IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 17 January 2024, accepted 17 March 2024, date of publication 21 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380480

RESEARCH ARTICLE

The Analysis of Deep Learning Recurrent Neural Network in English Grading Under the Internet of Things

DANDAN LI[1], WENLING LI[2], YANMEI ZHAO[3], AND XUTAO LIU[4]
[1]Department of Language and Humanities Education, Faculty of Educational Studies, Universiti Putra Malaysia, Serdang 43300, Malaysia
[2]Faculty of Educational Studies, Universiti Putra Malaysia, Serdang, Selangor 43300, Malaysia
[3]School of Foreign Languages, Yuxi Normal University, Yunnan 653100, China
[4]Department of Sport Studies, Faculty of Educational Studies, Universiti Putra Malaysia, Serdang 43300, Malaysia

Corresponding author: Yanmei Zhao (cherryxf99@yxnu.edu.cn)

ABSTRACT This work aims to investigate the use of the Recurrent Neural Network (RNN) in automated English grading. In order to achieve this, this work first constructs an automated English grading system based on the Internet of Things (IoT). Next, based on the variant of RNN called Gated Recurrent Unit (GRU), it introduces a self-attention mechanism into bidirectional GRU to form the Bidirectional-GRU_self-attention (Bi-GRU_Att) model. Simultaneously, an attention pooling (AP) mechanism is introduced into bidirectional GRU to form the Bidirectional-GRU_AP (Bi-GRU_AP) model. Comparative experiments are conducted using Chinese and English corpora to compare the performance of these two models. The results indicate that the Bi-GRU_AP model performs well on both Chinese and English datasets. On the Chinese dataset, compared to Bi-GRU_Att, Bi-GRU, and GRU, its accuracy is improved by 1.3%, 9.9%, and 19%, respectively. On the English dataset, compared to Bi-GRU_Att, Bi-GRU, and GRU, its accuracy is improved by 2.2%, 9.8%, and 19.2%, respectively. This suggests that introducing the AP module enables the model to better capture sentence information, thereby enhancing model performance. Additionally, after 20 iterations, the Bi-GRU_AP model exhibits good convergence and stability. The findings provide new insights for the development of automated English subjective grading systems based on IoT and deep learning.

INDEX TERMS Automated English grading, recurrent neural network, gated recurrent unit, self-attention mechanism, attention pooling.

I. INTRODUCTION

A. RESEARCH BACKGROUND AND MOTIVATIONS

With the advancement and widespread adoption of Internet of Things (IoT) technology, the interconnection of various devices and sensors has brought significant opportunities to the field of education [1], [2], [3]. English, as a globally used second language, faces limitations in its teaching and assessment methods, including subjectivity, inefficiency, and lack of personalization [4], [5], [6]. Therefore, from the perspective of IoT, the research on the application of Recurrent Neural Network (RNN) in English grading, combined with deep learning, is of great significance. Deep learning (DL), a crucial branch of machine learning, utilizes multi-layer neural network models for complex pattern learning and representation [7], [8], [9]. RNN, a specific type of neural network, has advantages in handling sequential data by capturing contextual information, thereby enhancing the understanding and representation of input data.

In the context of English grading, traditional methods rely on manual assessment, which is time-consuming and susceptible to subjective factors [10], [11]. Utilizing RNN-based DL allows for training based on large amounts of language data and learning more accurate and objective grading models. Furthermore, connecting various devices and sensors through IoT can collect real-time data on students' speech,

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Ming Chen.

writing, and reading, providing comprehensive and personalized assessments [12], [13], [14]. Therefore, this work aims to explore the application of RNN-based DL in English grading from an IoT perspective. It is hoped to revolutionize traditional English grading methods and provide more effective, accurate, and personalized solutions for English education assessment. Additionally, the application research of IoT-oriented RNN-based DL is expected to offer new perspectives and methods for the integration of IoT and the field of education.

B. RESEARCH OBJECTIVES

The research objective is to design an RNN-based DL model that can capture text information, thereby creating an English automatic grading system. In order to achieve this goal, this work develops an English automatic grading system based on IoT. It introduces self-attention mechanisms and attention pooling (AP) into the Bidirectional Gated Recurrent Unit (Bi-GRU) model to enhance the representation capability of input data and improve the accuracy of capturing sentence information. The performance of two GRU models is compared. Achieving these research objectives can bring significant innovation and improvement to the field of English education assessment and provide specific guidance and references for the construction of English subjective grading systems based on IoT and DL.

II. LITERATURE REVIEW

Technological progress leads to increasing research on automatic grading systems. Salim et al. constructed a ridge regression model to enhance the accuracy of automatic short-answer grading systems for Indonesian language students by improving early state-of-the-art models and methods. The results demonstrated that the system using bidirectional encoder representations from transformer model and fine-tuning methods could enhance the accuracy of short-answer grading [15]. Abdul Salam et al. proposed a hybrid approach that combined long short-term memory (LSTM) networks and grey wolf optimizer to automatically grade short-answer questions. Simulation results indicated better performance of this hybrid model, although the training time was longer [16]. Filighera et al. designed a black-box adversarial attack specifically for educational short-answer grading scenarios to study the robustness of grading models. In the attack, adjectives and adverbs were inserted into incorrect positions in student answers to deceive the model into predicting them as correct. Their study concluded with recommendations for the safer practical use of automatic grading systems [17].

