Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Performance prediction and importance analysis using Transformer

Akiyoshi SATAKE[a*], Hironobu FUJIYOSHI[b], Takayoshi YAMASHITA[c], Tsubasa HIRAKAWA[d], Atsushi SHIMADA[e]

[a]*MPRG, Chubu University, Japan*
[b]*MPRG, Chubu University, Japan*
[c]*MPRG, Chubu University, Japan*
[c]*MPRG, Chubu University, Japan*
[e]*Kyushu University, Japan*
*satake@mprg.cs.chubu.ac.jp

Abstract: The growth of online education has made it easier to capture learner activity. It is expected that detailed feedback to learners will lead to better performance. For this purpose, it is important to predict the performance of learners. Methods using classical machine learning and RNNs that take time series information into account have been proposed. In this paper, we propose a Transformer-based performance prediction method that aims to improve accuracy and extract important activity. The proposed method achieves more accurate performance prediction than conventional methods. In addition, we found that NEXT, SEARCH_JUMP and LINK_CLICK are important behaviors by analyzing the rationale of the Transformer.

Keywords: Transformer, score prediction, action, RNN

1. Introduction

In recent years, the growing usage of online education systems has made it easier to collect data on the learning activities of students on a large scale. For example, in a Massive Open Online Course (MOOC) course, they collect user clickstream data on the website. By collecting data on such learning activities, it is possible to learning activity analysis, pattern mining, predict grades, identifying at risk students, and support learning.

In the field of Natural Language Processing (NLP), Transformer, a deep learning model using only Attention, was proposed in 2017 and outperformed the accuracy of translation by traditional methods such as Recurrent Neural Network (RNN) [1].

When estimating student grades, classical machine learning and simple deep learning methods such as the multilayer perceptron are still often used. but, such methods cannot handle complex data and may lack important information for estimation. There are also some studies that have used RNNs, which are deep learning models that can take into account time-series information, to estimate student performance [2, 3]. However, when dealing with long-term time series information, RNNs tend to give more importance to the most recent information, making it difficult to capture the important information from the long-term information. If can make highly accurate estimating student grades by considering long-term information, we will be able to provide more detailed support to students in learning.

In this paper, we propose a method for estimating a student's test score from the student's learning activity data using the encoder part of the Transformer, which can take into account more long-term information. We will also discuss the more important information in inferring test scores from the Transformer's Attention.

2. Transformer

2.1 Outline of Transformer

Transformer was published in 2017 in the field of NLP and scored above the state of the art at that time. Many methods have been developed, including BERT [4], as application methods. In addition to NLP, there are methods used in other fields, such as Vision Transformer [5] in image recognition and Conformer [6] in speech recognition.

The Transformer is constructed using only the Attention layer without using RNN or Convolutional Neural Network (CNN). This allows for parallel computation, which speeds up the computation, and at the same time allows for the consideration of long-term time series information.

2.2 Structure of Transformer

The structure of the encoder part of the Transformer is shown in Figure 1. The encoder part is composed of Input Embedding, which compresses the input, Positional Encoding, which adds positional information, Multi-Head Attention, which calculates Attention, Feed Forward, and Add & Norm, which performs residual merging and normalization. Of these, the processing from Multi-Head Attention to Add & Norm after Feed Forward is repeated multiple times.

