IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 21, 2021, accepted September 3, 2021, date of publication September 6, 2021,
date of current version September 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3110683

A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring

JIN XUE, XIAOYI TANG, AND LIYAN ZHENG
School of Foreign Studies, University of Science and Technology Beijing, Haidian, Beijing 10083, China

Corresponding author: Xiaoyi Tang (gzutxy@gmail.com)

This work was supported in part by Beijing Social Science Fund under Grant 19YYB008, and in part by the Fundamental Research Funds for University of Science and Technology Beijing under Grant FRF-BY-19-007.

ABSTRACT The task of automated essay scoring (AES) continues to attract interdisciplinary attention due to its commercial and educational importance as well as related research challenges. Traditional AES approaches rely on handcrafted features, which are time-consuming and labor-intensive. Neural network approaches have recently given fantastic results in AES without feature engineering, but they usually require extensive annotated data. Moreover, most of the existing AES models only report a single holistic score without providing diagnostic information about various dimensions of writing quality. Focusing on these issues, we develop a novel approach using multi-task learning (MTL) with fine-tuning Bidirectional Encoder Representations from Transformers (BERT) for multi-dimensional AES tasks. As a state-of-the-art pre-trained language model, a BERT-based approach can improve AES tasks with limited training data. Meanwhile, we deal with long texts by proposing a hierarchical method and using the attention mechanism to automatically determine the contribution of different fractions of the input essay to the final score. For the multi-topic essay scoring tasks on the ASAP dataset, results reveal that our approach outperforms the average quadratic weighted Kappa (QWK) score by 4.5% compared with the strong baseline. We propose a self-collected dataset of **C**hinese **E**FL **L**earners' **A**rgumentation (CELA) to provide valuable information about writing quality from multiple rating dimensions, including holistic and five analytic scales. For the multi-rating dimensional essay scoring tasks on the CELA dataset, experimental results demonstrate that our model increases the average QWK score by 8.1% compared with the strong baseline.

INDEX TERMS Multi-dimensional essay scoring, transfer learning, BERT, multi-task learning.

I. INTRODUCTION

The task of automated essay scoring (AES) draws interdisciplinary interest in linguistics [1], [2], education [3]–[5] and natural language processing (NLP) [6]–[8]. Existing AES models can be mainly classified into two types: traditional approaches using handcrafted features and neural network approaches using automatic feature selection with raw texts. The disadvantages of the first subtype are that features must be manually chosen to fit the model and that extra effort is required to perform effectively on various tasks. Attempts to solve this dilemma have resulted in the development of the neural network approach. Recent advances in neural network approaches have yielded promising results without the use of handcrafted features [9]–[13]. However, deep neural networks (DNNs) require vast quantities of labeled data for specific tasks, which is not always accessible. A considerable amount of research has indicated that pre-trained models can be fine-tuned to fit various tasks without training new models from scratch. As the most advanced pre-trained language model [14], Bidirectional Encoder Representations from Transformers (BERT) [15] is based on a multi-layer bidirectional transformer, and has achieved fruitful results in a variety of language-based tasks. Little research has been conducted to utilize the pre-trained language model BERT for AES tasks.

Meanwhile, the vast majority of the existing AES tasks are based on holistic scoring, assigning a single score to an essay on a specific topic based on its overall impression. However, holistic scoring has been criticized for its weakness in providing specific feedback for writing improvement [16]. Providing scores from different analytic rating dimensions

The associate editor coordinating the review of this manuscript and approving it for publication was Easter Selvan Suviseshamuthu.

could help raters and students distinguish between various dimensions of writing quality and make improvements for each dimension.

Focusing on the problems and arguments mentioned above, we propose a BERT-based transfer learning approach to predict multi-dimensional scores jointly without any feature engineering. This joint learning effectively increases the training model's sample size compared with learning multiple tasks individually, thereby improving the AES models' generalization ability and performance.

The main contributions of this paper are as follows.

- We developed a novel method using multi-task learning with fine-tuning BERT for multi-dimensional essay scoring tasks. It can jointly incorporate various aspects of features without any additional handcrafted engineering.
- We dealt with long essays by proposing a hierarchical method and using the attention mechanism to automatically determine the contribution of different fractions of the input essay to the final score. Compared with the traditional truncation approach, the Hierarchical + Attention Pooling approach effectively enhances the performance since the model can capture the combined representation of the entire essay.
- We achieved state-of-the-art results for multi-topic scoring tasks on the widely used Automated Student Assessment Prize (ASAP) dataset. Experimental results indicate that our approach promotes performance through the underlying shared representation of different topics, thereby improving the generalization ability and performance of the AES model.
- We proposed a self-collected Chinese EFL Learners' Argumentation (CELA) dataset with multiple rating dimensions. To the best of our knowledge, the CELA dataset is the first multi-rating dimensional dataset for automated essay scoring tasks.

II. RELATED WORK

In the literature, a brief review of rating scales used for writing assessment and different approaches to AES are covered to demonstrate the new trends for AES tasks. Also, numerous studies applying transfer learning and multi-task learning in AES tasks are presented.

A. RATING SCALES USED FOR WRITING ASSESSMENT

A rational writing rubric can be essential for score reliability and model construction. Over the last few decades, numerous studies have been conducted to investigate the strengths and weaknesses of holistic and analytic dimensions for writing assessment.

Many assessment programs rely on assigning a single holistic score to an essay. However, the holistic score does not offer helpful guidelines about the different dimensions of writing quality, leading to validity and reliability issues [16]. In contrast, analytic scoring often has to incorporate different scores given to different essay traits. Scholars have suggested that analytic scoring highlights the different dimensions of writing quality. Therefore, raters can quickly understand the level of the writing quality. More recently, empirical evidence has shown that analytic scores are more trustworthy than holistic scores. In general, providing scores at different analytic scales of essay quality can help raters distinguish the writing quality from different dimensions and thus enhance the reliability and validity of the AES models.

B. APPROACHES TO AES

1) TRADITIONAL APPROACHES USING HANDCRAFTED FEATURES

As the earliest AES system, PEG utilizes regression methods to predict the essay quality with surface linguistic features. Traditional AES approaches mainly address holistic scoring, which extracts predefined surface linguistic features such as morphology, syntax, and semantics, and substitutes them into learning algorithms, such as linear regression [17]–[19], support vector regression [20]–[22], logistic regression [23], [24] and Bayesian network classification [25]. One limitation of the traditional approaches is that they usually rely on manually extracted features with deep-level linguistic information being ignored.

Recent developments in corpus linguistics and computational linguistics allow more linguistic features to be extracted from essays. If some of the techniques used in corpus-related studies (e.g., Coh-Metrix [26]) can be resorted to, hopefully, more linguistic features can be extracted so that a more justifiable model can be constructed for AES.

2) NEURAL NETWORK APPROACHES

Recently, there has been a turning point in AES tasks following the advent of the neural network approach. The neural network approaches do not need to extract features manually and can automatically learn semantic features. Scholars have diverted their attention to setting up AES models without handcrafted features. Taghipour and Ng [10] proposed a neural network approach based on long and short term memory (LSTM) for AES tasks and achieved significant results compared to the previous studies. The model considered the word sequence in the raw text as input and used the convolutional layer to extract features. Dong and Zhang [27] used a convolutional neural network (CNN) model to automatically learn syntactic and semantic features without external preprocessing. Subsequently, Dong *et al.* [11] used concentration at the word and sentence levels and demonstrated how the attention mechanism could improve the accuracy of AES.

Although neural network approaches have achieved more promising results than traditional approaches, they must be trained on a broad collection of training data.

C. TRANSFER LEARNING

In the field of NLP, we often come across tasks that suffer from data deficits and poor generalization ability. Transfer learning can enhance the performance by exploiting

TABLE 1. Detailed information of representative AES models.

