A MODULAR HIERARCHICAL MODEL FOR PAPER QUALITY EVALUATION

Xi Deng[1], Shasha Li[1], Jie Yu[1], Jun Ma[1], Bin Ji[1], Wuhang Lin[1], Shezheng Song[1] and Zibo Yi[2]

[1]College of Computer, National University of Defense Technology, Changsha, China
[2]Information Research Center of Military Science PLA Academy of Military Science, Beijing, China

ABSTRACT

Paper quality evaluation is of great significance as it helps to select high quality papers from the massive amount of academic papers. However, existing models needs improvement on the interaction and aggregation of the hierarchical structure. These models also ignore the guiding role of the title and abstract in the paper text. To address above two issues, we propose a well-designed modular hierarchical model(MHM) for paper quality evaluation. Firstly, the input to our model is most of the paper text, and no additional information is needed. Secondly, we fully exploit the inherent hierarchy of the text with three encoders with attention mechanisms: a word-to-sentence(WtoS) encoder, a sentence-to-paragraph(StoP) encoder, and a paper encoder. Specifically, the WtoS encoder uses the pre-trained language model SciBERT to obtain the sentence representation from the word representation. The StoP encoder lets sentences in the same paragraph interact and aggregates them to get paragraph embeddings based on importance scores. The paper encoder does interaction among different hierarchical structures of three modules of a paper text: the paper title, abstract sentences, and body paragraphs. Then this encoder aggregates new representations generated into a compact vector. In addition, the paper encoder models the guiding role of the title and abstract, respectively, generating another two compact vectors. We concatenate the above three compact vectors and additional four manual features to obtain the paper representation. This representation is then fed into a classifier to obtain the acceptance decision, which is a proxy for papers' quality. Experimental results on a large-scale dataset built by ourselves show that our model consistently outperforms the previous strong baselines in four evaluation metrics. Quantitative and qualitative analyses further validate the superiority of our model.

KEYWORDS

Paper quality evaluation, Modular, Hierarchical, Attention mechanisms, Interact.

1. INTRODUCTION

As thousands of papers appear, evaluating their quality helps to select high-quality papers from them quickly. Naturally, this task can help readers choose good papers rapidly, assist reviewers in reviewing papers, and allow authors to self-check their papers. However, the research on this novel task still has a long way to go. As we all know, measuring innovativeness and contribution accurately is the key to evaluating the paper's quality. To the best of our knowledge, there is no publicly available expert-rated large-scale dataset on the innovativeness and contribution of papers in this field. More importantly, the models need to be pre-trained with sufficiently large

David C. Wyld et al. (Eds): ARIN, CSITA, ISPR, NBIOT, DMAP, MLCL, CONEDU -2023
pp. 09-23, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130702

real-world data of mathematical, computer, and other types of subject expertise. Even the state-of-the-art models have less rich prior knowledge about the paper than the reviewers. Therefore, models can hardly truly understand and evaluate the innovation and contribution of a paper from its text alone. We leave it to future researchers to fully understand the innovativeness and contribution of papers. However, the experimental results in this paper demonstrate that our model works well in evaluating the quality of papers. That is, the model has an excellent ability to evaluate papers after training.

In reality, the quality of a paper is tough to measure accurately with specific values. Thus most work will take the papers' acceptance as a proxy for quality evaluation. Since conferences at different levels and domains have different quality standards for papers, researchers will carry out their work on a specific level conference dataset in a particular field (e.g., top conferences in Artificial Intelligence). Or they will directly use a specific conference as a dataset to unify the quality standards of the model. Following the convention, we conduct our experiments on the ICLR conference dataset produced ourselves. But in theory, our model will also be able to find papers that meet this quality standard after being trained on any reasonable dataset with uniform quality standards.

In the literature, early studies [4–6] focus on collecting domain-specific manual features to build deterministic models to predict paper acceptance, which is a proxy for papers’ quality. However, feature engineering is time-consuming and labor-intensive, let alone the domain knowledge required. Recently, with the rapid development of deep learning, numerous neural models can be leveraged to extract these features automatically, such as MILAM [7], MHCNN [8], DeepSentiPeer [9], and HabNet [10]. Despite the much higher prediction accuracy, these models rely on extra information such as author backgrounds and review comments. However, we demonstrate that the extra information is inaccessible in double-blind reviews, hindering the model application. More importantly, these models pay less attention to the quality of the paper itself.

Some studies formulate the paper acceptance prediction task as a text classification task. A paper has multiple modules, such as the title, abstract, and body. Within each module, there are multiple hierarchical structures such as words, sentences, and paragraphs. There exists interaction and aggregation in the hierarchical structure. However, previous works [10–15] does not fully consider the position and contextual information of elements in interaction, and elements’ importance in aggregation. Meanwhile they operate on the overall representation of the module. But we demonstrate that the model works better at different levels of the three modules: title, abstract sentences, and body paragraphs. Also, textual representations in these models have difficulty capturing the information in figures, tables, formulas, and references. Therefore, we improve on the problems in the above work.

In this paper, we propose a novel modular hierarchical model (MHM) for paper quality evaluation. Our model contains three encoders from the bottom-up to capture the hierarchical structure of paper texts. Specifically, we first divide a paper text into three modules: title, abstract, and body. Then we apply the word-to-sentence (WtoS) encoder to the sentences contained in the three paper modules and obtain their sentence-level representations. Additionally, this encoder is a fine-tuned SciBERT [16] model. Next, we apply the sentence-to-paragraph (StoP) encoder to the sentence-level representations of each body paragraph. Following the DiSAN[17] framework, the StoP encoder allows sentence-level representations to interact contextually at the paragraph level and aggregate these representations into a paragraph-level representation through bi-directional self-attention and multi-dimensional attention. After that, the paper encoder takes sentence-level representations of the title and abstract sentences, and paragraph-level representations of body paragraphs as inputs, and it outputs a compact vector.