In the field of English automatic grading, Zhang et al. developed a model for semi-open-ended short-answer questions without reference answers. The model integrated general and domain-specific information, utilized LSTM to learn representations in the classifier, and considered word sequence information [18]. Qi et al. proposed an active learning algorithm that combined Gaussian mixture models and sparse Bayesian learning for strategic sample selection and labeling. This approach aimed to build a classifier with joint sample distribution features to enhance students' language skills, creating a teaching assessment model suitable for current educational modes [19]. Agarwal and Chakraborty comprehensively analyzed English computer-assisted pronunciation training tools, categorized them into four types based on the technology used and studied the significant features of each category [20].

The studies above analyze automatic short-answer grading systems from various perspectives, employ different methods to enhance these systems, and achieve varying degrees of success. However, fewer studies consider the contextual information of long sentences and the mutual influence between words. Therefore, this work considers introducing the attention mechanism and AP, and constructs a new neural network model incorporating these two modules to improve the model's detection performance.

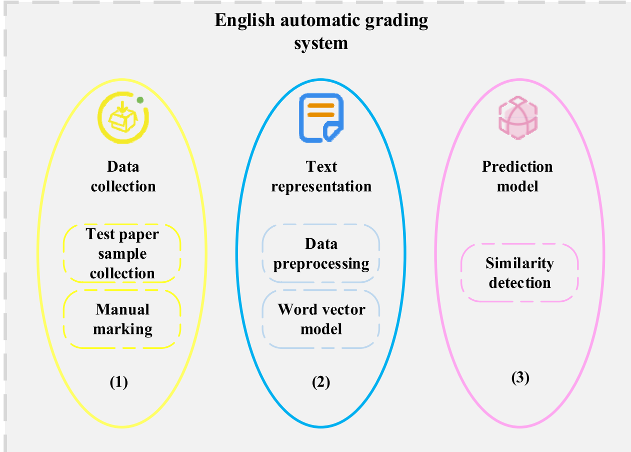

FIGURE 1. Functional modules of the english automatic grading system.

III. RESEARCH METHODOLOGY

A. AN ENGLISH AUTOMATIC GRADING SYSTEM BASED ON THE IOT

With the rise of the IoT, it is possible to leverage IoT technology to construct an intelligent and efficient English automatic grading system [21], [22]. The English automatic grading system primarily consists of three modules. They are the data collection, text representation, and prediction modules. Simultaneously, this system makes full use of IoT technologies such as sensors and cloud computing to enhance the efficiency and accuracy of the system [23], [24], [25], [26]. Figure 1 illustrates the modular structure of the system.

In Figure 1, data from student answers and standard answers are first collected using IoT sensors. Graders assess student answers based on the standard answers, assigning scores using a 0-1 label, where 1 indicates similarity and 0 indicates dissimilarity. Subsequently, the process comes to the machine processing stage. In grading subjective questions in English exams, key aspects include matching similar vocabulary, semantic relevance, and lexical matching [27], [28], [29]. In the text representation stage, the input text undergoes preprocessing, including removing spaces, numbers, and punctuation. Then, the text is tokenized, converting it from a character language to a numerical language. Following this, word vectors for text pairs are generated for training and testing, facilitating subsequent DL processing [30], [31], [32], [33]. With the preprocessing completed, a similarity measurement algorithm is chosen, and model training begins. This work utilizes RNN to construct the model. This work strategically incorporates IoT into the framework of DL RNN for English grading. The decision to integrate IoT is based on its potential advantages in enhanced data collection, connectivity, and overall system efficiency. By fully leveraging IoT technology, this work aims to provide a more comprehensive and context-aware evaluation of English language skills, thereby enhancing the robustness and effectiveness of the proposed model. This strategic use of IoT aligns with the ongoing trend of technological development, contributing to the sophistication of language assessment systems.

B. ANALYSIS OF THE BI-GRU MODEL

The core of automatic grading lies in researching text similarity, and therefore, neural network-based models are used for English text similarity detection. The choice here is the RNN algorithm. RNN can capture contextual information in sequential data, meaning that inputs from both preceding and succeeding moments influence the output at the current moment [34], [35]. This makes RNN perform well in tasks like natural language processing and speech recognition, where considering contextual relationships is crucial [36], [37], [38]. However, RNN suffers from the vanishing gradient and exploding gradient problems. When the gradient is too large, model training may fail to converge. To address this issue, researchers proposed the GRU. GRU is a variant of RNN that introduces gating mechanisms to control the flow of information and memory updates, aiming to solve the problem of long-term dependencies [39], [40], [41].

Although GRU resolves the problem of long-distance information loss, it is still unidirectional, focusing only on the forward information of the text and neglecting subsequent text information. The Bi-GRU model is proposed to comprehensively consider contextual information. Bi-GRU simultaneously processes the forward and backward information of the input sequence. The forward GRU calculates in the normal time sequence, while the backward GRU processes the input sequence in reverse. Through this bidirectional propagation, the model can utilize future and past information to reference the current time step, thereby better capturing contextual relationships in sequential data [42], [43], [44]. Figure 2 illustrates the structure of Bi-GRU.

