SHORT VIDEO RECOMMENDATION SYSTEM: A GCN MODEL INCORPORATING BI-LSTM AND MULTIHEADED ATTENTION MECHANISM

YangJie Lu[1*], Dr. Norriza Hussin[2], Assoc. Prof. Dr. Rajamohan Parthasarathy[3]

Abstract—

With the popularity of short video platforms, recommending personalized short videos for users has become a challenging task. Among them, graph neural network (GNN)based recommendation algorithms have attracted much attention due to their ability to capture the complex relationships between users and items. However, how to better utilize the historical behavioral data of users and how to optimize GNN algorithms are still hot research topics. To this end, this paper proposes a Graph Convolutional Network (GCN) model that incorporates Bidirectional Long Short-Term Memory (Bi-LSTM) and Multi-Head Attention (MHA), referred to as BiL-MA-GCN. BiL-MA-GCN model combines the user's viewing history and the textual information of short videos, and uses GCN networks to describe the relationship between short videos and users to generate a personalized recommendation list. Meanwhile, this paper conducts experiments using three publicly available short video datasets Kuaishou, TikTok and MovieLens. The experimental results prove that the BiL-MA-GCN algorithm can effectively improve the accuracy of recommendations.

Keywords—video recommendation, Bi-LSTM, GCN, Multihead attention, Personalized recommendation

[1*]MSc Research Scholar, Faculty of Engineering, Built Environment and Information Technology SEGi University Kota Damansara, Malaysia nhjie@163.com
[2]Senior Lecturer, Faculty of Engineering, Built Environment and Information Technology SEGi University, norriza@segi.edu.my
[3]PhD Research Supervisor, Faculty of Engineering, Built Environment and Information Technology SEGi University Kota Damansara, Malaysia Kota Damansara, Malaysia, prajamohan@segi.edu.my

***Corresponding Author:** YangJie Lu
*MSc Research Scholar, Faculty of Engineering, Built Environment and Information Technology SEGi University Kota Damansara, Malaysia nhjie@163.com

DOI: - 10.48047/ecb/2023.12.si5a.0304

INTRODUCTION

In today's digital era, digital content as well as forms are constantly enriched. Short videos have become an important part of people's entertainment life with the advantages of simplicity of production, diversity of genres, and timely distribution [1].However, the massive amount of videos on short video platforms is often overloaded for users, and how to accurately recommend short videos of interest to users has become an urgent problem. The short video recommendation system aims to improve users' viewing experience by learning their interests and semantic information of videos and building a reasonable algorithm to recommend personalized short videos to users [2].

Currently, common recommendation system models include collaborative filtering-based models, content-based recommendation models, deep learning models, etc. [3].However, these models still have some problems.

Collaborative filtering-based models rely on similarity among users and cannot effectively deal with the cold-start problem and sparsity. Content-based recommendation models can solve the cold-start problem, but they require a large amount of domain knowledge and labeled data and cannot handle the evolution of users' interests well. The short video recommendation system is a direction of great interest in the recommendation field, and some researchers have studied the combination of neural networks and factorization machines to accomplish the prediction of the corresponding click-through rate [4].Some researchers this is done by studying the combination of deep confidence networks (DBN) and collaborative filtering algorithms for video recommendation [5].Some researchers have studied LSTM recurrent neural networks to accomplish personalized movie recommendations [6].But, these researchers are missing modeling information about the sequence of user interaction behaviors and do not capture the relationship between short videos and users well.

In this paper, we propose a graph convolutional neural network model incorporating Bidirectional Long Short-Term Memory and multi-headed attention mechanism, referred to as BiL-MA-GCN model, for short video recommendation systems.

Compared with traditional recommendation systems, this study focuses on the relationship between short videos and users. The content text information of short videos and users' short video viewing history are converted into vector representations, and the relationship between short videos and users is modeled using GCN models to better predict users' future behavior, more comprehensively understand the relationship between users' interests and short videos, and provide more accurate recommendations.