AES Models	Datasets	Rating Dimensions	Approaches	Evaluation Metrics
Phandi *et al.* [28]	ASAP	Holistic	Handcrafted Features + Correlated Bayesian Linear Ridge Regression Model	QWK
Chen *et al.* [29]	ASAP	Holistic	CNN-Based Model	QWK
Dimitrios *et al.* [9]	ASAP	Holistic	LSTM-Based Model	QWK
Uto *et al.* [30]	ASAP	Holistic	Hybrid Method: Feature-Engineering Approach + Neural Network Approach	QWK
Our Proposed Model	ASAP + CELA	Holistic + Analytic	Hierarchical + Transformer-Based Model + Multi-Task Learning	QWK

pre-trained language models and applying the knowledge acquired to similar or related tasks.

For AES tasks, Phandi *et al.* [28] used the Correlated Bayesian Linear Ridge Regression to explore essay scoring on different topics. Cummins *et al.* [31] developed a constrained preference learning approach that can jointly performed AES tasks from different topics and rating scales.

As a specific kind of transfer learning, BERT achieves fantastic results in a wide range of challenging tasks. Beltagy *et al.* [32] released SciBERT for scientific texts in the scientific domain. Besides, BioBERT [33] and BioELMo [34] were pre-trained and applied to develop the most influential biomedical text processing models. Although BERT has performed remarkable results in numerous challenging NLP tasks, it is still pre-trained on a large corpus, and in the process of AES, the whole model architecture needs to be modified for the specific task, in our case, the score prediction.

D. MULTI-TASK LEARNING

In contrast to single-task learning, multi-task learning (MTL) involves the simultaneous learning of several interrelated tasks, which can promote performance through the underlying shared representation and enhance the generalization performance of the target domain [35], [36]. MTL has been applied to solve various challenges in the NLP field, such as part-of-speech tagging and syntactic component division [37]–[40]. Collobert and Weston [41] explored six standardized NLP tasks to enhance generalization performance through MTL. Changpinyo *et al.* [42] investigated the application of MTL in sequence tagging tasks.

For AES tasks, Cummins and Rei [43] developed a multi-task neural network approach that can jointly optimize for the tasks of grammatical error detection and AES, showing that the accuracy of the predicted score can be greatly improved with the help of MTL. The advent of the MTL approach for AES has significantly impacted the simultaneous learning of multi-dimensional essay scoring tasks. In our research, MTL involves multi-dimensional essay scoring tasks from different perspectives concerning each dataset. Multi-dimensional essay scoring refers to multi-topic scoring tasks for the ASAP dataset as we explore eight different topics jointly. For the CELA dataset, multi-dimensional essay scoring refers to multi-rating dimensional tasks as we simultaneously investigate holistic and analytic rating dimensions.

Based on the overview of studies above, a summary of representative models is provided for a better understanding of the contributions of our approach. Table 1 compares the most representative AES models from different aspects including the datasets, the rating dimensions, the approaches and the evaluation indicators. Although neural network approaches have been applied to AES tasks, other aspects of writing quality, such as different analytic dimensions providing detailed diagnostic information, are ignored. Besides, no study has utilized the pre-trained language model to improve multi-dimension essay scoring tasks with small training datasets. The motivation of our study is to develop a novel approach using multi-task learning with fine-tuning BERT for multi-topic scoring tasks on the ASAP dataset and multi-rating dimensional tasks on the self-collected CELA dataset.

III. DATASETS

This section gives valuable information about the widely used ASAP dataset and the self-collected CELA dataset. Details about the inter-rater reliability of the self-collected CELA dataset are provided to validate the consistency of human raters.

A. THE ASAP DATASET

The ASAP dataset[1] is sponsored to explore affordable and effective ways for AES tasks. The ASAP dataset contains essays written by American students from grades 7 to 10. There are eight sets of essays from different topics and genres. Topics 1 and 2 are argumentative essays requiring writers to state their opinions on a specific topic. Topics 3 to 6 are response essays where writers are expected to read an extract and respond according to the material. Topic 7 and 8 are narrative essays where writers tell a story based on a particular situation. The scoring range varies from each topic. At least two human raters grade all essays. The average length of each topic is different, ranging from 150 to 650 words. Table 2 shows the statistical details of each topic. In our experiment, we compare the multi-topic essay scoring approach on the ASAP dataset with the single-topic essay scoring approach, which helps understand how the generalization ability and performance of the AES model could be improved.

B. THE CELA DATASET

To address the need to provide valuable information about writing quality, we collected 144 argumentative essays from undergraduates of non-English majors in China at multiple rating dimensions. Participants were required to write a

[1] https://www.kaggle.com/c/asap-aes

TABLE 2. Details of the ASAP dataset. There are 12978 essays with different genres including argumentation, response, and narration. The maximum length of essays in this dataset is 983. Topics 1 and 2 are argumentative essays that require writers to express their ideas about the influence of computers and censorship in libraries. Topics 3 to 6 are response essays that ask writers to read extracts from different topics and respond to each topic according to the requirements. Topics 7 and 8 are narrative essays that require writers to write stories about patients and laughter.

Topics	1_Computers	2_Censorship	3_Cyclist	4_Hibiscus	5_Mood	6_Dirigibles	7_Patience	8_Laughter
Number of Essays	1,783	1,800	1,726	1,772	1,805	1,800	1,569	723
Avg Length	350	350	150	150	150	150	250	650
Max Length	911	118	395	383	452	489	659	983
Min Score	2	1	0	0	0	0	0	0
Max Score	12	6	3	3	4	4	30	60
Unique Score	11	6	4	4	5	5	23	34

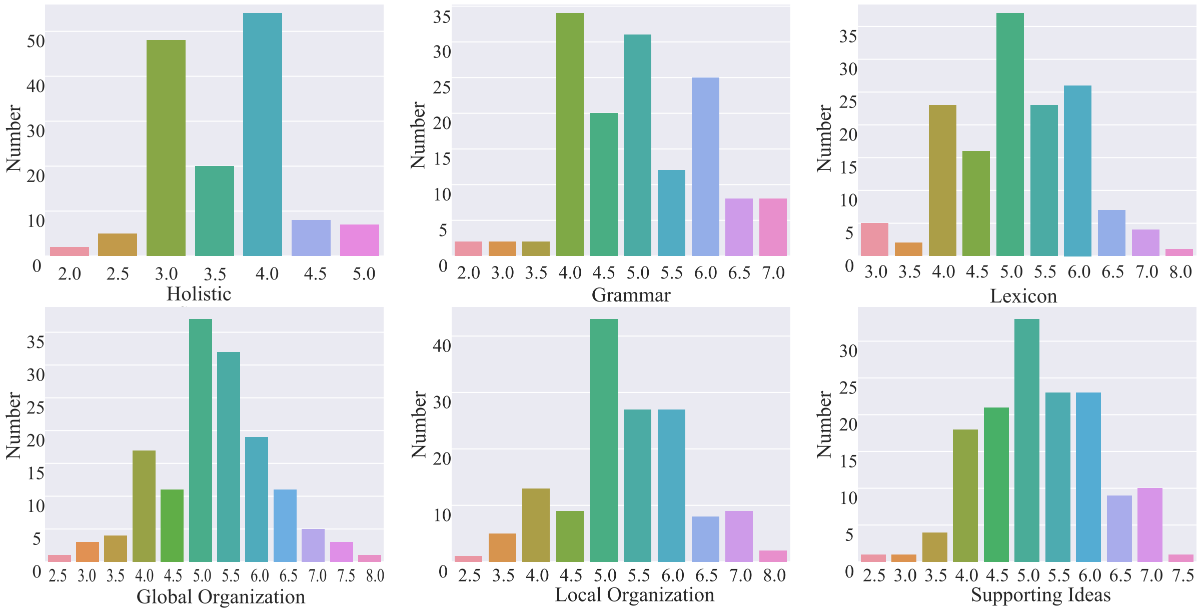


FIGURE 1. Distribution of both holistic and analytic scores of our self-collected Chinese EFL learners' argumentation (CELA) dataset.