The computation process of Bi-GRU is as follows:

$$\overrightarrow{k_t} = f(\overrightarrow{\omega} x_t + \overrightarrow{u}\, \overrightarrow{k_{t-1}} + \overrightarrow{b}) \quad (1)$$

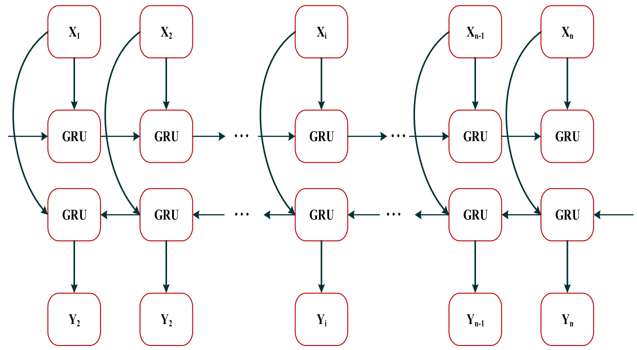

FIGURE 2. Schematic diagram of the Bi-GRU model structure.

$$\overleftarrow{k_t} = f(\overleftarrow{\omega} x_t + \overleftarrow{u}\, \overleftarrow{k_{t-1}} + \overleftarrow{b}) \quad (2)$$

$$k_t = [\overrightarrow{k_t}, \overleftarrow{k_t}] \quad (3)$$

x_t represents the input layer, $\overrightarrow{k_t}$ is the forward propagation layer, $\overleftarrow{k_t}$ is the backward propagation layer, and k_t is the output layer. The [] denotes the concatenation between two vectors. ω represents the weights of the GRU, u represents the weights of the GRU gating unit, and b refers to the bias term of the GRU. The forward and backward propagations handle past and future information, respectively. They use information from different time points to determine the output at each moment. A more comprehensive and complex set of text features is obtained for each node by concatenating the results of the forward and backward layers. This process enables a context-based comprehensive judgment [45], [46], [47].

C. BI-GRU MODEL BASED ON ATTENTION MECHANISM

The attention mechanism introduces additional learnable parameters, allowing the model to dynamically focus on relevant parts of the input sequence based on the content it is currently generating. This mechanism is widely used in natural language processing. The self-attention mechanism is an improvement upon the attention mechanism, particularly suitable for learning from long sequences [48], [49]. Incorporating the self-attention mechanism into the Bi-GRU model results in the fused Bi-GRU_Att model. By combining GRU and the attention mechanism, this model can effectively extract long-term temporal features and dependencies of sentences. It measures the attention levels of different words within a sentence to capture global information. Simultaneously, the self-attention mechanism can capture complex interactions among words within a sentence, providing the model with more information. The Bi-GRU_Att model excels in handling such tasks. Figure 3 illustrates its structure.

In Figure 3, y represents the output text representation after passing through the Bi-GRU layer, and M is the text representation after self-attention. A is a multi-level attention matrix. The Bi-GRU model fused with the self-attention mechanism features a double-layer GRU structure and can capture complex semantic features of the text and dependencies between

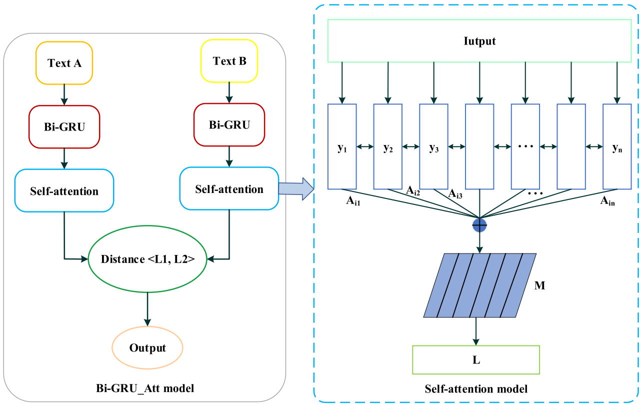

FIGURE 3. **Structure of Bi-GRU_Att model.**

forward and backward words. However, this model treats text pairs separately and does not fully consider their mutual influence. In order to address this issue, AP is introduced, forming the Bi-GRU model fused with AP (Bi-GRU_AP) to further enhance model performance. The key characteristic of AP is the construction of an AP matrix W, which generates maximum row pooling vectors and maximum column pooling vectors. The purpose is to retain the most significant features and reduce the computational complexity of the model [50], [51], [52]. Figure 4 illustrates the structure of the Bi-GRU_AP model.

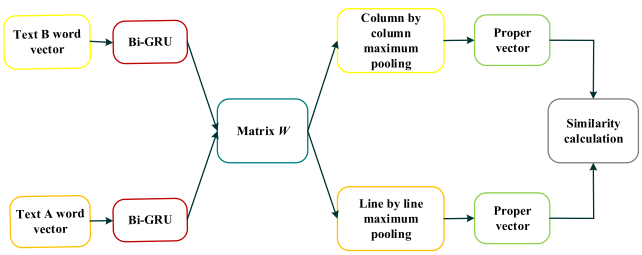

FIGURE 4. **The structure of the Bi-GRU_AP model.**

The calculation method for the matrix W is as follows:

$$W = tanh(A^T VB) \tag{4}$$

V is a parameter matrix learned through the neural network, and it can be optimized and adjusted during the model training process. Each element of the matrix W reflects the degree of matching for each word in the text pair. Performing row-based max pooling and column-based max pooling on matrix W yields two vectors, w^a and w^b. The m-th element of w^a indicates the influence of the context around the m-th word in text A on text B. Attention weight vectors φ^a and φ^b are obtained by applying the Softmax normalization to these two pooled vectors. With φ^a as an example, the i-th element of φ^a is given by:

$$[\varphi^a]_i = \frac{e^{[w^a]_i}}{\sum_{1<l<t} e^{[w^a]_l}} \tag{5}$$

t represents the length of the text, and l is the index variable. Multiplying the original text matrix by the normalized attention weight vector, the text pair representation vector is obtained as follows:

$$z^a = A\varphi^a \tag{6}$$
$$z^b = B\varphi^b \tag{7}$$

z^a and z^b are the representation vectors for texts A and B, respectively. Finally, the similarity δ between the text pairs is calculated using the cosine similarity:

$$\cos(\delta) = \frac{< z^a, z^b >}{\|z^a\| \, \|z^b\|} \tag{8}$$

D. MODEL EVALUATION INDEXES

Common model evaluation indexes include accuracy, precision, recall, and F1 score [53], [54]. Among them, accuracy represents the proportion of correctly predicted samples out of the total number of samples, and its calculation is as follows:

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

Recall suggests the proportion of true positive samples among the actual positive samples. The calculation reads:

$$Rec = \frac{TP}{TP + FN} \tag{10}$$

Precision represents the proportion of true positive samples among the samples predicted as the positive class. The calculation for precision reads:

$$Pre = \frac{TP}{TP + FP} \tag{11}$$

F1 score is the harmonic mean of precision and recall, balancing precision and recall. The calculation for the F1 score reads:

$$F1 = 2 * \frac{Pre * Rec}{Pre + Rec} \tag{12}$$

True Positive (TP) refers to correctly predicted samples of this class; False Negative (FN) refers to the prediction of this type of label as a sample of another type; False Positive (FP) refers to mistakenly predicting other classes as samples of this class; True Negative (TN) refers to samples that do not belong to this category and are predicted to be not in this category.

IV. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

A. DATASETS COLLECTION

This experiment uses two datasets: a Chinese dataset and an English dataset. The Chinese dataset consists of exercises from five school English translation exams, primarily focusing on English-to-Chinese translation. The dataset may contain samples with invalid or no student answers, requiring cleaning. The method for constructing similar text pairs involves selecting student answers similar to the standard

TABLE 1. The experimental environment and parameter settings.

The experimental environment/parameters	Model/Setting Values
Operating System	64-bit Windows10
Central Processing Unit (CPU)	Intel(R)Core(TM)i7-7700 3.60GHz
Memory	16GB
Graphics Card (GPU)	NVIDIA GeForce GTX 1080×4
Video Memory	8GB
Programming Language	Python 3.9
DL framework	Tensorflow
Initial learning rate	0.001
Loss function	Cross-Entropy Function
Optimization function	Adam
Dropout	0.5

answer, combining them with the standard answer, and randomly permuting them. The pairs labeled as 1 indicate similarity between them. Student answers from different questions are randomly combined and labeled as 0 for dissimilar text pairs, indicating dissimilarity. A total of 1100 text pairs are constructed.

This work also utilizes the English dataset Sentences Involving Compositional Knowledge (SICK) to enhance the reliability and accuracy of the English automatic grading system. The SICK dataset comprises 9927 pairs of English sentences, each with a similarity score between 1 and 5. In order to align the data labels with the Chinese dataset, sentence pairs with similarity scores of 4 and 5 are selected from the SICK dataset and labeled as 1. Similarly, sentence pairs with similarity scores of 1 and 2 are labeled as 0, creating the English corpus for this work.

B. EXPERIMENTAL ENVIRONMENT AND PARAMETERS SETTING

This experiment utilizes the Numpy library for data processing, and the jieba segmentation is applied to tokenize the Chinese corpus. Table 1 contains the remaining experimental environment and parameter settings.

C. PERFORMANCE EVALUATION

1) COMPARISON OF RESULTS FROM DIFFERENT MODELS

Using the same Chinese and English corpus, a comparison is made among the Bi-GRU_Att model with self-attention mechanism, Bi-GRU_AP model with introduced AP, Bi-GRU model, and GRU model. Figure 5 illustrates the results on the Chinese dataset.

Figure 5 reveals that among the models based on the Bi-GRU model, Bi-GRU_AP has the best performance in terms of accuracy, recall, and F1 score, with values of 83.8%, 78.8%, and 81.3%, respectively. Compared to Bi-GRU, the accuracy of Bi-GRU_Att, which incorporates attention mechanisms, increases by 8.6%. Compared to GRU, Bi-GRU with bidirectional information improves accuracy by 9.1%.

Figure 6 presents the results on the English dataset.

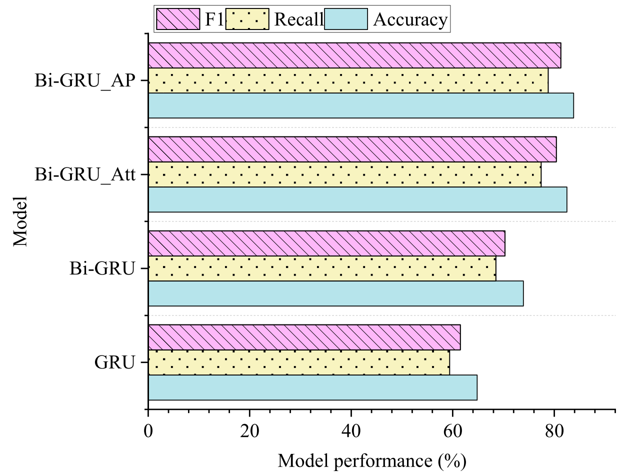

FIGURE 5. Comparison of detection results for different algorithms on the chinese dataset.