For modeling the guiding role of the title, the paper encoder does cross-attention between the title and abstract sentences, generating a compact vector. For modeling the guiding role of the abstract, the paper encoder does cross-attention between the abstract sentences and body paragraphs, generating another compact vector. For a more comprehensive representation, we concatenate the above three compact vectors and four manual features to obtain the paper representation. The four features represent the number of figures, tables, formulas, and references, respectively. Finally, the paper representation is fed into a classifier to obtain the acceptance decision, a proxy for the papers' quality.

The contributions of our work are summarized as follows:

- We produce a standard large-scale dataset of ICLR conference papers. This dataset expands the research resources in the field and is available to researchers for extensive research.
- We propose a well-designed modular hierarchical model for paper quality evaluation. This model takes into account the interactions and aggregations in the hierarchical structure of the paper more fully than existing models. And we are the first to model the guiding role of the title and abstract.
- Experimental results on ICLR conference dataset show that our model consistently outperforms the previous strong baselines in three evaluation metrics. Quantitative and qualitative analyses further validate the superiority of our model.

2. RELATED WORK

Paper quality evaluation is a novel task proposed in 2018[4, 8]. The acceptance of papers in a specific field at a particular level of the conference can be a proxy for quality assessment. All existing research in this area fails to enable models to truly understand the innovativeness and contribution of the paper, even though both are core to the assessment. Researchers use language models to extract linguistic features such as sentence readability and contextual consistency from paper texts. The models can also determine the innovation and contribution of a paper very roughly from specific words such as "SOTA," "outperform," "the first," "code publicly available," etc. Since models do not have the same extensive domain knowledge as reviewers, and the ability to reproduce the experimental results of a paper, they can easily be deceived. The model will likely assess a well-written paper that falsifies innovative and contributory results as good quality. Researchers cannot discern this deception now, leaving it for future work. The current studies assume that the authors are honest and that the paper's content is authentic and trustworthy. Research in this area is limited and still very much in its infancy, with much room for future exploration.

The length of a paper is typically thousands of words. Directly using CNNs, RNNs, or attention mechanisms on such long texts will be limited by memory and computing power. So researchers generally divide long texts to build hierarchical networks according to the paper's inherent structure. From word to full text, the structure in aggregation and interaction has three levels: word, sentence, and module.

The first is the hierarchical structure for aggregating words. Researchers generally aggregate the representations of words in a sentence to obtain an effective sentence representation. Shen et al. [12] propose a joint model combining text content with a visual rendering for document quality assessment. The model averages the word vectors directly to obtain the sentence vectors, following Shen et al. [18]. This model ignores the position and the varying significance of each word in the sentence when capturing the meaning of long sentences.

The second is the hierarchical structure for aggregating sentences. Wenniger et al. [14] propose to use HANs combined with structure tags. They use bi-directional LSTM for sentences, , which prevents parallelization. Leng et al. [13] let model learning the semantic, grammar, and innovative features of an article by three main well-designed components simultaneously. However, only sentences within a fixed-size receptive field can interact with each other. Lu et al. [15] propose MV-HATrans, a text representation model that combines multi-viewpoint information. But they ignore interaction of sentence-level representation at the paragraph level.

The third is the hierarchical structure for aggregating modules. Paper consists of the title, abstract, and body. Researchers further subdivide the body into introduction, related work, method, experimental results, and conclusion. As mentioned above, words are aggregated into sentences, and sentences are then aggregated to get modules. Researchers aggregate modules to obtain a paper representation. Yang et al. [8] propose a novel modularized hierarchical convolutional neural network to achieve automatic academic paper rating. However, they do not consider the interaction among different hierarchical structures of modules. Qiao et al. [11] improve on Yang's work by using an LSTM between the modules. Nevertheless, the one-way LSTM only allows the module to fuse its previous content.

In summary, the existing work needs to improve in terms of the position and contextual information of elements in interaction, and elements' importance in aggregation. Besides, these works do not consider the guiding role of the title and abstract.

To address the above two problems in existing work, we propose a well-designed modular hierarchical model. Firstly, we take into account as much detail as possible in the interaction and aggregation process on the hierarchical structure. Inspired by Shen [17], we use our improved version of bi-directional self-attention mechanisms to consider the location and contextual information during the interaction. We utilize multi-dimensional source2token self-attention proposed by Shen [17] to ensure distinct weights of elements in aggregation on each hierarchical structure. Secondly, we are the first to model the guiding role of the title and abstract. We leverage cross-attention mechanisms across hierarchical structures to generate two full-text representations guided by the title or abstract respectively.

3. METHODOLOGY

In this section, we first describe the problem setting, next explain the intuition behind the model, then introduce two pre-defined components, and finally present the details of our proposed model for paper quality evaluation.

3.1. Problem Setting

Our study takes the papers' acceptance as a proxy for quality evaluation. In the experiment, two assumptions may have some deviations from the real world. Assuming the content of the papers in the dataset is authentic and credible and that the authors did not falsify contribution or innovativeness. Assuming a specific conference has consistent standards of review for quality over six years. We examine the problem of predicting papers' acceptance/rejection based on a dataset D containing k papers. That is, $\mathrm{D} = \{(p_1 , \mathrm{y}_1),\ \dots, (p_\mathrm{k} , \mathrm{y}_k)\}$, where p_i is the i-thpaper's text and y_i is its corresponding conference-specific true decision. $y_i \in \{0,1\}$, where 0 means the paper is rejected, and 1 means the paper is accepted. Concretely, in the case of a paper, $p = \{t, a, b\}$, where t, a, b represent the title, abstract, and body of the paper, respectively. Assume that the title has l words, i.e.,$w^t = \{w_1^t, \dots, w_l^t\}$, where w_i^t denotes the embedding of i-th word in the title. Assume the abstract has m sentences, $s^a = \{s_1^a, \dots, s_m^a\}$, each sentence has n words. Let

$w^a_{i,j}$ with $i \in [1, m]$, $j \in [1, n]$ denotes the embedding of j-th word in the i-th sentence of the abstract. Assume that the body contains u paragraphs, $pa^b = \{pa^b_1, \ldots, pa^b_u\}$, each paragraph contains v sentences, $s^b_i = \{s^b_{i,1}, \ldots, s^b_{i,v}\}$, and each sentence contains w words. Let $w^b_{i,j,k}$ with $i \in [1, u]$, $j \in [1, v]$, $k \in [1, w]$ denotes theembedding of k-th word in the j-th sentence in the i-th paragraph of the body. Given the text p of a new paper, our goal is to predict the corresponding decision class y, which is a proxy for papers' quality. Here, we treat papers' acceptance/rejection prediction as a binary classification problem, where the class labels are the decisions y.