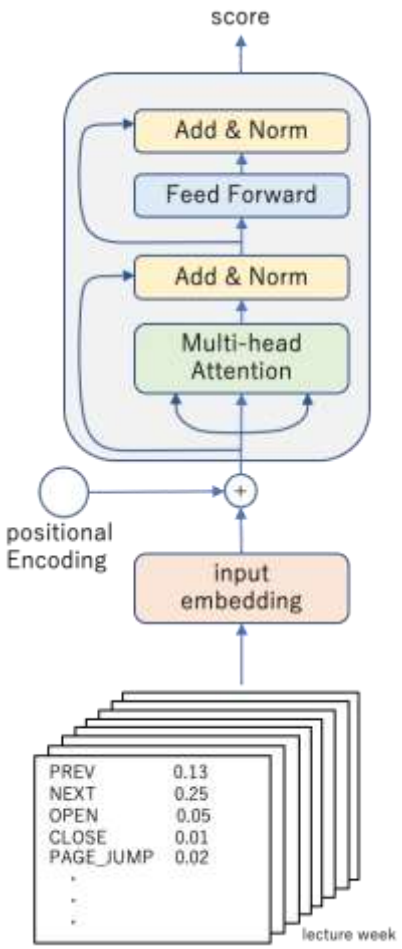

Figure 1. Encoder part of Transformer

2.2.1 Positional Encoding

Since Transformer does not use recursion or convolution to capture the positional information of the input data, the order of the input data is not considered. Therefore, after the input is passed through the Embedding layer, the position of the input is added to the embedded representation using Positional Encoding. The positional encodings to add have the same dimension d_{model} as the embeddings. Specifically, the values of sin and cos functions with different frequencies expressed by the following formulas are added to the vector.

$$PE_{(pos,2i)} = sin\left(pos/10000^{2i/d_{model}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(pos/10000^{2i/d_{model}}\right)$$

Where pos is the position and i is the dimension. In this way, the positional encoding is able to capture the positional information of the input data even if parallel processing is performed for each element.

2.2.2 Multi-Head Attention

Attention is a score that indicates the importance of which data should be paid more attention to when understanding the features of the input data, so that the features can be efficiently obtained from the data. In Transformer, Self-Attention is used to acquire the importance within the same input data, and it is calculated from the three vectors of Query, Key, and Value that each input has as follow.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k is the dimension of Query and Key. In Multi-Head Attention, multiple queries, keys, and values are created for each input, and calculations are performed on each of them to obtain more useful information.

3. Experiment

We estimate test scores from student activity data in university lectures using Transformer and other machine learning methods.

3.1 Experimental setting

In this experiment, we use the learning activity data of students who attended lectures on information-related subjects at Kyushu University in Japan. The learning activity data was collected by the M2B learning support system [7]. The lectures were held eight times, and different lecture materials were prepared for each lecture session. The lecture materials are used online as electronic learning materials, and the type of action taken, the page number of the subject, and the time stamp are recorded. In addition, students who attended the course took one test with a score of 100 points after the completion of all lectures.

In this experiment, we calculate features related to each student's learning activities from the collected data and estimate test scores as a regression problem. The number of times each student performed 16 different actions on the lecture materials is counted for each of the 8 lecture materials. The types of actions are as follows.

- PREV : Return to previous page
- NEXT : Proceed to next page
- OPEN : Open lecture materials
- CLOSE : Close lecture materials
- PAGE_JUMP : Jump to the specified page
- ADD_BOOKMARK : Add page to bookmark
- DELETE_BOOKMARK : Delete page to bookmark
- BOOKMARK_JUMP : Jump to the specified bookmark
- ADD_MARKER : Add marker to the page
- DELETE_MARKER : Delete marker to the page
- ADD_MEMO : Add memo to the page
- DELETE_MEMO : Delete memo to the page
- CHANGE_MEMO : Change the contents of the memo
- SEARCH : Search for a phrase
- SEARCH_JUMP : Jump to the location of the searched word
- LINK_CLICK : Click the link

The total number of features is 128. The largest value calculated for each action is set to 1, and the number of times all students performed the action is normalized to be the input data.

We use LSTM, Multilayer Perceptron (MLP), and Multiple Regression Analysis (MRA) as comparison methods. Transformer uses only the encoder part and treats the output from the encoder as the output value. In the LSTM, the hidden layer is 16-dimensional, and the features from the first lecture are input in order, and the output from the LSTM with the features from the eighth lecture is made 1-dimensional in the output layer. The MLP consists of two 128-dimensional intermediate layers and an output layer. The data of 1200 students who attended the lecture are used for training the network, and the data of 100 students are used for evaluation.

3.2 Comparison results with conventional methods

Table 1 shows the Root Mean Squared Error (RMSE) of the estimated test scores for the evaluation data after 10 training runs of 200 epochs with each method. Transformer has the highest accuracy among the compared methods. From this results, it can be seen that Transformer is the best compared to other machine learning methods.