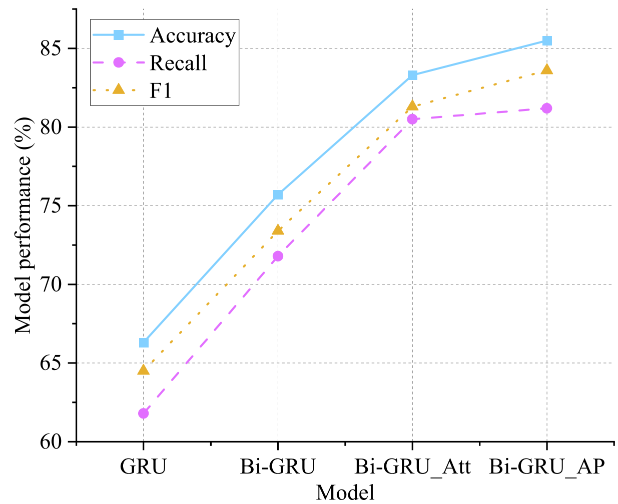

FIGURE 6. Comparison of detection results for different algorithms on the english dataset.

Figure 6 demonstrates consistent results with the Chinese dataset, where Bi-GRU_AP performs the best in terms of accuracy, recall, and F1 score, with values of 85.5%, 81.2%, and 83.4%, respectively. Moreover, combined with Figure 5, these data suggest that models need sufficient flexibility and depth for long sentences. The purpose is to capture cross-sentence correlations, context, backward information, and the interaction between words in different sentences, thereby improving accuracy and semantic understanding capabilities.

2) CONVERGENCE AND STABILITY ANALYSIS OF THE BI-GRU_AP MODEL

It is confirmed that the Bi-GRU_AP model performs the best. This model's loss values and accuracy after 20 iterations are researched to assess the convergence and performance stability. Figure 7 illustrates the loss values of the Bi-GRU_AP model after 20 iterations on the Chinese and English datasets.

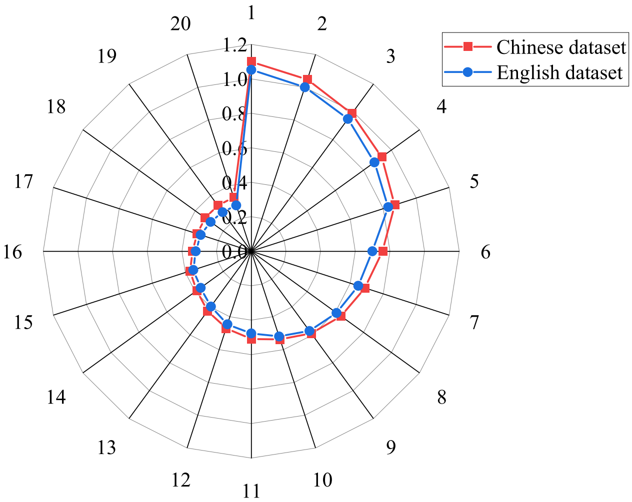

FIGURE 7. Loss values after 20 iterations of the Bi-GRU_AP model.

Figure 7 reveals that as the number of iterations increases, the model's loss values gradually decrease, stabilizing at 0.328 for the Chinese dataset and 0.280 for the English dataset. This indicates that the model gradually learns better parameter configurations during training to minimize the error between predicted and actual values. After 18 iterations, the model tends to converge, showing no significant improvement.

Figure 8 depicts the accuracy of the Bi-GRU_AP model after 20 iterations.

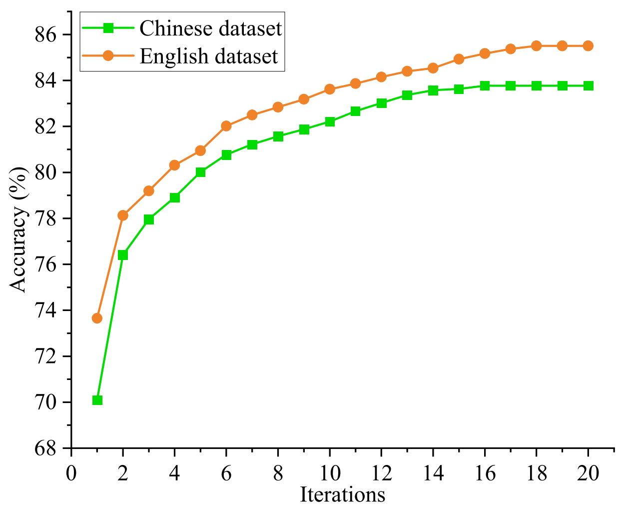

FIGURE 8. Accuracy of the Bi-GRU_AP model after 20 iterations.

Figure 8 demonstrates that as the number of iterations increases, the model's accuracy gradually improves, stabilizing at 85.5% for the Chinese dataset and 83.8% for the English dataset. This indicates that the model's performance on the grading task for both Chinese and English datasets gradually improves, and after 18 iterations, it achieves a relatively stable performance level. Considering both Figure 7 and Figure 8, the Bi-GRU_AP model demonstrates good convergence and stability.

D. DISCUSSION

Additionally, recent research works in deep neural networks indicate that neural network models incorporating attention mechanisms indeed exhibit better performance. Huang et al. proposed a lightweight tongue image segmentation network for an essential step in automatic tongue diagnosis and introduced an attention mechanism to enhance crucial features and suppress irrelevant ones. The results showed that the network demonstrated competitive performance on two tongue image datasets and accurately extracted the tongue, meeting practical application requirements [55]. Ma et al. introduced a DL model, CTR_YOLOv5n, which combined coordinate attention mechanisms and a Swin Transformer for identifying common corn leaf diseases (leaf spot, gray leaf spot, and rust) in mobile applications. Experimental results indicated that CTR_YOLOv5n achieved an average recognition accuracy of 95.2%, a 2.8 percentage point improvement over the original model [56]. Zhang et al. addressed the issue of tobacco leaf maturity recognition. They proposed a lightweight method for tobacco leaf maturity recognition based on the MobileNetV1 model, feature pyramid network, and attention mechanism. By introducing the attention mechanism, this model effectively integrated high-level semantic information and low-level positional information, identifying tobacco leaf boundaries [57].