3.2. The Intuition Behind the Model

By treating the acceptance of a paper as a proxy for its quality, the paper quality evaluation task becomes a binary classification task that predicts whether the paper will be accepted or not. Unlike ordinary classification tasks, a paper has a length of several thousand words or even tens of thousands. CNN, RNN, and attention mechanisms are limited by current computing power resources and cannot directly use the full text as input. The time complexity of the attention mechanism used extensively in language models is quadratic in the sequence length. Common truncation on long documents breaks long-distance dependencies between tokens, resulting in performance degradation. Therefore, we refer to Yang's [19] idea of the hierarchical model to construct our model for paper quality evaluation, preserving the long-distance dependencies between tokens as much as possible.

Based on the inherent structure of the paper, we divide the paper into three modules: title, abstract, and body. Under the idea of the hierarchical composition of sentences from words, we generate context-aware sentence representations for titles and abstracts, which is what our modular hierarchical model does in the W to S encoder. According to the hierarchical idea that paragraphs consist of sentences, we generate context-aware paragraph representations for the body of the paper. We utilize the S to P encoder in our model to do this.

The paper encoder is the core of our model and the most complex part. We do not interact with the title, abstract, and body at the same level (sentence or paragraph) as in previous work. When analyzing the paper carefully, we find that the title and abstract contain much more information than the body at the same level. Specifically, although the title is at the sentence level, it summarizes the main content of the abstract. The abstract as a whole is at the paragraph level and contains the core content of the paper. Each sentence in the abstract is at the sentence level and outlines the information in one or more essential body paragraphs. The BiDiSAN module in paper encoder helps us to accomplish this interaction and obtain a full-text representation by fusing the title, abstract and body information.

However, due to the model's complexity and the dataset's insufficient size, the model relying only on the above full-text representation is inadequate to capture all the critical information. Therefore, we use two modules, T-CrossAN and A-CrossAN, to enhance the full-text representation. We use the cross-attention of the title and the abstract sentences to obtain the core representation of the abstract sentences. When the summary sentence is more relevant to the headline, it is more central. We use the cross-attention of the abstract sentence and the body paragraph to obtain the core representation of the body paragraph. The more relevant the body paragraph is to the abstract sentence, the more central it is. These two core representations containing the critical information of the paper are employed to enhance the full-text representation of the BiDiSAN module.

Our model has two drawbacks. First, the model ignores some features of the papers. The complete features of a paper include the title, abstract, body, and references. The body consists of text, tables, images, and formulas. The linguistic model with text-only input cannot access the visual information of the figures and tables in the paper. It is also difficult for the model to understand complex formulas and references. Therefore, we extract the number of figures, tables, references, and formulas to consider the features missed by the model most simply. In the subsequent work, we will let the model learn these four features end-to-end. Second, for innovation and contribution, which best reflect the quality of the paper, we can currently only judge it simply by the descriptive language of the paper. This judgment relies heavily on the integrity and writing ability of the author. We leave this to the future development of artificial intelligence. But from the experimental results in this paper, our model can still assess the quality of papers very well.

The model in this paper is well-designed for paper quality evaluation. It also can be used for other long document modeling tasks with minor modifications.

3.3. Two Pre-defined Component

Before describing the details of our approach, we briefly introduce the two pre-defined components that will be used several times in the model.

3.3.1. Improved Version of the Bi-directional Self-Attention

We improve slightly on bi-directional self-attention [17] to address the need for our paper quality evaluation task. First, computing separate attention scores for each dimension between tokens on such long texts as the paper would take much memory and computation time. Therefore, we replace the original multi-dimensional token2token self-attention with a multi-head self-attention [20], preserving positional masks. In this way, we can reduce the memory by a multiple of the token dimension size when computing the attention mechanism. It also reduces the model complexity and can be trained faster. Second, the fusion gate in the original bi-directional self-attention mechanism dynamically controls the ratio of both the original representation and the context-aware representation. This ratio is a learnable parameter of the same dimensional size as the token. It is hard to train such a complex model for a dataset of our size. Thus we replace the original fusion gate with a residual structure [21] to reduce the number of parameters and decrease the model complexity

We introduce our improved version of the bi-directional self-attention mechanism using S to P encoder as an example. The effect of this mechanism in this encoder is to allow the sentences within a paragraph to interact to obtain a context-aware sentence representation. Suppose the input is s_i^b,which is the set of all sentence representations of the i-th paragraph in the body. Specifically, $s_i^b = \{s_{i,1}^b, \dots, s_{i,v}^b\}$, $s_{i,j}^b \in \mathbb{R}^{d_e}, j \in [1, v]$.$v$ is the number of sentences contained in the i-th paragraph.$s_{i,j}^b \in \mathbb{R}^{d_e}, j \in [1, v]$. $s_{i,j}^b$ is the embedding of the j-th sentence of the i-th paragraph in the body. Q , K , and V are obtained by making different, learned linear transformations on s_i^b. We compute the dot products of Q and K, divide each by $\sqrt{d_k}$following Transformer [20]. Here we choose from two different positional masks,$M^{fw}$and M^{bw}, to use. The former makes each sentence only focus on its preceding sentences, while the latter makes each focus only on its following ones. Then we apply a softmax function to obtain the attention weights on V. Attention(Q, K, V, M) is the weighted sum on V according to the attention weights.