TABLE 3. Statistical estimates of holistic and analytic scores of our self-collected Chinese EFL learners' argumentation (CELA) dataset.

Rating Dimensions	Mean	Min	Max	Standard Deviation
Holistic Score	3.59	2.00	5.00	0.63
Grammar	5.01	2.00	7.00	1.01
Lexicon	5.10	3.00	8.00	0.92
Global Organization	5.24	2.50	8.00	0.98
Local Organization	5.34	2.50	8.00	0.92
Supporting Ideas	5.21	2.50	7.50	0.95

300-word essay in response to the following prompt: "*Should a government be allowed to limit the number of children a family can have?*". The holistic and analytic writing rubrics for the CELA dataset can be seen in Tables 12 and 13. Two expert raters were required to score essays regarding the holistic and five analytic rating dimensions, such as *grammar*, *lexicon*, *global organization*, *local organization* and *supporting ideas*. Each rater had to give six scores according to both holistic and analytic scales. The final score of each essay was the average score of the two raters. Figure 1 indicates that the distributions of the holistic and analytic rating dimensions tend to be a little bit above the medium level. Table 3 presents the actual statistical estimates of the holistic and analytical scores of the CELA dataset.[2]

The correlation coefficient and QWK were calculated to evaluate the consistency between human raters' scores. The inter-rater reliability of the CELA dataset was quite acceptable (See Table 4). The correlations between the two raters across different scales were significant ($p < 0.001$). The high agreement between these two raters laid a solid foundation for score prediction in the modeling process.

IV. METHODS

Figure 2 demonstrates the description of our proposed approach, which contains three modules: the input representation module, the weight sharing module, and the output module. The input representation module is responsible for generating vector representations of the essay. The weight sharing module includes BERT and the attention pooling layer, sharing parameters for training different rating dimensions on the CELA dataset. Finally, the output module makes

[2] The CELA dataset is available at https://github.com/gzutxy/CELA.

TABLE 4. Inter-rater reliability between raters of our self-collected Chinese EFL learners' argumentation (CELA) dataset (*** $p < 0.001$).

Rating Dimensions	Holistic	Analytic				
		Grammar	Lexicon	Global Organization	Local Organization	Supporting Ideas
r	0.75***	0.87***	0.81***	0.82***	0.77***	0.75***
QWK	0.75	0.87	0.80	0.79	0.72	0.74

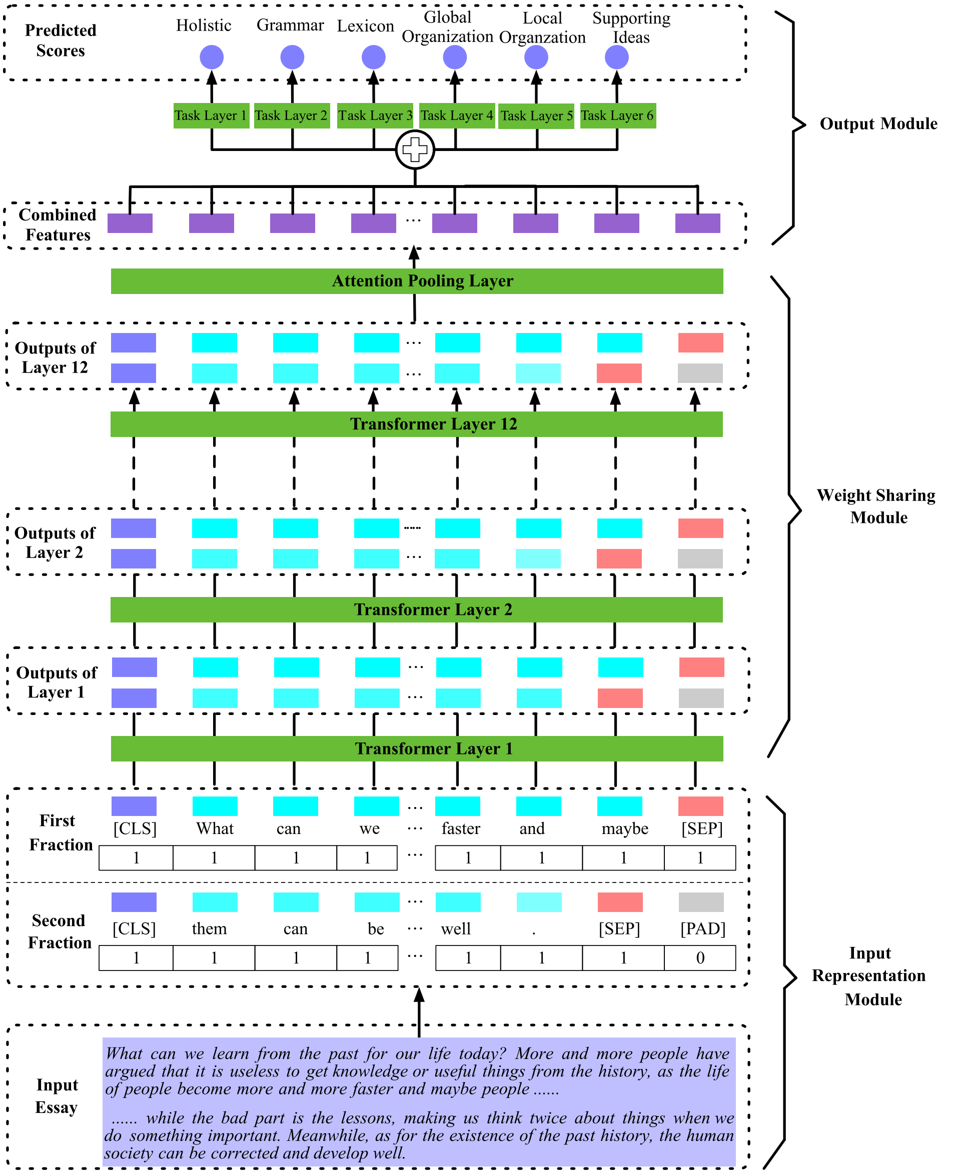

Predicted Scores
Holistic
Grammar
Lexicon
Global Organization
Local Organzation
Supporting Ideas
Task Layer 1
Task Layer 2
Task Layer 3
Task Layer 4
Task Layer 5
Task Layer 6
Output Module
Combined Features
Attention Pooling Layer
Outputs of Layer 12
Transformer Layer 12
Weight Sharing Module
Outputs of Layer 2
Transformer Layer 2
Outputs of Layer 1
Transformer Layer 1
First Fraction
[CLS] What can we ⋯ faster and maybe [SEP]
1 1 1 1 ⋯ 1 1 1 1
Second Fraction
[CLS] them can be ⋯ well . [SEP] [PAD]
1 1 1 1 ⋯ 1 1 1 0
Input Representation Module
Input Essay
What can we learn from the past for our life today? More and more people have argued that it is useless to get knowledge or useful things from the history, as the life of people become more and more faster and maybe people ......
...... while the bad part is the lessons, making us think twice about things when we do something important. Meanwhile, as for the existence of the past history, the human society can be corrected and develop well.

FIGURE 2. A hierarchical BERT-based transfer learning approach for multi-rating dimensional tasks on the Chinese EFL learners' argumentation (CELA) dataset. The model contains three modules: the input representation module, the weight sharing module, and the output module. The input representation module is responsible for generating vector representations of the essay. The weight sharing module includes BERT and the attention pooling layer, sharing parameters for training different rating dimensions on the CELA dataset. Finally, the output module makes predictions for different rating dimensions. For the multi-topic essay scoring tasks on the ASAP dataset, the weight sharing module shares parameters for training different topics and the output module makes predictions for different topics.

predictions for different rating dimensions. For multi-topic essay scoring tasks, the weight sharing module shares parameters for training different topics and the output module makes predictions for different topics.