RELATED TECHNOLOGY INTRODUCTION

Global Vectors for Word Representation (GloVe)

GloVe is a word vector representation method based on global word frequency statistics, whose core idea is to represent each word as a vector by statistical analysis of global word frequency, and let the inner product of this vector be equal to the log probability that these two words occur together, so that these vectors can represent the semantic relationship between words [7]. Specifically, the training process of GloVe is divided into two steps. First, it counts the number of co-occurrence of each word pair in the global and uses this co-occurrence information to construct a word-word co-occurrence matrix X. GloVe uses least squares as the loss function optimization algorithm and adds offset terms to the rows and columns in the co-occurrence matrix X [8], as in:

$$J = \sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \quad (1)$$

|V| represents the size of the lexicon, w_i and w_j are the word vectors of the context, and b is the offset value of the cooccurrence matrix. And the weight function f(x) is as follows, when x=0, f(0)=0. Among them, the GloVe model author Pennington's experiments are known to have the best feature representation for x_{max}=100 and α=0.75, for which this value is also used in this paper. The weighting function f(x) is as in:

$$f(x) = \begin{cases} (x/x_{max})^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Bidirectional Long Short-Term Memory (Bi-LSTM)

Bidirectional long and short term memory network (BiLTSM) is a recurrent neural network (RNN) based model with two LSTM layers, forward and backward [9]. This network structure is able to learn long-term dependencies in sequences, and since it processes both past and future inputs, we can use it in recommendation applications for short videos to extract vectors of users' history of watching short videos. The computation in the LSTM layer is divided into three gates: the input gate (how much of the input value is retained at the current moment), the forgetting gate (which determines how much of the previous moment's cell state is retained to the current moment) and the output gate (how much of the controlled cell state is output). The computation of these gates consists of a series of neural network layers, where each neuron has a state value and an output value. The inflow and outflow of information can be controlled, thus controlling the flow of information in the network. the LSTM is composed of a series of LSTM units, as shown in Figure 1 below.

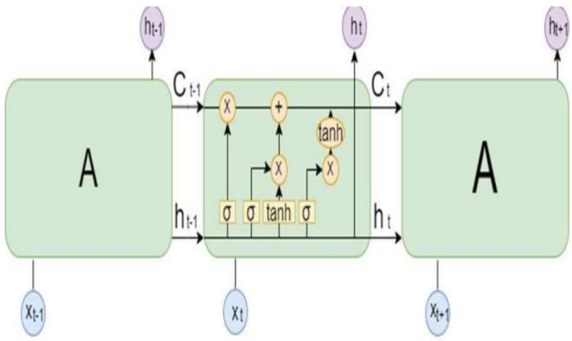

Fig. 1. LSTM unit diagram

In the forgetting gate (W_f), x_t and h_{t-1} are taken as input information and data discarding is done by sigmoid activation function, b_f is the bias condition. As in:

$$f_{\mathrm{t}} = \sigma(W_f\,[h_{\mathrm{t-1}}, x_{\mathrm{t}}] + b_{\mathrm{f}}) \qquad (3)$$

In the input gate (W_t), the sigmoid activation function updates the corresponding state, which is then processed by the tanh function to output a brand new C_t by combining the current memory C_t and the long-term memory C_{t-1}. As in:

$$i_{\mathrm{t}} = \sigma(W_{\mathrm{t}}\,[h_{\mathrm{t-1}}, x_{\mathrm{t}}] + b_{\mathrm{i}}) \qquad (4)$$

$$\overline{\mathrm{C}}_{\mathrm{t}} = \tanh(W_C[h_{t-1}, x_t] + b_c) \quad (5)$$

$$\mathrm{C}_{\mathrm{t}} = \mathrm{f}_{\mathrm{t}} \times \mathrm{C}_{\mathrm{t}} + \mathrm{i}_{\mathrm{t}} \times \overline{\mathrm{C}}_{\mathrm{t}} \quad (6)$$