$$Q = W^{(1)} s_i^b \qquad K = W^{(2)} s_i^b \qquad V = W^{(3)} s_i^b \tag{1}$$

$$\mathrm{Attention}(Q,K,V,M) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{2}$$

where $W_1 \in \mathbb{R}^{d_e \times d_k}$, $W_2 \in \mathbb{R}^{d_\mathrm{e} \times d_k}$, $W_3 \in \mathbb{R}^{d_\mathrm{e} \times d_v}$. Among these, $d_k = d_v = d_e$, $d_e = 768$. M is either M^{fw} or M^{bw}.M^{fw} is a strictly lower triangular with order v.M^{bw} is a strictly upper triangular with order v.The value of all elements within the M^{fw} and M^{bw} is 1 except for 0.

In the multi-head attention mechanism, we linearly project the Q, K and V to $\widetilde{d_k}$, $\widetilde{d_k}$ and $\widetilde{d_v}$ dimensions, respectively. Then, we parallel perform scaled dot-product attention with masks in Eq.(2) on each of these projected versions. Each attention function produces $\widetilde{d_v}$-dimensional output values. They are concatenated and once again projected to obtain the final values.

$$\begin{gathered}\mathrm{MultiHead}(Q,K,V,M) = \text{ Concat }(\mathrm{head}_1, \dots, \mathrm{head}_{\,h})W^O \\ \text{where } \mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V, M\right)\end{gathered} \tag{3}$$

where the number of heads $h = 2$. $W_i^Q \in \mathbb{R}^{d_k \times \widetilde{d_k}}$, $W_i^K \in \mathbb{R}^{d_\mathrm{k} \times \widetilde{d_k}}$, $W_i^V \in \mathbb{R}^{d_\mathrm{v} \times \widetilde{d_v}}$ and $W_i^O \in \mathbb{R}^{h\widetilde{d_\mathrm{v}} \times d_e}$. Among these, $\widetilde{d_k} = \widetilde{d_v} = d_e/h$. M is either M^{fw} or M^{bw}.

Forward mask M^{fw} and backward mask M^{bw} are input into Eq.(3) separately to use bi-directional attention. We finally employ a residual connection on the results obtained by concatenation operationin Eq.(4). In this way, we add the weighted representation of the context obtained by the attention mechanism to the original representation of the token to derive a context-aware representation.

$$se_i^b = [MultiHead(Q,K,V,M^{fw})||MultiHead(Q,K,V,M^{bw})] + V \tag{4}$$

where || denotes concatenation operation. $se_i^b = \{se_{i,1}^b \dots se_{i,v}^b\}$, $se_{i,j}^b \in \mathbb{R}^{2d_e}, j \in [1, v]$. $se_{i,j}^b$ is context-aware embedding of the j-th sentence of the i-th paragraph in the body.

3.3.2. Multi-Dimensional Source2token Self-attention

Suppose the input isse_i^b calculated from Eq.(4), $se_i^b = \{se_{i,1}^b \dots se_{i,v}^b\}$,$se_{i,j}^b \in \mathbb{R}^{2d_e}, j \in [1, v]$. se_i^bareall context-aware sentence representations of the i-th paragraph in the body. We use the function $f\left(se_{i,j}^b\right)$ to calculate the dependency between $se_{i,j}^b$ and the entire sequence se_i^b. The attention weight $\alpha_{i,j}, j \in [1, v]$, of each sentence $se_{i,j}^b$ is obtained by applying softmax function on the $f\left(se_{i,j}^b\right)$. The output $pa_i^b$of this module is the weighted sum of the inputsse_i^b according to the attention weights. Formally, we have:

$$f\left(se_{i,j}^b\right) = W^T\sigma\left(W^{(1)}se_{i,j}^b + b^{(1)}\right) + b \tag{5}$$

$$\alpha_{i,j} = \frac{\exp\left(f\left(se_{i,j}^b\right)\right)}{\sum_{j=1}^{y} \exp\left(f\left(se_{i,j}^b\right)\right)} \tag{6}$$

$$pa_i^b = \sum_j \alpha_{i,j} se_{i,j}^b \tag{7}$$

where $f\left(se_{i,j}^{b}\right) \in \mathbb{R}^{2d_e}$ is a vector with the same length as $se_{i,j}^{b}$, $\sigma(\cdot)$ is an activation function, and the weight matrices $W, W^{(1)} \in \mathbb{R}^{2d_e \times 2d_e}$. b and $b^{(1)}$ are bias terms. $W, W^{(1)}, b$, and $b^{(1)}$ are all trainable parameters. $pa_i^b \in \mathbb{R}^{2d_e}$ is the embedding of the i-th paragraph in the body.

3.4. Our Approach

We divide each paper into three modules: title, abstract and body.Our proposed model takes the paper text as input and consists of four main components: WtoS encoder, StoP encoder, paper encoder, and decision predictor, as shown in Fig. 1. The following is a detailed description of the modular hierarchical model.

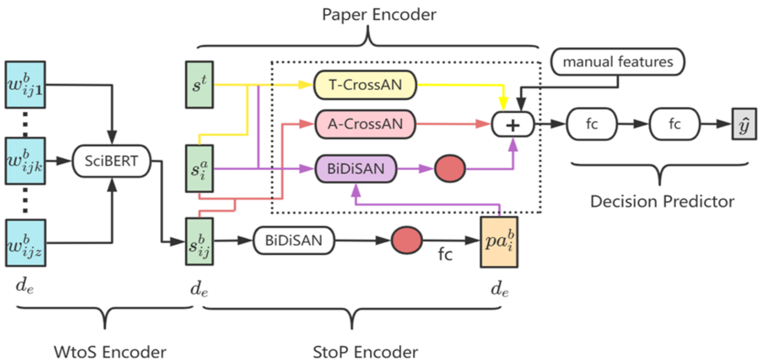

Figure 1.The architecture of our model. Note that the part inside the rectangular dashed box is the paper encoder. The part below the dashed box is the S to P encoder. The red solid circles represent the multi-dimensional source2token self-attention module.