A. INPUT REPRESENTATION MODULE

As the maximum length is set during the pre-training process, the BERT model can only accept 512 tokens and cannot handle long texts. However, the input length of the AES task is usually longer than the maximum length. When applying BERT to long text tasks, a common approach is to truncate the input to the maximum length. It will reduce performance since the model cannot capture the entire essay's long dependencies and global information.

1) HIERARCHICAL METHOD

To process the long texts in the ASAP dataset, we propose a hierarchical method that divides the input text into $k = L/510^3$ fractions and feeds these fractions into the BERT model. $X = \{X^1, X^2, \cdots X^k\}$ refers to dividing the input text into k fractions, then converting the word sequences of different fractions into embedding vectors. In the subsection IV-B, we will describe how to combine representations of all fractions by attention pooling.

B. WEIGHT SHARING MODULE

BERT includes 12 encoders with 12 bidirectional self-attention heads, and 768 hidden units. In the weight sharing part, 12 encoders with a self-attention mechanism are applied to the embedding vectors of different fractions in the subsection IV-A to collect text information. The representation of each fraction H^i is the hidden state of the final layer of input text X^i, and H^i is the concatenation of $(h_0, h_1, \cdots h_n)$ for each token.

$$H^i = [h_0, h_1, \cdots h_n] \tag{1}$$

1) ATTENTION POOLING

After obtaining the representation H^i of different input fractions of the essay processed by BERT, the final text representation H is learned through the attention pooling layer [44]. The attention pooling layer is defined in the equations 2 and 3, where W_m is weight matrix, w_v is weight vector, b is bias vector.

$$a_i = \frac{e^{w_v \cdot \tanh(W_m \cdot H_i + b)}}{\sum e^{w_v \cdot \tanh(W_m \cdot H_j + b)}} \tag{2}$$

$$H = \sum a_i \cdot H^i \tag{3}$$

C. OUTPUT MODULE

On top of the shared BERT layer, we use a fully connected layer to score each task. During the training process, the BERT model and task-specific layers are fine-tuned through the multi-task objective function.

As shown in the equation 3, H is the semantic representation of the input essay X. The predicted score of each essay X is calculated by logistic regression with sigmoid function:

$$p(c|H) = sigmoid(W_i H), \tag{4}$$

where W_i is the task-specific parameter for different rating dimensions i. We jointly fine-tune all the BERT parameters and W by maximizing the log probability of the human raters' scores.

Algorithm 1: Training Procedures of Our Proposed Approach

```
Initialize model parameters $\theta$
    foreach layer in BERT do
        Copy layer parameters in BERT;
    end
    foreach layer in task-specific layers do
        Initialize parameters randomly;
    end
end
Create mini-batches $D$ by merging different rating
dimensions in the dataset; while $epoch < max\_epoch$ do
    Randomly Shuffle mini-batches $D$;
    $epoch = epoch + 1$;
    while $batch_i$ in mini-batches $D$ do
        Calculate Cross-Entropy loss of $batch_i$;
        Calculate Slope: $\nabla(\theta)$
        Update model parameters: $\theta = \theta - \eta\nabla(\theta)$
    end
end
```

TABLE 5. Hyper-parameters of our proposed approach.

Hyper-Parameters	Values
Activation Function of Task Layers	ReLU [45]
Dropout Rate	0.1
Base Learning Rate	$2e^{-5}$
Warm-Up Proportion	0.1
Optimizer	Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$
Batch Size	16
Training Epochs	5

D. MULTI-TASK TRAINING PROCEDURES

The training procedures of the proposed model consist of two stages: pre-training BERT and multi-task learning. Algorithm 1 shows the detailed procedures of our proposed approach.

Firstly, the pre-trained BERT model is employed to initialize the parameters of the shared layer and randomly initialize the parameters of specific task layers. Then, our proposed models are trained by merging mini-batch data of different scoring dimensions in the dataset. Finally, cross-entropy loss (equation 5) is used to adjust the model weights, where t_j is the true label of score class j, p_j is the probability of score class j in the equation 4. Similar to the previous study which trains the model for multiple natural language understanding tasks [46], mini-batches stochastic gradient descent is trained to update the parameters when training our proposed models.

$$L(\theta) = -\sum_{j=1}^{n} t_j log(p_j) \tag{5}$$

E. EXPERIMENTAL SETUP

1) EVALUATION METRIC

Quadratic Weighted Kappa (QWK) [47] is used to quantify the consistency between raters' scores and predicted scores.

[3] The sequence always contains a specific classification embedding [CLS] and another unique token for separating segments [SEP].

The ASAP competition treats QWK as the golden standard for evaluation. Other experiments using the ASAP dataset (e.g. [11], [48], [49]) also adopt this evaluation as an indicator of performance. Since we use the ASAP dataset for evaluation in our experiments, we thus regard QWK as the evaluation standard. This indicator usually varies from 0 (no consistency between raters) to 1 (perfect consistency between raters). For details about the QWK formula, please refer to Zhao *et al*. [48].

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j}O_{i,j}}{\sum_{i,j} W_{i,j}E_{i,j}} \tag{6}$$

2) IMPLEMENTATION DETAILS

We used 5-fold cross-validation to evaluate our model, where 60% of the data was used for training, 20% of the data was used for validation, and 20% for testing. The optimal model was selected according to the results of the validation data. The detailed information about hyper-parameters used in our proposed approach can be seen in Table 5. We set ReLU [45] as the activation function of the task layer. The dropout rate [50] of all the task-specific layers was to 0.1. The base learning rate was set to $2e^{-5}$. We set the warm-up proportion to 0.1 and used the Adam optimizer. The PyTorch framework [51] was used to train the model on the NVIDIA 3090 GPU, with a batch size of 16. We set the maximum training epochs to 5. The settings mentioned above were applied for the experiments on the multi-topic scoring and the multi-dimensional scoring tasks.

V. RESULTS

This section demonstrates the experimental results of the multi-topic scoring and multi-rating dimensional tasks. Comparisons with the baselines and detailed analysis are presented to validate the efficiency of our proposed approach.

A. MULTI-TOPIC SCORING TASKS ON THE ASAP DATASET

1) COMPARISONS WITH THE BASELINES

Table 6 (rows 2-5) shows the QWK scores of the baselines on topics of the ASAP dataset. MN [48] predicts the score by calculating the correlation between the non-graded response and each selected response in memory. LSTM CNN-att [11] stands for constructing a hierarchical model to represent the sentences and applying the attention mechanism to decide the weights automatically. TSLF-ALL [49] denotes a two-stage learning framework that combines the strengths of both feature-engineered and end-to-end AES methods. HISK + BOSWE and v-SVR [22] presents an approach that includes both low-level types of features and high-level semantic feature representation.

BERT-finetune here means that we fine-tune BERT for AES tasks with each topic on the ASAP dataset, while *BERT-MTL-finetune* represents that all the tasks of AES are exploited simultaneously with eight topics.

It can be seen from Table 6 that the QWK scores of our proposed approach are much higher than the baselines. Our model *BERT-MTL-finetune* surpasses the strong baseline HISK + BOSWE and v-SVR [22] in terms of all topics and improves the average QWK score by 4.5%. We conduct paired *t*-test to explore whether the QWK scores of our model are significantly higher than the strong baseline. Results show that *BERT-MTL-finetune* improved performance at the 5% significance level ($p = 0.033$) for HISK + BOSWE and v-SVR [22].

We also observe that *BERT-MTL-finetune* outperforms six-eighths of topics and leads the average performance compared with *BERT-finetune*. As shown in Table 6, *BERT-finetune* outperforms *BERT-MTL-finetune* in topics 1 and 5. One interpretation for this phenomenon is that the number of essays on each topic is sufficient enough. Therefore, *BERT-finetune* can achieve good results without learning the AES tasks jointly and focus on training the data for the specific topic. Another explanation lies in that *BERT-MTL-finetune* trains all topics jointly in one model and different topics will contribute to the modeling effect. For example, for topic 8 with the least amount of data, the QWK score of *BERT-MTL-finetune* increases the most; while for topic 5 with the largest amount of data, the QWK score of *BERT-MTL-finetune* has a slight decrease. However, *BERT-MTL-finetune* improves the performance from about 0.815 to 0.830 in the average QWK scores. The results imply that when the training data is large enough, *BERT-finetune* achieves the best performance as it focuses on training the specific topic only. Alternatively, when the training data is small, *BERT-MTL-finetune* performs better as it trains all topics jointly and provides shared representation for different topics in modeling.

