazhong University of Science and Technology, in 2006, and the Ph.D. degree in engineering from Wuhan University, in 2013. He is currently a Professor with the School of Mathematics and Computer Science, Wuhan Polytechnic University. He is also a Postdoctoral Research Fellow with Alto University, Finland. His research interests include the perception characteristics of acoustic spatial parameters, artificial intelligence, and the application of 3D audio and video in virtual reality.

CONG ZHANG received the bachelor's degree in automation engineering from Huazhong University of Science and Technology, in 1993, the master's degree in computer application technology from Wuhan University of Technology, in 1999, and the Ph.D. degree in computer application technology from Wuhan University, in 2010. He is currently a Professor with the School of Electrical and Electronic Engineering, Wuhan Polytechnic University. His research interests include multimedia signal processing, multimedia communication system theory and application, and pattern recognition.

SHUAI ZHANG received the B.E. degree from Hubei Polytechnic University, Huangshi, China, in 2021. He is currently pursuing the M.S. degree in software engineering with Wuhan Polytechnic University, Wuhan. His research interests include artificial intelligence technology and its application.

• • •