3.4.1. W to S Encoder

The W to S encoder aims to capture the relationships between words in a sentence and the importance of each word to the meaning of the sentence. It has been demonstrated that BERT [22] models perform effective knowledge transfer from self-supervised tasks with large-scale training data. The SciBERT [16] is a variant of BERT [22]. Trained on numerous academic papers, SciBERT contains rich prior knowledge intensely relevant to our task. We tokenize each sentence using vocabulary with a maximum sequence length $L = 512$ following SciBERT [16]. We add [CLS] at the beginning and [SEP] at the end of the token sequence of each sentence. If the length of the token sequence is less than L, we append the extra [pad] token to the end. Then we feed the words token sequence in each sentence contained in the title, abstract, and body into SciBERT, taking the embedding of the [CLS] token as the sentence representation. In summary, the W to S encoder lets words in the same sentence interact and aggregates them to get sentence embeddings based on importance scores.

3.4.2. S to P Encoder

The S to P encoder aims to capture the relationships between sentences in a paragraph and the importance of each sentence to the meaning of the paragraph. The input is the sentence embedding $s_{i,j}^{b}$ of body generated by the WtoS encoder, where $i \in [1, u], j \in [1, v]$. It first generates context-aware embeddings $se_{i,j}^{b}$ for each sentence using the bidirectional self-attention as shown in Eq.(4). Based on these context-aware sentence embeddings $se_{i,j}^{b}$, the multi-dimensional source2token self-attention computes attention scores $\alpha_{i,j}$ as shown in Eq.(6). It indicates the importance of each feature of the sentence to the paragraph. This paragraph embedding is a weighted sum of the context-aware sentence representations according to the attention scores $\alpha_{i,j}$as shown in Eq.(7). We obtain pa_i^b after reducing the dimensionality through a full connection layer. This paragraph representation incorporates all the sentence information, the relationship between sentences, and the importance of sentences in the paragraph. In summary, the S to P encoder lets sentences in the same paragraph interact and aggregates them to get paragraph embeddings based on importance scores.

3.4.3. Paper Encoder

There are two crucial observations here. First, there is an interaction of information among the title, abstract sentences, and body paragraphs. Second, the title guides the abstract's core ideas and core sentences. Similarly, the abstract guides the body's core ideas and core paragraphs. Based on these two observations, the paper encoder is divided into three modules from top to bottom: T-CrossAN, A-CrossAN, BiDiSAN. (see Fig. 1)

The T-CrossAN module models the title's guiding role for the abstract. This cross-attention module first calculates the relevance scores α_i^1 of the title s^t and abstract sentences $s_i^a, i \in [1, m]$. Then it outputs a weighted sum of the embeddings of abstract sentences. The output A is the semantic representation of the core abstract guided by the title. Formally, we have:

$$Q^1 \quad = W_1^{(1)} s^t K_i^1 = W_1^{(2)} s_i^a V_i^1 = W_1^{(3)} s_i^a \tag{8}$$

$$\alpha_i^1 \quad = \text{softmax}\left(\frac{Q(K_i^1)^T}{\sqrt{d_k}}\right) \qquad A = \sum_i \alpha_i^1 V_i^1 \tag{9}$$

where $W_1^{(1)} \in \mathbb{R}^{d_e \times d_k}$, $W_1^{(2)} \in \mathbb{R}^{d_e \times d_k}$, $W_1^{(3)} \in \mathbb{R}^{d_e \times d_v}$. $d_k = d_v = d_e$.

The A-CrossAN module models the abstract's guiding role for the body. This cross-attention module calculates the relevance score between each abstract sentence $s_i^a, i \in [1, m]$ and each body paragraph $p_j^b, j \in [1, u]$. We combined the relevance scores for the abstract and body with the relevance scores for the title and abstract to obtain the final scores α_j^2. The module outputs a weighted sum of the embeddings of body paragraphs according to α_j^2. This output B is the semantic representation of the core body text guided by the abstract sentences. The more relevant an abstract sentence is to the title, the more guidance it has. Formally, we have:

$$Q_i^2 = W_2^{(1)} s_i^a K_j^2 = W_2^{(2)} p_j^b \qquad V_j^2 = W_2^{(3)} p_j^b \tag{10}$$

$$\alpha_j^2 = \sum_i \text{softmax}\left(\frac{Q_i^2 \left(K_j^2\right)^T}{\sqrt{d_k}}\right) \alpha_i^1 \qquad B = \sum_j \alpha_j^2 V_j^2 \tag{11}$$

where $W_2^{(1)} \in \mathbb{R}^{d_e \times d_k}$, $W_2^{(2)} \in \mathbb{R}^{d_e \times d_k}$, $W_2^{(3)} \in \mathbb{R}^{d_e \times d_v}$. $d_k = d_v = d_e$.

The BiDiSAN module models the interaction of the title, abstract sentences and body paragraphs. Its structure is the same as the StoP encoder. Its inputs are representation for the title, abstract sentences and body paragraphs. The output of this module is C, a weighted sum of the context-aware full-text representations according to the attention scores. C incorporates the relationship among the title, abstract and body, as well as capturing their importance in the paper.

The paper encoder concatenates A, B,and C from the above three modules and four manual features as output. These four well-designed manual features are the number of figures, tables, formulas, and references in the paper. We consider the number of figures and tables to be a rough representation of the adequacy of the paper's experiments. The number of formulas partially represents the theoretical correctness of the paper. The number of references correlates with the adequacy of the authors' research on related work. These manual features are currently tricky for deep models to learn from the long textual content. So we combine these four features with the full-text representation consisting of A, B, and C， which adds richer information.