Furthermore, in the research on automated scoring systems, Süzen et al. enhanced the reliability of human grading, particularly for short-answer questions, by applying technologies such as data mining to automated scoring. They aimed to provide valuable feedback to students and developed a similarity-based score prediction model [58]. Tan et al. also focused on short-answer questions, utilizing a dual-layer graph convolutional network to encode non-directed heterogeneous graphs of all student responses. This approach aimed to improve the performance of the automated short-answer scoring system, and their results indicated that their model outperformed other models [59], [60]. In comparison to these two studies, this work achieves a notable advantage by introducing an AP module in a Bi-GRU network, achieving an accuracy of 85.5%. In summary, an increasing number of studies suggest that integrating attention mechanism modules into other deep neural networks can lead to better results, thereby enhancing the feasibility of practical applications.

V. CONCLUSION

A. RESEARCH CONTRIBUTION

This work aims to investigate the English automatic grading system based on IoT and DL. It utilizes the variant of RNN, GRU, as a foundation to develop two models. The first is the Bi-GRU_Att model, which introduces self-attention mechanisms into the Bi-GRU network. The second is the Bi-GRU_AP model, incorporating AP into the Bi-GRU network.

The performance of these two models is experimentally compared, leading to the following conclusions:

1) On both Chinese and English datasets, the Bi-GRU_AP model demonstrates superior performance. On the Chinese dataset, compared to Bi-GRU_Att, Bi-GRU, and GRU, its accuracy increases by 1.3%, 9.9%, and 19%, respectively. On the English dataset, compared to Bi-GRU_Att, Bi-GRU, and GRU, its accuracy increases by 2.2%, 9.8%, and 19.2%, respectively. This suggests that introducing the AP module allows the model to better capture sentence information, thereby improving accuracy and semantic understanding.

2) With an increase in the number of iterations, the Bi-GRU_AP model's loss gradually decreases, stabilizing at 0.328 for the Chinese and 0.280 for the English datasets. After 18 iterations, the model converges, indicating good convergence properties.

3) As the number of iterations increases, the Bi-GRU_AP model's accuracy gradually improves, stabilizing at 85.5% for the Chinese dataset and 83.8% for the English dataset after 18 iterations. This indicates that the model has a relatively stable performance level.

B. FUTURE WORKS AND RESEARCH LIMITATIONS

Although the proposed model has shown better performance than other models, there are still areas for improvement in practical applications. During text preprocessing, manual labeling of the training set is required, which introduces a certain workload. Future research could explore more precise representation methods, such as dynamic and static word vectors. Additionally, various factors influence the scores of subjective English questions, such as students' handwriting, which the DL algorithm currently does not consider. It may be worthwhile to explore approaches like computer-based answers to mitigate the impact of handwriting on the evaluation.

ACKNOWLEDGMENT

(Dandan Li and Wenling Li are co-first authors.)

REFERENCES

[1] W. Yang, "Artificial intelligence education for young children: Why, what, and how in curriculum design and implementation," *Comput. Educ., Artif. Intell.*, vol. 3, Sep. 2022, Art. no. 100061.

[2] P. K. Malik, R. Singh, A. Gehlot, S. V. Akram, and P. Kumar Das, "Village 4.0: Digitalization of village with smart Internet of Things technologies," *Comput. Ind. Eng.*, vol. 165, Mar. 2022, Art. no. 107938.

[3] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two decades of artificial intelligence in education," *Educ. Technol. Soc.*, vol. 25, no. 1, pp. 28–47, 2022.

[4] L. Zhou, "Construction of English writing hybrid teaching model based on machine learning automatic composition scoring system," *Proc. Comput. Sci.*, vol. 208, pp. 384–390, Jan. 2022.

[5] S. Pathak, R. Raj, K. Singh, P. K. Verma, and B. Kumar, "Development of portable and robust cataract detection and grading system by analyzing multiple texture features for tele-ophthalmology," *Multimedia Tools Appl.*, vol. 81, no. 16, pp. 23355–23371, Mar. 2022.

[6] W. Liawrungrueang, P. Kim, V. Kotheeranurak, K. Jitpakdee, and P. Sarasombath, "Automatic detection, classification, and grading of lumbar intervertebral disc degeneration using an artificial neural network model," *Diagnostics*, vol. 13, no. 4, p. 663, Feb. 2023.

[7] P. Limna, S. Jakwatanatham, S. Siripipattanakul, P. Kaewpuang, and P. Sriboonruang, "A review of artificial intelligence (AI) in education during the digital era," *Adv. Knowl. Executives*, vol. 1, no. 1, pp. 1–9, Jul. 2022.

[8] E. del Gobbo, A. Guarino, B. Cafarelli, L. Grilli, and P. Limone, "Automatic evaluation of open-ended questions for online learning. A systematic mapping," *Stud. Educ. Eval.*, vol. 77, Jun. 2023, Art. no. 101258.

[9] M. D. Johansyah, B. Foster, V. Maharani, G. Prakarsa, and S. Purnama, "Improving the value of lobster selling with grading method using machine vision technology," *Linguistics Culture Rev.*, vol. 6, no. 1, pp. 596–605, Jan. 2022.

[10] M. H. Mustaqim, S. Akmal, M. Marwan, R. Sylvia, and S. Siraj, "Perspective of English lecturers on reflective paper," *Innov., J. Social Sci. Res.*, vol. 3, no. 2, pp. 14991–14999, Jun. 2023.