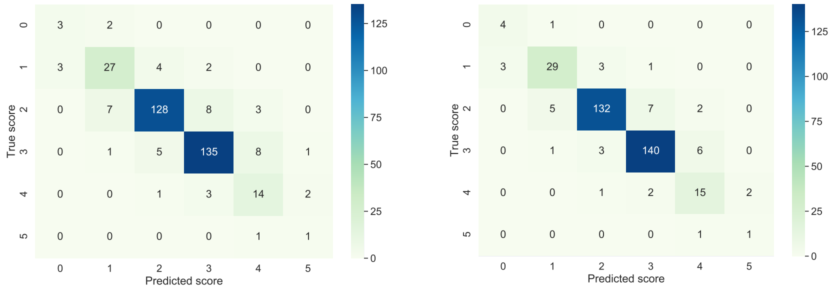
We plot the confusion matrices between true and predicted scores of *BERT-finetune* and *BERT-MTL-finetune* for the ASAP dataset topic 2 (see Figure 3). The confusion matrices indicate that the predicted scores of *BERT-MTL-finetune* tend to be closer to true scores, which corroborate with our above findings.

To better understand the reasons for accurate and inaccurate predictions, we analyze two sample essays from topic 2 in the ASAP dataset. The sample essay of inaccurate prediction (Essay ID 3705, true score 2, predicted score 4) uses complex sentence structures but with various errors in morphosyntax (e.g., spelling and verb form). Besides, it is pretty hard for readers to understand its arguments and ideas. In contrast, the sample text of accurate prediction (Essay ID 3291, true score 4, predicted score 4) uses a variety of complex syntactic structures with errors in spelling that will not cause confusion. Moreover, it is developed with several examples to support the topic. An explanation for inaccurate prediction is that *BERT-MTL-finetune* tends to assign the mean score to essays that are of low or high quality because predicting the mean score will minimize the penalty of the loss and improve the accuracy of the model.

The above findings prove that *BERT-MTL-finetune* improves the accuracy of AES tasks by simultaneously fine-tuning BERT with all the topics on the ASAP dataset. More

TABLE 6. Comparison of QWK scores on the ASAP dataset with our proposed approach and the baselines.

Topics	1_Computers	2_Censorship	3_Cyclist	4_Hibiscus	5_Mood	6_Dirigibles	7_Patience	8_Laughter	Avg
MN [48]	0.830	0.720	0.720	0.820	0.830	0.830	0.790	0.680	0.780
LSTM CNN-att [11]	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
TSLF-ALL [49]	0.852	0.736	0.731	0.801	0.823	0.792	0.762	0.684	0.773
HISK + BOSWE and v-SVR [22]	0.845	0.792	0.684	0.829	0.833	0.830	0.804	0.729	0.785
BERT-finetune	**0.882**	0.786	0.802	0.839	**0.848**	0.826	0.815	0.721	0.815
BERT-MTL-finetune	0.853	**0.836**	**0.828**	**0.843**	0.840	**0.832**	**0.826**	**0.784**	**0.830**



(a) BERT-finetune

(b) BERT-MTL-finetune

FIGURE 3. The confusion matrices between true and predicted scores of *BERT-finetune* and *BERT-MTL-finetune* for the ASAP dataset Topic 2 Censorship.

TABLE 7. Results of different methods for dealing with long texts in the ASAP dataset.

Models	Methods	Avg QWK
BERT-finetune	Head-Only	0.801
	Tail-Only	0.798
	Head + Tail	0.794
	Hierarchical + Average Pooling	0.812
	Hierarchical + Attention Pooling	**0.815**
BERT-MTL-finetune	Head-Only	0.813
	Tail-Only	0.808
	Head + Tail	0.814
	Hierarchical + Average Pooling	0.817
	Hierarchical + Attention Pooling	**0.830**

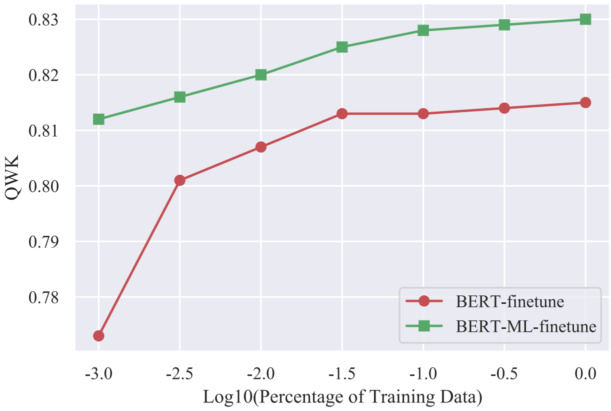


FIGURE 4. Results of the QWK scores on the ASAP dataset with different proportions of training samples.

importantly, the results validate that *BERT-MTL-finetune* effectively increases the sample size of the training model and promotes performance through the underlying shared representation, thereby improving the generalization ability and performance of the AES model.

2) DEALING WITH LONG TEXTS

As shown in Table 2, the maximum sequence length of the ASAP dataset is above 512. Therefore, the primary problem of applying BERT to AES is to deal with the long texts in the ASAP dataset. We use the following ways to deal with long texts: Head-Only; Tail-Only; Head + Tail [52]; and the hierarchical method mentioned in subsection IV-A.The first three ways are the conventional truncation methods, which keep the key and essential information of each essay. Head-Only means that only the first 510 tokens are reserved. Tail-Only means that the last 510 tokens are reserved. Head + Tail means that the first 128 and the last 382 tokens are reserved. In the ASAP dataset, the maximum length of the essay is 983. Therefore, we set the k value to 2 in this experiment and divide the essay into two input fractions.

Table 7 indicates that the Hierarchical + Attention Pooling approach outperforms the traditional truncation approach on the average QWK value, which shows great power in dealing with long texts. The findings are understandable because the traditional truncation approach could not capture the features of the whole essay and tends to lead to missing data problems in the modeling process. Alternatively, the hierarchical method provides good integrity support and can capture the combined representation of the entire essay. Compared with the Hierarchical + Average Pooling approach, the Hierarchical + Attention Pooling approach effectively enhances performance since the model can capture the combined representation of the entire essay. Potential explanations are that the attention pooling can capture the critical information in

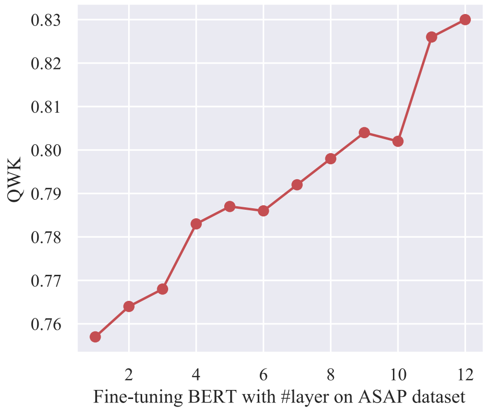


FIGURE 5. Results of the QWK scores of fine-tuning BERT with different layers.

different fractions and better highlight the impact of crucial parts of the essay. Unlike average pooling, with each fraction being evaluated equally, which contradicts to the human raters' evaluation process. We obtain the best performance by automatically learning the relevance of different fractions of each essay through the attention pooling layer. It demonstrates how the attention mechanism aids in locating crucial elements that contribute to judging the writing quality of each essay.

3) EXPERIMENTS ON SMALL DATASETS

The key strength of the pre-trained language model is that we can train models for particular tasks with small datasets. We conduct experiments using 0.1% to 100% of the training data in the ASAP dataset. We evaluate *BERT-finetune* and *BERT-MTL-finetune* on different proportions of training samples.