3.4.4. Decision Predictor

We design a decision predictor to finally predict the acceptance of the academic paper, which is a proxy for papers' quality. Specifically, we take the compact representation from the paper encoder as its input. Let it pass through a fully connected layer with ELU function. This time its dimensionality is reduced from high dimension to 128. Then we feed it into the fully connected layer with Sigmoid function to get the final prediction result $\hat{y}$. Although existing research on this task, including the work in this paper, is still at a preliminary stage of exploration, we believe that paper quality evaluation will be of great value in the academic field.

4. Experiments and Results

4.1. Dataset

It should be noted that the currently widely used dataset is PeerRead [4], whose largest subset is a mixture of several top conferences. The dataset currently widely used for paper quality evaluation is the largest subset of PeerRead[4], arXiv. Papers judged as rejected by the rules in arXiv may not be truly negative samples. And with arXiv as a collection of several top conferences, quality standards that vary between conferences may affect this task. Therefore, we have produced our large dataset containing only ICLR papers. Among the top conferences, only ICLR officially provides complete, accurate data on rejected papers. In theory, our proposed model can be trained on all datasets of the same quality standard to obtain the ability to evaluate at this quality standard.

Table 1. Statistics of ICLR dataset

Year	#paper	#Acc/#Rej
ICLR2017	450	177/273
ICLR2018	804	308/496
ICLR2019	1227	449/778
ICLR2020	2024	645/1379
ICLR2021	2386	802/1584
ICLR2022	2420	1004/1416
Total	9311	3385/5926

We conduct experiments on a six-year papers dataset from the top conference ICLR to verify the validity of our model for predicting the acceptance of papers. Specifically, using the official interface provided by OpenReview, we download the main conference papers of ICLR from 2017 to 2022. Papers with an original category label of Accept (Oral), Accept (Spotlight), Accept (Talk), or Accept (Poster) are uniformly labeled as 1 (accept). Papers with an original category label of Reject or Invite to Workshop Track are uniformly labeled as 0 (reject). Then we utilize the open-source project GROBID to read the pdf file for the structured text of the paper. After cleaning the data and removing samples with text lengths longer than 8000, the dataset contains a total of 9311 samples. Table 1 shows the detailed statistics of this dataset. All samples are randomly shuffled before dividing the dataset. We follow a 7:2:1 division ratio, meaning that 6516, 1864, and 931 papers are used as the training, validation, and test sets, respectively.

4.2. Experimental Settings

All experiments are carried out on a machine with two RTX 3090 Ti having 24G of GPU memory. We use the text of the paper as the input for all models. All models use the deep learning framework PyTorch. All deep learning models are trained using BCE With Logits Loss loss function and AdamW optimizer with default parameters. For all models, we adopt the 10-fold cross-validation, using the average of their ten results for evaluation. Other hyperparameters are adjusted based on empirical settings and experimental results. The following are the detailed settings of our model. Due to memory limitations, we only fine-tune the last three layers of SciBERT in our model. The learning rate of these three layers is set to $1e$-5, while the other layers of SciBERT are frozen. The learning rate of the rest of our model is set to $3e$-4. We use a linear warm-up strategy to optimize these two learning rates, where warm-up steps are 0.2 times of total steps.Also, due to memory limitations, the batch size is only 1. Therefore, we use the gradient accumulation strategy, i.e., we update the network parameters once every 8 batches. In addition, the number of heads of multi-head attention is 2. And the epoch of the experiment is set to 20. With the data in parallel on two GPUs, the total running time of our model is 8 hours.

4.3. Baselines and Evaluation Metrics

There has been some research work on the task of predicting the acceptance of academic papers. However, these works do not publish the source code and its specific implementation details are not clear. Therefore, we do not compare our work with these due to the complexity of the reimplementation.

To verify the validity of our model, we compare our proposed model with ten baselines according to the experimental settings described above. The baseline models are specified as follows: (1) Five flat baselines: Two are traditional text classification models, including logistic regression (LR) and Bernoulli’s Bayesian (BernoulliNB). Three are deep learning models for text classification, including TextCNN[23], TextRCNN[24] and DPCNN [25]. (2) Five pre-trained models, including GPT2 [26], SciGPT2 [27], BERT-base [22], BERT-large [22] and SciBERT [16].

We use Area Under Curve (AUC), Accuracy (ACC), Macro-F1 (Ma-F1) and Micro-F1 (Mi-F1) to evaluate all models on the task of predicting the acceptance of papers. The AUC is the area under the ROC curve enclosed by the coordinate axis. Mathematically, a pair of positive and negative samples are randomly selected, and the probability that the classifier scores a positive sample more than a negative sample is AUC. Because AUC is not affected by the imbalance between positive and negative data volumes, we use it as the primary metric for our experiments. Since the model predicts values from 0 to 1, we need to set a threshold manually. We consider samples above this threshold as positive samples and those below this threshold as negative samples. As different thresholds produce different ACC, Ma-F1 and Mi-F1, we develop the criteria for selecting the thresholds in this experiment. We traverse the thresholds in 0.01 steps from [0.01,0.99] to find the optimal threshold with the goal of maximizing Mi-F1. Then ACC and Ma-F1 were calculated at that threshold. Note that in binary classification, ACC and Mi-F1 are equal.