Table 1. *Comparison result of RMSE*

methods	Transformer	LSTM	MLP	MRA
RMSE	**9.10±0.69**	9.34±0.52	9.37±0.53	10.02±0.52

3.3 Comparison result of input data shape

In the experiment described in Section 3.2, the number of times 16 different actions were performed as input data to the Transformer was calculated for each of the 8 lecture materials. In other words, eight 16-dimensional feature values are input. In this experiment, we change the input data to the number of actions in the 8 lecture materials for each of the 16 types of actions, i.e., we input 16 8-dimensional feature values.

Table 2 shows the RMSE of the estimated test scores for the evaluation data after 10 training runs of 200 epochs, before and after changing the input data shape. The results show that the input data shape before the change (as in sec. 3.2) was more accurate than the input data shape after the change. This suggests that the information on which lecture session the students were more active is more useful than the information on what actions they took throughout the entire test.

Table 2. *Accuracy of input data changing shape*

input shape	8 × (16 dim)	16 × (8 dim)
RMSE	**9.10±0.69**	9.35±0.64

4. Investigating of Attention

The Transformer uses the Attention to the input data, and the data with strong Attention is the data with high importance for inference. In other words, data with a strong Attention may indicate behavior that greatly affects the test score. This may help improve student behavior.

Table 3 shows the Attention to the evaluation data by the Transformer trained on the input data in the form calculated for each of the eight lecture materials. The values of the Attention are the values when the largest value is 100. From the table, we can see that the activity data in the latter lecture is emphasized and strong Attention is generated. In other words, the activity in the lecture immediately before the test is more important for estimating the test score in this experiment.

Table 3. *Attention score of each lecture*

# of lecture	1	2	3	4	5	6	7	8
Attention	45.64	47.80	50.15	49.86	62.66	73.12	78.53	**100.00**

In addition, Table 4 shows the Attention of 16 types of actions when input as series information. The frequency in the table is the ratio of the number of times the action was performed to the total number of actions. The strongest Attention was paid to "NEXT", which has the highest number of actions. The next strongest actions were "SEARCH_JUMP" and "LINK_CLICK" with very few actions. This indicates that "SEARCH_JUMP" and "LINK_CLICK", which occur less frequently, are also important for performance prediction.

Table 4. *Attention score of each activity action*

Action	PREV	NEXT	OPEN	CLOSE	PAGE_JUMP
Frequency (%)	29.61	59.53	1.13	0.60	2.35
Attention	31.93	**100.00**	18.07	30.35	22.51

ADD_ BOOKMARK	DELETE_ BOOKMARK	BOOKMARK _JUMP	ADD_ MARKER	DELETE_ MARKER	ADD_MEMO
0.44	0.07	0.33	4.55	0.66	0.47
20.75	30.99	17.87	34.53	41.53	23.02

DELETE_ MEMO	CHANGE_ MEMO	SEARCH	SEARCH_ JUMP	LINK_CLICK
0.02	0.12	0.06	0.03	0.03
21.79	25.61	42.26	68.71	58.90

5. Conclusion

In this paper, we proposed a Transformer-based performance prediction method.
Compared with conventional machine learning methods, the RMSE of the proposed method can be reduced. It was also found that the feature values calculated for each of the lecture material are superior to each of the action, as input to the Transformer.

In addition, we analyzed the behaviors that are useful for performance prediction. As a result, the strongest Attention was paid to "NEXT", which has the highest number of actions. In addition, we found that "SEARCH_JUMP" and "LINK_CLICK" are important behaviors in spite of their low frequency of occurrence.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP18H04125, Japan.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems*, 6000–6010.

[2] Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). A Neural Network Approach for Students' Performance Prediction. *7th International Learning Analytics and Knowledge Conference*, 598-599.

[3] Ding, M., Yang, K., Yeung, D., & Pong, T. (2019). Effective Feature Learning with Unsupervised Learning for Improving the Predictive Models in Massive Open Online Courses. *9th International Learning Analytics and Knowledge Conference*, 135-144.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186

[5] Dosoviskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterhiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.
[6] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. *Interspeech*. 5036–5040.
[7] Ogata, H., Taniguchi, Y., Suehiro, D., Shimada, A., Oi, M., Okubo, F., Yamada, M., & Kojima, K. (2017). M2BSystem: A Digital Learning Platform for Traditional Classrooms in University. *7th International Learning Analytics and Knowledge Conference*, 155–162.