Figure 4 demonstrates that *BERT-MTL-finetune* brings a significant improvement to small datasets. After carefully examining the data, we speculate that even if just 0.1% of the training data is utilized when employing the pre-trained language model, the QWK score of *BERT-finetune* is 0.77, which is superior to the baseline LSTM CNN-att [11]. More specifically, *BERT-MTL-finetune* outperforms all the baselines in Table 1 even if just 0.1% of the training data is utilized. This demonstrates that fine-tuning the pre-trained language model in score prediction offers significant advantages, such as not requiring a vast quantity of labeled data and ensuring reliable model performance.

4) FEATURES FROM DIFFERENT LAYERS

Since each layer of the BERT model contains distinct characteristics of the input text, we explore the validity of features from each layer of the BERT model. Figure 5 outlines the performance of fine-tuning BERT with each layer on the ASAP dataset. The top-level features of BERT obtain the best performance.

B. MULTI-RATING DIMENSIONAL TASKS ON THE CELA DATASET

1) COMPARISONS WITH THE BASELINES

To build the baselines, we first use BERT to extract features from the original data and then use different regression approaches (rows 2-5) to score the features. The parameter settings of the above methods are shown in Table 8. *BERT-finetune* in Table 9 means that we fine-tune BERT for AES tasks at holistic and analytic dimensions separately, while *BERT-MTL-finetune* means that all the tasks of AES are exploited simultaneously at multiple dimensions.

As indicated in Table 9, our model outperforms the baselines on all these six rating dimensions. Results indicate that our models (*BERT-finetune* and *BERT-MTL-finetune*) improve the average QWK score by 6.0% and 8.1% respectively, compared with the strong baseline BERT + Neural Networks. Compared with *BERT-finetune*, which fine-tunes BERT for AES tasks at holistic and analytic dimensions separately, our proposed approach *BERT-MTL-finetune* still improves the average QWK score by 2.1%. This proves that using a BERT-based transfer learning approach brings benefits to multi-dimensional scoring through the underlying shared representation of each rating dimension. Paired *t*-test is conducted to explore whether the averaged QWK score of our proposed approach is significantly higher than the strong baseline BERT + Neural Networks. Results show that *BERT-MTL-finetune* improved performance at the 0.1% significance level ($p = 6.93\ e^{-13}$) for the strong baseline BERT + Neural Networks.

Different from the results of the ASAP dataset, *BERT-MTL-finetune* outperforms *BERT-finetune* at all the multi-rating dimensional tasks. An interpretation for the variation of results on the CELA and ASAP datasets is that the number of essays used for training differs. For the ASAP dataset, the training data for *BERT-finetune* is the specific topic, while the training data for *BERT-MTL-finetune* contains all the eight topics. Therefore, when the training data is large enough, *BERT-finetune* achieves the best performance as it focuses on training the specific topic only. Alternatively, when the training data is small, *BERT-MTL-finetune* performs better as it trains all topics jointly and provides shared representation for different topics in modeling. For the CELA dataset, *BERT-finetune* uses single-dimensional scores for training, while *BERT-MTL-finetune* uses multi-dimensional scores for training. The difference lies in the number of training essays of these two approaches remains unchanged, but *BERT-MTL-finetune* has more correlated scoring labels. Therefore, the QWK score of each rating dimension has been improved.

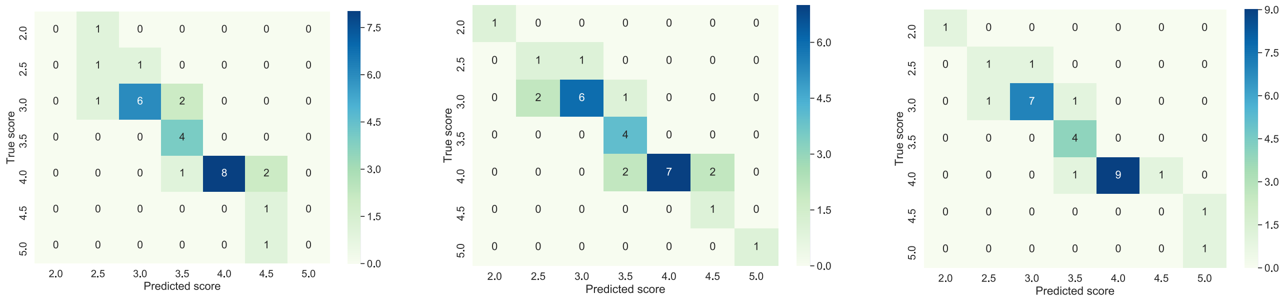
We plot the confusion matrices between true and predicted scores of the baseline BERT + Neural Networks, *BERT-finetune* and *BERT-MTL-finetune* for the CELA dataset at holistic rating dimension. The confusion matrices (see Figure 6) show that The BERT + Neural Networks approach assigns a higher score than the human raters' score compared with our proposed approach. Also, the confusion matrices

TABLE 8. Baseline parameter settings of the self-collected Chinese EFL learners' argumentation (CELA) dataset.

Methods	Parameter Settings
Linear Regression	copy X = 1, fit intercept = 1, normalize = False
Decision Tree	criterion = 'mse', min samples leaf = 1, min samples split = 2
Random Forest	bootstrap = 1, criterion = 'mse', min samples leaf = 1, min samples split = 2, n estimators = 10
Neural Networks	optimizer = Adam(), lr = 0.001, loss fn = L1 Loss(), batchsize = 16

TABLE 9. Comparison of QWK scores on the self-collected Chinese EFL learners' argumentation (CELA) dataset with our proposed approach and the baselines.

Rating Dimensions	Holistic	Analytic				
		Grammar	Lexicon	Global Organization	Local Organzation	Supporting Ideas
BERT + Linear Regression	0.652	0.694	0.624	0.671	0.635	0.610
BERT + Decision Tree	0.676	0.718	0.648	0.695	0.659	0.634
BERT + Random Forest	0.681	0.723	0.653	0.701	0.664	0.639
BERT + Neural Networks	0.702	0.744	0.674	0.721	0.684	0.660
BERT-finetune	0.762	0.804	0.734	0.781	0.745	0.720
BERT-MTL-finetune	**0.783**	**0.825**	**0.755**	**0.802**	**0.766**	**0.741**



(a) BERT + Neural Networks

(b) BERT-finetune

(c) BERT-MTL-finetune

FIGURE 6. Confusion matrices between true and predicted scores of BERT + Neural Networks, ***BERT-finetune*** and ***BERT-MTL-finetune*** for the CELA dataset at holistic rating dimension.

TABLE 10. Accurate and inaccurate predictions of sample essays in the CELA dataset.

Rating Dimensions		Grammar	Lexical	Global Organization	Local Organization	Supporting Ideas	Holistic
Inaccurate Prediction ID 1	True	2	3	2.5	2.5	2.5	2
	Predicted	3	4	2.5	3.5	3	3
Accurate Prediction ID 49	True	6.5	6	7	7	6.5	5
	Predicted	6.5	5.5	7	7	6.5	5

demonstrate that the predicted scores of *BERT-MTL-finetune* tend to be closer to true scores than the baselines, which corroborate with our above findings.

For a comprehensive understanding of the reasons for accurate and inaccurate predictions, we discuss sample essays in the CELA dataset (see Table 10). The sample essay of inaccurate prediction uses simple sentence structures with frequent errors in morphosyntax (e.g., spelling, punctuation, and consistency). Besides, it is an off-topic essay with little developed examples to support the topic. In contrast, the sample text of accurate prediction uses a variety of complex syntactic structures with rate errors in lexicon or grammar. Moreover, it is well-organized and developed with successfully developed examples to support the main idea. Generally speaking, *BERT-MTL-finetune* achieves fantastic results in multi-rating dimensional essay scoring tasks. Even for the sample of inaccurate predictions, the differences between the predicted and true score are about 0.5 to 1. An explanation for inaccurate prediction is that the off-topic perspective is not explored.