4.4. Experimental Results

Table 2. Performance results of all models on OpenReview datasets

	Model	AUC	ACC/Mi-F1	Ma-F1
Flat Baseline	NB	0.628	0.668	0.598
	LR	0.624	0.654	0.456
	TextCNN	0.629	0.649	0.405
	TextRCNN	0.708	0.708	0.601
	DPCNN	0.705	0.708	0.600
Pre-trained Baseline	GPT2	0.642	0.668	0.536
	SciGPT2	0.678	0.665	0.611
	BERT-base	0.768	0.733	0.675
	BERT-large	0.755	0.732	0.661
	SciBERT	0.778	0.742	0.692
	MHM	**0.806**	**0.754**	**0.701**

The experimental results are shown in Table 2. MHM achieves the best performance results in each of the metrics. This demonstrates the effectiveness of our model and its generalizability over the test set. Specifically: (1) NB and LR models have the worst results on the dataset. This may be since their structures are too simple to learn deeper features. (2) Among the three deep learning models, we find that the results of TextCNN are almost close to LR. Moreover, DPCNN performs significantly better than TextCNN, which indicates that CNNs with more layers can learn deeper features and thus perform better. TextRCNN, with fewer network layers and faster training, can achieve almost the same performance as DPCNN. This fully demonstrates the advantages of RNNs in modeling long texts. The RNN model is more suitable for dealing with long-distance dependencies. (3) Among the five pre-trained models, GPT2 and SciGPT2 perform poorly. This may be because the generative models are not good at solving long text classification. Both BERT-base and BERT-large achieve excellent experimental results. It reflects the fact that pre-trained language models trained on a large-scale corpus can be of great help for downstream tasks. SciBERT, a variant of BERT, achieves the most advanced performance among the baseline models due to its rich scientific domain knowledge. (4) Compared to SciBERT, MHM exceeds 2.8%, 1.2%, and 0.9% on AUC, Ma-F1, and Mi-F1, respectively. This demonstrates the effectiveness of our model on the task of predicting the acceptance of academic papers. In addition, the decisions predicted by MHM are currently not entirely correct but still can be helpful as an aid tool for people.

4.5. Ablation Study

We perform the following ablation study on MHM to evaluate the contribution of each component. The results are shown in Table 3. The full version of MHM outperforms all variants with individual components removed, which indicates that each component is indispensable. Specifically, (1) the variant with the WtoS encoder removed has the worst performance. This indicates that words, as the most fundamental structure in a paper, play the most critical role in the overall semantics and subsequent performance of the model. (2) Removing the StoP encoder has the second-worst impact on the model. This illustrates the importance of a reasonable generation of paragraph representations. We know that the body paragraphs take up the most words in a paper. (3) The results of removing the paper encoder show that the paper encoder also has a vital role in the model. It can do interaction among the title, abstract sentences, and body paragraphs. Furthermore, it can model the guiding role of the title and abstract to obtain the core semantics of the abstract and body. (4) Although the effect of removing the feature is not as significant, it still provides useful auxiliary information for the model. The model currently has difficulty learning this information from the input text. In conclusion, all three encoders and features are crucial in improving the model's prediction performance.

Table 3. Ablation Results of MHM

Model	AUC	ACC/Mi-F1	Ma-F1
MHM	**0.806**	**0.754**	**0.701**
-WtoS Encoder	0.717 (-0.089)	0.702 (-0.052)	0.623 (-0.078)
-StoP Encoder	0.778 (-0.028)	0.731 (-0.023)	0.681 (-0.020)
-Paper Encoder	0.783 (-0.023)	0.733 (-0.021)	0.682 (-0.019)
-feature	0.789 (-0.017)	0.742 (-0.012)	0.696 (-0.005)

4.6. Case Study

We randomly sample 20 samples and visualize the attention scores to obtain the following findings: (1) In the StoP encoder, the model generally pays most attention to the first or last sentence or both the first and last sentences of the paragraph. (2) In the T-CrossAN module, the model mainly focuses on the last sentence. (3) In the A-CrossAN module, the model focuses more on the conclusion and the paragraphs near the section headings. (4) In the BiDiSAN module of the paper encoder, the model gives higher attention scores overall to both title and abstract sentences and lower attention scores to both body text. Specifically, in the abstract, the model pays the most attention to the end sentences, generally the parts describing the experiment. In the body text, the model pays more attention to the conclusion and the parts closely linked to the title. Consistent with the idea of human review, it suggests that the model does capture essential information. However, such a boilerplate analysis may miss excellent papers that do not follow this standard writing style.

We then analyze the top 10 highest-rated true positives. We find that most of them have keywords such as "code is available", "acknowledgment", "appendix" and "state-of-the-art". There is also a specific division of labor descriptions. This reflects, to some extent, the paper's innovation, contribution, meticulousness, etc.

We investigate the error cases that our model does not predict correctly on the ICLR dataset and find that: (1) 77.3% of them are false negatives; (2) 22.7% of them are false positives. Our analysis of the highest-scoring false-positive reveals that this paper included the keywords and standard structure that an excellent paper has. It looks like an outstanding paper. However, the publicly available human review comments on OpenReview indicate that this paper is too simple and does not contribute enough. It suggests that the model has difficulty learning in-depth features, such as contribution, requiring extensive domain knowledge to judge. Our analysis of the lowest-scoring false-negative finds that the model misidentifies the core sentences in the abstract. This misidentification may be why the model does not capture the correct semantics of the paper.

5. CONCLUSIONS

In this paper, we propose a modular hierarchical model(MHM) for paper quality evaluation. Specifically, we first divide the paper into three modules: title, abstract, and body. Then we use three hierarchical encoders in the model: the W to S encoder, the S to P encoder, and the paper encoder. We consider fuller interaction and aggregation in the hierarchy than in existing models. We are also the first to model the guiding role of titles and abstracts to generate core representations. Experimental results on the large-scale dataset we produced show that our model outperforms the strong baseline model by a large margin on all evaluation metrics. But our model can hardly learn the features of figures, tables, formulas and references without manual features. It also fails to truly understand the contribution and innovation of the paper. With the development of document images understanding[28,29], we plan to investigate the combination of images and text in papers in the future better learn all features of papers. We also hope that in subsequent work, the model will be able to better understand and evaluate the paper's innovation and contribution.