[11] Z. Yang, X. Kou, and Y. Wang, "Research on English teaching reading quality assessment based on cognitive diagnostic assessment," *Int. J. Continuing Eng. Educ. Life-Long Learn.*, vol. 1, no. 1, pp. 388–402, Jun. 2024.

[12] X. Zhai, P. He, and J. Krajcik, "Applying machine learning to automatically assess scientific models," *J. Res. Sci. Teaching*, vol. 59, no. 10, pp. 1765–1794, Apr. 2022.

[13] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, "Interrater reliability of sleep stage scoring: A meta-analysis," *J. Clin. Sleep Med.*, vol. 18, no. 1, pp. 193–202, Jan. 2022.

[14] S. S. Abdullah and M. P. Rajasekaran, "Automatic detection and classification of knee osteoarthritis using deep learning approach," *La Radiologia Medica*, vol. 127, no. 4, pp. 398–406, Mar. 2022.

[15] H. R. Salim, C. De, N. D. Pratamaputra, and D. Suhartono, "Indonesian automatic short answer grading system," *Bull. Electr. Eng. Informat.*, vol. 11, no. 3, pp. 1586–1603, Jun. 2022.

[16] M. Abdul Salam, M. A. El-Fatah, and N. F. Hassan, "Automatic grading for Arabic short answer questions using optimized deep learning model," *PLoS One*, vol. 17, no. 8, Aug. 2022, Art. no. e0272269.

[17] A. Filighera, S. Ochs, T. Steuer, and T. Tregel, "Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs," *Int. J. Artif. Intell. Educ.*, vol. 23, pp. 1–31, Jul. 2023.

[18] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions," *Interact. Learn. Environments*, vol. 30, no. 1, pp. 177–190, Jan. 2022.

[19] S. Qi, L. Liu, B. S. Kumar, and A. Prathik, "An English teaching quality evaluation model based on Gaussian process machine learning," *Expert Syst.*, vol. 39, no. 6, p. e1286, Jul. 2022.

[20] C. Agarwal and P. Chakraborty, "A review of tools and techniques for computer aided pronunciation training (CAPT) in English," *Educ. Inf. Technol.*, vol. 24, no. 6, pp. 3731–3743, Jul. 2019.

[21] E. Abdelmaksoud, S. El-Sappagh, S. Barakat, T. Abuhmed, and M. Elmogy, "Automatic diabetic retinopathy grading system based on detecting multiple retinal lesions," *IEEE Access*, vol. 9, pp. 15939–15960, 2021.

[22] Y. Yu, S. A. Velastin, and F. Yin, "Automatic grading of apples based on multi-features and weighted K-means clustering algorithm," *Inf. Process. Agricult.*, vol. 7, no. 4, pp. 556–565, Dec. 2020.

[23] E. Park, K. Lee, T. Han, and H. S. Nam, "Automatic grading of stroke symptoms for rapid assessment using optimized machine learning and 4-limb kinematics: Clinical validation study," *J. Med. Internet Res.*, vol. 22, no. 9, Sep. 2020, Art. no. e20641.

[24] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022.

[25] C. Bhardwaj, S. Jain, and M. Sood, "Deep learning-based diabetic retinopathy severity grading system employing quadrant ensemble model," *J. Digit. Imag.*, vol. 34, no. 2, pp. 440–457, Mar. 2021.

[26] H. Chopra, H. Singh, M. S. Bamrah, F. Mahbubani, A. Verma, N. Hooda, P. S. Rana, R. K. Singla, and A. K. Singh, "Efficient fruit grading system using spectrophotometry and machine learning approaches," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16162–16169, Jul. 2021.

[27] T. Zhang, "Design of English learning effectiveness evaluation system based on K-means clustering algorithm," *Mobile Inf. Syst.*, vol. 2021, pp. 1–9, Nov. 2021.

[28] N. Narwianta, D. A. L. Bharati, and D. Rukmini, "The evaluation of higher order thinking skills in English school nationally standardized examination at state senior high school 6 Semarang," *English Educ. J.*, vol. 9, no. 3, pp. 316–326, Jun. 2019.

[29] P. Gao, "VIKOR method for intuitionistic fuzzy multi-attribute group decision-making and its application to teaching quality evaluation of college English," *J. Intell. Fuzzy Syst.*, vol. 42, no. 6, pp. 5189–5197, Apr. 2022.

[30] P. Li, F. Zhang, A. Yu, and X. Zhao, "Language history questionnaire (LHQ3): An enhanced tool for assessing multilingual experience," *Bilingualism, Lang. Cognition*, vol. 23, no. 5, pp. 938–944, Nov. 2020.

[31] 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.

[32] X. Li, W. Chen, Q. Zhang, and L. Wu, "Building auto-encoder intrusion detection system based on random forest feature selection," *Comput. Secur.*, vol. 95, Aug. 2020, Art. no. 101851.

[33] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: A survey and trends," *J. Intell. Inf. Syst.*, vol. 54, no. 2, pp. 391–424, Apr. 2020.

[34] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *Int. J. Forecasting*, vol. 37, no. 1, pp. 388–427, Jan. 2021.

[35] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, "Recurrent neural network wave functions," *Phys. Rev. Res.*, vol. 2, no. 2, Jun. 2020, Art. no. 023358.

[36] 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.

[37] K. Shang, Z. Chen, Z. Liu, L. Song, W. Zheng, B. Yang, S. Liu, and L. Yin, "Haze prediction model using deep recurrent neural network," *Atmosphere*, vol. 12, no. 12, p. 1625, Dec. 2021.