TABLE 11. Results of hierarchical v.s. non-hierarchical approach for dealing with short essays in the CELA dataset.

Models	Methods	Avg QWK
BERT-finetune	Non-Hierarchical	0.759
	Hierarchical + Average Pooling	0.612
	Hierarchical + Attention Pooling	**0.762**
BERT-MTL-finetune	Non-Hierarchical	0.782
	Hierarchical + Average Pooling	0.681
	Hierarchical + Attention Pooling	**0.783**

TABLE 12. The writing rubric for the CELA dataset at holistic rating dimension.

Score	Description
5	Is well organized and developed with clear and appropriate explanations, examples, and/or detailed information; complex syntactic diversity and appropriate word selection, although it may occasionally be wrong.
4	Is roughly well organized and developed with appropriate and adequate explanations, examples, and/or detailed information; shows facility in the language uses; demonstrates the diversity of syntax and vocabulary, although there are minor errors in the form, but will not interfere with the meaning.
3	Uses some developed explanations to support or illustrate an idea; is adequately organized and developed; demonstrates sufficient but probably inconsistent syntactic and word usages.
2	Insufficient supporting ideas; Inappropriate or unrelated examples, explanations, and/or detailed information; obviously inappropriate word usages.
1	Severe confusion or underdevelopment; severe and persistent errors in sentence structures or word usages.

TABLE 13. The writing rubric for the CELA dataset at analytic rating dimensions.

Rating Scales	Score	Level	Description
Grammar	1-2	POOR	Uses simple sentence structures; but there are still serious and frequent errors in mrophosyntax (e.g. tense, consistency).
	3-4	FAIR	Uses a small range of syntactic patterns with limited uses of complex and subordinate clauses; but there are only a few errors in morphosyntax.
	5-6	GOOD	Uses a variety of complex syntactic structures and longer sentences, with minor/occasional errors that do not interfere with meaning.
	7-8	EXCELLENT	Uses various complex constructions effectively and accurately, although it may have rare errors in lexicon or grammar.
Lexicon	1-2	POOR	With an extremely limited range of vocabulary and obviously inappropriate word usages.
	3-4	FAIR	Satisfactory usage of basic vocabulary, but may have limited awareness word formation, which leads to ambiguous meaning.
	5-6	GOOD	Masters a wide range of vocabulary, but has minor errors in word selection, word formation and/or spelling.
	7-8	EXCELLENT	With a wide range of vocabulary with very sophisticated and effective use; produces minor errors.
Global organization	1-2	POOR	Severely disorganized or underdeveloped throughout the whole essay, and difficult for readers to understand the logic of supporting ideas.
	3-4	FAIR	Satisfactory organization of the whole essay, but the order of the paragraphs may be inadequate and occasionally be problematic or difficult to understand.
	5-6	GOOD	Well organized and developed for the whole essay (both sections and paragraphs).
	7-8	EXCELLENT	Complex organization of sections and paragraphs, with clear and adequate details.
Local organization	1-2	POOR	Relies primarily on very limited cohesive devices and basic conjunctions in sentences within each paragraph.
	3-4	FAIR	Provides information with sentences within each paragraph; but there may be insufficient, inaccurate or excessive use of cohesive devices, which may not affect the understanding of ideas.
	5-6	GOOD	Appropriate uses of a wide range of cohesive devices and logical ideas that present a clear central topic within sentences in each paragraph.
	7-8	EXCELLENT	Manages all aspects of cohesion well; sequence information and ideas logically; uses sophisticated cohesive devices to make readers clearly understand sequencing of ideas.
Supporting ideas	1-2	POOR	Uses little examples, explanations and facts to support or illustrate the pros/cons of one'position.
	3-4	FAIR	Insufficient examples and facts to express a position, may have difficulties in choosing relevant examples/facts to present ideas; insufficiently illustrating those examples/facts.
	5-6	GOOD	Successfully developed explanations and details to support one's position, appropriate uses of examples and facts.
	7-8	EXCELLENT	Clear and adquate explanations, examples and/or detailed information; well organized and convincing ideas.

2) HIERARCHICAL V.S. NON-HIERARCHICAL APPROACH FOR DEALING WITH SHORT ESSAYS

For the CELA dataset, although the length of all essays is less than 512, we still apply the Hierarchical approach in the training process to improve the generalization ability of our approach. In doing so, our proposed model could deal with long texts in the context of writing tests. As in the ASAP dataset, we set the *k* value to 2. The experimental results are shown in Table 11. The results indicate that the QWK score of the Hierarchical + Attention approach is slightly higher than the Non-Hierarchical approach, because the input of the second fraction is filled with 0 in the Hierarchical + Attention approach for essays less than 510 tokens. Besides, the Hierarchical + Attention approach only involves one more non-linear attention layer than the Non-Hierarchical approach. Therefore, the Hierarchical + Attention Pooling approach improves performance by a little bit. The Hierarchical + Average Pooling approach is inferior because the mean layer averages the features of these two fractions, which affects the final scoring effect with the features of the second part being zero.

VI. CONCLUSION

In this study, we propose a BERT-based transfer learning approach for multi-dimensional essay scoring. To the best of our knowledge, this is the first study to fine-tune BERT for multi-topic essay scoring tasks on the ASAP dataset and multi-rating dimensional essay scoring tasks on the self-collected Chinese EFL learners' argumentation (CELA) dataset.

For multi-topic scoring tasks on the widely used ASAP dataset, the experimental results show that our proposed approach *BERT-MTL-finetune* significantly improves the performance by the underlying shared representation of each topic. We also deal with long essays by proposing a hierarchical method and using the attention mechanism. The results demonstrate that the Hierarchical + Attention approach effectively enhances performance since the model can capture the combined representation of each topic.

For multi-rating dimensional tasks on the CELA dataset, the experimental results indicate that our proposed approach *BERT-MTL-finetune* benefits multi-dimensional scoring through the underlying shared representation of each rating dimension. Combing both holistic and analytic rating dimensions increases the validity and reliability of score prediction.

In future work, we will further explore multi-dimensional AES tasks from the following four perspectives: 1) Other pre-trained models XLNet [53], GPT [54], and GPT-2 [55] could be applied in multi-dimensional AES tasks. 2) The potential effects of the extent of different rating dimensions accounting for the overall writing quality could be explored. 3) The traditional approaches based on hand-crafted features and the neural approaches could be incorporated to provide diagnostic feedback for learners to improve their writing quality. 4) Different genres and topics can be included to enrich the diversity of the self-collected CELA dataset.

APPENDIX A WRITING RUBRICS AT THE HOLISTIC AND ANALYTIC SCALES

A. WRITING RUBRICS AT THE HOLISTIC SCALE

See Table 12.

B. WRITING RUBRICS AT THE ANALYTIC SCALES

See Table 13.

REFERENCES

[1] E. Amorim, M. Cançado, and A. Veloso, "Automated essay scoring in the presence of biased ratings," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 229–237. [Online]. Available: https://aclanthology.org/N18-1021

[2] S. M. Eid and N. M. Wanas, "Automated essay scoring linguistic feature: Comparative study," in *Proc. Int. Conf Adv. Control Circuits Syst. (ACCS) Syst. Int. Conf New Paradigms Electron. Inf. Technol. (PEIT)*, Nov. 2017, pp. 212–217.

[3] J. Shin and M. J. Gierl, "More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms," *Lang. Test.*, vol. 38, no. 2, pp. 247–272, Apr. 2021, doi: 10.1177/0265532220937830.

[4] J. C. Machicao, "Higher education challenge characterization to implement automated essay scoring model for universities with a current traditional learning evaluation system," in *Proc. Int. Conf. Inf. Technol. Syst.*, Feb. 2019, pp. 835–844.

[5] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *J. Comput. Higher Educ.*, early access, pp. 1–20, Jun. 2021, doi: 10.1007/s12528-021-09283-1.