ACKNOWLEDGEMENTS

This work is supported by Hunan Provincial Natural Science FoundationProject (No.2022JJ30668) and (No.2022JJ30046).

REFERENCES

[1] R. E. Page, (2013) “Stories and social media: Identities and interactio”,in Routledge, New York,USA, 2013.

[2] Q. V. Dang and C. L. Ignat, (2016)“Measuring quality of collaboratively edited documents: The case of Wikipedia”, in Proc. CIC, Pittsburgh, PA, USA, pp266-275.

[3] A. Shen, B. Salehi, J. Qi and T. Baldwin, (2020) “A multimodal approach to assessing document quality”, Journal of Artificial Intelligence Research, Vol. 68, pp607-632.

[4] D. Kang, W. Ammar, B. D. Mishra, M. V. Zuylen, S. Kohlmeier et al., (2018)“A dataset of peer reviews (peerread): Collection, insights and nlp applications”, in Proc. NAACL HLT, New Orleans, pp1647–1661.

[5] M. Skorikov and S. Momen, (2020)“Machine learning approach to predicting the acceptance of academic papers”, in Proc. IAICT, Bali, Indonesia, pp113–117.

[6] P. Vincent-Lamarreand V. Larivi`ere, (2021) “Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome”, Quantitative Science Studies, Vol. 2, No.2, pp662–677.

[7] K. Wangand X. Wan, (2018)“Sentiment analysis of peer review texts for scholarly papers”, in Proc. SIGIR, Ann Arbor, MI, USA, pp175–184.

[8] P. Yang, X. Sun, W. Li and S. Ma, (2018) “Automatic academic paper rating based on modularized hierarchical convolutional neural network”, in Proc. ACL, Melbourne, Australia, pp496– 502.

[9] T. Ghosal, R. Verma, A. Ekbal and P. Bhattacharyya, (2019)“DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions”, in Proc. ACL, Florence, Italy, pp1120–1130.

[10] Z. Deng, H. Peng, C. Xia, J. Li, L. He and P. S. Yu, (2020)“Hierarchical bi-directional self-attention networks for paper review rating recommendation”, in Proc. COLING, Barcelona, Spain, pp6302–6314.

[11] F. Qiao, L. Xu and X. Han, (2018)“Modularized and attention-based recurrent convolutional neural network for automatic academic paper aspect scoring”, in Proc. WISA, Taiyuan, China, pp68–76.

[12] A. Shen, B. Salehi, T. Baldwin and J. Qi, (2019)“A joint model for multimodal document quality assessment”, in Proc. JCDL, Champaign, IL, USA, pp107–110.

[13] Y. Leng, L. Yu and J. Xiong, (2019)“Deepreviewer: Collaborative grammar and innovation neural network for automatic paper review”, in Proc. ICMI, Suzhou, China, pp395–403.

[14] G. M. de Buy Wenniger, T. van Dongen, E. Aedmaa, H. T. Kruitbosch, E. A. Valentijnet al., (2020)“Structure-tags improve text classification for scholarly document quality prediction”, in Proc. Proceedings of the First Workshop on Scholarly Document Processing, Online, pp158–167.

[15] Y. Lu, J. Luo, Y. Xiaoand H. Zhu, (2021) “Text representation model of scientific papers based on fusing multi-viewpoint information and its quality assessment”,Scientometrics, Vol. 126, No.8, pp6937–6963.

[16] I. Beltagy, K. Lo and A. Cohan, (2019) “SciBERT: A pretrained language model for scientific text”, in Proc. EMNLP-IJCNLP, Hong Kong, China, pp3615–3620.

[17] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan and C. Zhang, (2018) “Disan: Directional self-attention network for rnn/cnn-free language understanding”, in Proc. AAAI, Louisiana, USA, pp5446–5455.

[18] A. Shen, J. Qiand T. Baldwin, (2017) "A hybrid model for quality assessment of Wikipedia articles", in Proc. Proceedings of the Australasian Language Technology Association Workshop, Brisbane, Australia, pp43–52.

[19] Z. Yang, D. Yang, C. Dyer, X.He, A. Smolaet al., (2016) "Hierarchical attention networks for document classification", in Proc. NAACL-HLT,San Diego, California, pp1480–1489.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Joneset al., (2017)"Attention is all you need", in Proc. NIPS, Long Beach, CA, USA, pp5998-6008.

[21] K. He, X. Zhang, S. Ren and J. Sun, (2016)"Deep residual learning for image recognition", in Proc. CVPR, Las Vegas, NV, USA, pp770–778.

[22] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, (2019) "BERT: Pre-training of deep bidirectional transformers for language understanding", in Proc. NAACL-HLT, Minneapolis, Minnesota, pp4171–4186.

[23] Y. Kim, (2014)"Convolutional neural networks for sentence classification", in Proc. EMNLP, Doha, Qatar, pp1746–1751.

[24] S. Lai, L. Xu, K. Liu and J. Zhao, (2015)"Recurrent convolutional neural networks for text classification", in Proc. AAAI, Austin, Texas, USA, pp2267-–2273.

[25] R. Johnson and T. Zhang, (2017)"Deep pyramid convolutional neural networks for text categorization", in Proc. ACL, Vancouver, Canada, pp562–570.

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodeiet al., (2019) "Language models are unsupervised multitask learners",OpenAI blog, Vol. 1, No.8, pp9.

[27] K. Luu, X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola et al., (2021) "Explaining relationships between scientific documents", in Proc. ACL-IJCNLP, Online, pp2130–2144.

[28] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang et al., (2022) "DiT: Self-supervised Pre-training for Document Image Transformer", in Proc. ACM MM, Lisboa, Portugal, pp3530–3539.

[29] Y. Huang, T. Lv, L. Cui, Y. Lu and F. Wei, (2022) "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking", in Proc. ACM MM, Lisboa, Portugal, pp4083–4091.