[38] M. A. Khan, "HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system," *Processes*, vol. 9, no. 5, p. 834, May 2021.

[39] A. Dutta, S. Kumar, and M. Basu, "A gated recurrent unit approach to Bitcoin price prediction," *J. Risk Financial Manage.*, vol. 13, no. 2, p. 23, Feb. 2020.

[40] S. Haidong, C. Junsheng, J. Hongkai, Y. Yu, and W. Zhantao, "Enhanced deep gated recurrent unit and complex wavelet packet energy moment entropy for early fault prognosis of bearing," *Knowl.-Based Syst.*, vol. 188, Jan. 2020, Art. no. 105022.

[41] S. U. Rehman, M. Khaliq, S. I. Imtiaz, A. Rasool, M. Shafiq, A. R. Javed, Z. Jalil, and A. K. Bashir, "DIDDOS: An approach for detection and identification of distributed denial of service (DDoS) cyberattacks using gated recurrent units (GRU)," *Future Gener. Comput. Syst.*, vol. 118, pp. 453–466, May 2021.

[42] J. Zhang, Y. Jiang, S. Wu, X. Li, H. Luo, and S. Yin, "Prediction of remaining useful life based on bidirectional gated recurrent unit with temporal self-attention mechanism," *Rel. Eng. Syst. Saf.*, vol. 221, May 2022, Art. no. 108297.

[43] A. Kumar, S. Abirami, T. E. Trueman, and E. Cambria, "Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit," *Neurocomputing*, vol. 441, pp. 272–278, Jun. 2021.

[44] J. Yang, F. Yang, Y. Zhou, D. Wang, R. Li, G. Wang, and W. Chen, "A data-driven structural damage detection framework based on parallel convolutional neural network and bidirectional gated recurrent unit," *Inf. Sci.*, vol. 566, pp. 103–117, Aug. 2021.

[45] Q. Zhou, C. Zhou, and X. Wang, "Stock prediction based on bidirectional gated recurrent unit with convolutional neural network and feature selection," *PLoS One*, vol. 17, no. 2, Feb. 2022, Art. no. e0262501.

[46] H. Gul, N. Javaid, I. Ullah, A. M. Qamar, M. K. Afzal, and G. P. Joshi, "Detection of non-technical losses using SOSTLink and bidirectional gated recurrent unit to secure smart meters," *Appl. Sci.*, vol. 10, no. 9, p. 3151, Apr. 2020.

[47] Pamir, N. Javaid, U. Qasim, A. S. Yahaya, E. H. Alkhammash, and M. Hadjouni, "Non-technical losses detection using autoencoder and bidirectional gated recurrent unit to secure smart grids," *IEEE Access*, vol. 10, pp. 56863–56875, 2022.

[48] D. Soydaner, "Attention mechanism in neural networks: Where it comes and where it goes," *Neural Comput. Appl.*, vol. 34, no. 16, pp. 13371–13385, May 2022.

[49] S. Rao, A. K. Verma, and T. Bhatia, "Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data," *Expert Syst. Appl.*, vol. 217, May 2023, Art. no. 119594.

[50] R. Xu, Y. Zheng, X. Wang, and D. Li, "Person re-identification based on improved attention mechanism and global pooling method," *J. Vis. Commun. Image Represent.*, vol. 94, Jun. 2023, Art. no. 103849.

[51] M. Patrick, D. Campbell, Y. M. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12493–12506.

[52] L. Sun, S. Cheng, Y. Zheng, Z. Wu, and J. Zhang, "SPANet: Successive pooling attention network for semantic segmentation of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4045–4057, 2022.

[53] M.-D. Yang, H.-H. Tseng, Y.-C. Hsu, and H. P. Tsai, "Semantic segmentation using deep learning with vegetation indices for Rice lodging identification in multi-date UAV visible images," *Remote Sens.*, vol. 12, no. 4, p. 633, Feb. 2020.

[54] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, Jan. 2020.

[55] X. Huang, L. Zhuo, H. Zhang, X. Li, and J. Zhang, "Lw-TISNet: Lightweight convolutional neural network incorporating attention mechanism and multiple supervision strategy for tongue image segmentation," *Sens. Imag.*, vol. 23, no. 1, p. 6, Jan. 2022.

[56] L. Ma, Q. Yu, H. Yu, and J. Zhang, "Maize leaf disease identification based on YOLOv5n algorithm incorporating attention mechanism," *Agronomy*, vol. 13, no. 2, p. 521, Feb. 2023.

[57] Y. Zhang, Y. Zhu, X. Liu, Y. Lu, C. Liu, X. Zhou, and W. Fan, "In-field tobacco leaf maturity detection with an enhanced MobileNetV1: Incorporating a feature pyramid network and attention mechanism," *Sensors*, vol. 23, no. 13, p. 5964, Jun. 2023.

[58] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," *Proc. Comput. Sci.*, vol. 169, pp. 726–743, Jan. 2020.

[59] H. Tan, C. Wang, Q. Duan, Y. Lu, H. Zhang, and R. Li, "Automatic short answer grading by encoding student responses via a graph convolutional network," *Interact. Learn. Environments*, vol. 31, no. 3, pp. 1636–1650, Apr. 2023.

[60] X. Li, B. Li, and S.-J. Cho, "Empowering Chinese language learners from low-income families to improve their Chinese writing with ChatGPT's assistance afterschool," *Languages*, vol. 8, no. 4, p. 238, Oct. 2023, doi: 10.3390/languages8040238.