[6] V. S. Kumar and D. Boulanger, "Automated essay scoring and the deep learning black box: How are rubric scores determined?" *Int. J. Artif. Intell. Educ.*, early access, pp. 1–47, Sep. 2020, doi: 10.1007/s40593-020-00211-5.

[7] M. Beseiso and S. Alzahrani, "An empirical analysis of BERT embedding for automated essay scoring," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 204–210, 2020.

[8] R. Gunawansyah, R. Rahayu, Nurwathi, B. Sugiarto, and Gunawan, "Automated essay scoring using natural language processing and text mining method," in *Proc. 14th Int. Conf. Telecommun. Syst., Services, Appl.*, Nov. 2020, pp. 1–4.

[9] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," 2016, *arXiv:1606.04289*. [Online]. Available: http://arxiv.org/abs/1606.04289

[10] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proc. EMNLP*, 2016, pp. 1882–1891.

[11] F. Dong, Y. Zhang, and J. Yang, "Attention-based recurrent convolutional neural network for automatic essay scoring," in *Proc. CoNLL*, 2017, pp. 153–162.

[12] C. Jin, B. He, K. Hui, and L. Sun, "TDNN: A two-stage deep neural network for prompt-independent automated essay scoring," in *Proc. ACL*, 2018, pp. 1088–1097.

[13] Y. Tay, M. C. Phan, L. A. Tuan, and S. Hui, "Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring," in *Proc. AAAI*, 2018, pp. 1–8.

[14] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[16] S. C. Weigle, *Assessing Writing*. Stuttgart, Germany: Ernst Klett Sprachen, 2002.

[17] B. B. Klebanov and M. Flor, "Word association profiles and their use for automated scoring of essays," in *Proc. ACL*, 2013, pp. 1148–1158.

[18] A. Faulkner, "Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure," in *Proc. FLAIRS Conf.*, 2014, pp. 1–6.

[19] B. Beigman Klebanov, M. Flor, and B. Gyawali, "Topicality-based indices for essay scoring," in *Proc. BEA@NAACL-HLT*, 2016, pp. 63–72.

[20] I. Persing, A. Davis, and V. Ng, "Modeling organization in student essays," in *Proc. EMNLP*, 2010, pp. 229–239.

[21] I. Persing and V. Ng, "Modeling argument strength in student essays," in *Proc. ACL*, 2015, pp. 543–552.

[22] M. Cozma, A. M. Butnaru, and R. T. Ionescu, "Automated essay scoring with string kernels and word embeddings," 2018, *arXiv:1804.07954*. [Online]. Available: http://arxiv.org/abs/1804.07954

[23] N. Farra, S. Somasundaran, and J. Burstein, "Scoring persuasive essays using opinions and their targets," in *Proc. BEA@NAACL-HLT*, 2015, pp. 64–74.

[24] H. V. Nguyen and D. Litman, "Argument mining for improving the automated scoring of persuasive essays," in *Proc. AAAI*, 2018, pp. 1–8.

[25] L. M. Rudner and T. Liang, "Automated essay scoring using Bayes' theorem," *J. Technol., Learn. Assessment*, vol. 1, no. 2, pp. 1–22, 2002.

[26] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-Metrix: Analysis of text on cohesion and language," *Behav. Res. Methods, Instrum., Comput.*, vol. 36, no. 2, pp. 193–202, May 2004.

[27] F. Dong and Y. Zhang, "Automatic features for essay scoring—An empirical study," in *Proc. EMNLP*. Austin, TX, USA: Association for Computational Linguistics, Nov. 2016, pp. 1072–1077. [Online]. Available: https://www.aclweb.org/anthology/D16-1115

[28] P. Phandi, K. M. A. Chai, and H. T. Ng, "Flexible domain adaptation for automated essay scoring using correlated linear regression," in *Proc. EMNLP*, 2015, pp. 431–439.

[29] Z. Chen and Y. Zhou, "Research on automatic essay scoring of composition based on CNN and OR," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 13–18.

[30] M. Uto, Y. Xie, and M. Ueno, "Neural automated essay scoring incorporating handcrafted features," in *Proc. COLING*, 2020, pp. 6077–6088.

[31] R. Cummins, M. Zhang, and T. Briscoe, "Constrained multi-task learning for automated essay scoring," in *Proc. ACL*, 2016, pp. 789–799.

[32] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," 2019, *arXiv:1903.10676*. [Online]. Available: http://arxiv.org/abs/1903.10676

[33] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[34] Q. Jin, B. Dhingra, W. Cohen, and X. Lu, "Probing biomedical embeddings from language models," in *Proc. 3rd Workshop Evaluating Vector Space Represent. NLP*, 2019, pp. 82–89.

[35] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.

[36] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, early access, p. 1, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9392366, doi: 10.1109/TKDE.2021.3070203.

[37] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[38] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," *CoRR*, vol. abs/1511.06114, pp. 1–10, Mar. 2016.

[39] H. Guo, R. Pasunuru, and M. Bansal, "Soft layer-specific multi-task summarization with entailment and question generation," in *Proc. ACL*, 2018, pp. 687–697.

[40] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," in *Proc. AAAI*, 2019, pp. 4822–4829.

[41] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. ICML*, 2008, pp. 160–167.

[42] S. Changpinyo, H. Hu, and F. Sha, "Multi-task learning for sequence tagging: An empirical study," 2018, *arXiv:1808.04151*. [Online]. Available: http://arxiv.org/abs/1808.04151

[43] R. Cummins and M. Rei, "Neural multi-task learning in automated assessment," 2018, *arXiv:1801.06830*. [Online]. Available: http://arxiv.org/abs/1801.06830

[44] H. Zhang and D. Litman, "Co-attention based neural network for source-dependent essay scoring," in *Proc. BEA@NAACL-HLT*, 2018, pp. 399–409.

[45] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2018, *arXiv:1803.08375*. [Online]. Available: http://arxiv.org/abs/1803.08375

[46] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. ACL*, 2019, pp. 4487–4496.

[47] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.

[48] S. Zhao, Y. Zhang, X. Xiong, A. Botelho, and N. Heffernan, "A memory-augmented neural model for automated grading," in *Proc. 4th ACM Conf. Learn. Scale*, Apr. 2017, pp. 189–192.

[49] J. Liu, Y. Xu, and Y. Zhu, "Automated essay scoring based on two-stage learning," 2019, *arXiv:1901.07744*. [Online]. Available: http://arxiv.org/abs/1901.07744

[50] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

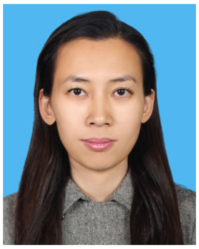
[51] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8026–8037.

[52] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham, Switzerland: Springer, 2019, pp. 194–206.

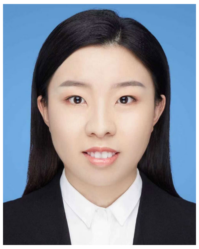
[53] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. NeurIPS*, 2019, pp. 1–11.

[54] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," OpenAI Blog, Tech. Rep., 2018. [Online]. Available: https://openai.com/blog/language-unsupervised/ and https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[55] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, Tech. Rep., 2019. [Online]. Available: https://openai.com/blog/better-language-models/ and https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf



JIN XUE received the Ph.D. degree from Beijing Normal University. She is currently a Professor with the Center for the Advancement of Language Science, University of Science and Technology Beijing, China. Her research interests include second language teaching and learning, language assessment, psycholinguistics, and neurolinguistics.



XIAOYI TANG is currently pursuing the Ph.D. degree with the Center for the Advancement of Language Science, University of Science and Technology Beijing, China. Her research interests include automated essay scoring, natural language processing, language assessment, and computational linguistics.

LIYAN ZHENG received the master's degree from the University of Science and Technology Beijing, China. Her research interests include writing assessment, and language learning and teaching.