In the forgetting gate (W_f), x_t and h_{t-1} are taken as input information and data discarding is done by sigmoid activation function, b_f is the bias condition. As in:

$$O_{\mathrm{t}} = \sigma(W_0\,[h_{\mathrm{t-1}}, x_{\mathrm{t}}] + b_0) \ (7)$$

$$h_{\mathrm{t}} = O_{\mathrm{t}} \times \tanh(C_{\mathrm{t}}) \ (8)$$

The Bi-LSTM network is superimposed by the forward LSTM and the reverse LSTM. The role of capturing past and future data characteristics is combined, as shown in Figure 2 below.

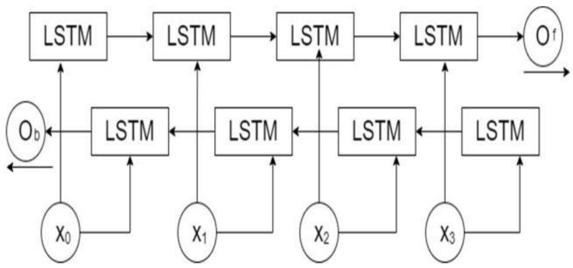

Fig. 2. Bi-LSTM network structure diagram

Multi-headed attention mechanism

In past studies, researchers have found that the human brain does not attend to all information without differences when processing visual information. Instead, they focus on a visual area of interest, for which attentional mechanisms have been proposed [10]. And on the short video recommendation system, we can use the attention mechanism to extract the relationship between short videos and users.

The attention mechanism is generally divided into a query vector Q, a key matrix K and a value matrix V [11]. The query vector Q and K values are calculated using the similarity function, which is generally used as Scaled Dot-Product Attention, with d_K being the length of the word vector, as in:

$$Attention(Q, K, V) = softmax(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d}_{\mathrm{K}}}})\,V \quad (9)$$

In considering the magnitude of the error between predicted and actual values, the mean square error can be used and the corresponding scoring matrix of the iterations can be detected by this function to minimize the total error. In addition, the implicit feature vector of P_u users and the implicit feature vector of Q_i items are added to act as regularization terms, which prevents the corresponding models from overlearning and thus overfitting. And the corresponding λ regularization coefficients can be obtained by always cross iterating the validation [10], as in:

$$Q = XW^{\mathrm{Q}} \qquad (10)$$

$$K = XW^{\mathrm{K}} \qquad (11)$$

$$V = XW^{\mathrm{V}} \qquad (12)$$

As for the multi-headed attention mechanism, for the same input matrix X, we can define multiple groups of different training weight matrices, such as W_0^Q , W_0^K , W_0^V , W_1^Q , W_1^K , W_1^V ,compute the Q/K/V of each group, and finally learn different parameters, as shown in Figure 3 below.

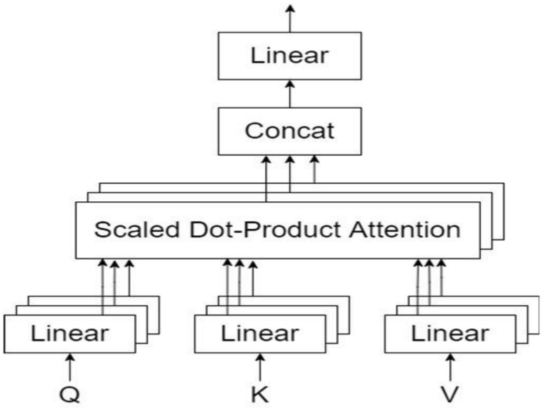

Fig. 3. Multi-headed attention structure chart

Graph Convolutional Neural Network (GCN)

The graph neural network demonstrates its powerful processing power in modeling the relationship between data nodes. For entity recognition as well as for relationship extraction, graph convolutional networks are used more often [12]. GCN is an efficient graph-based neural network model that allows representation of nodes and graphs. The update of each node depends only on the feature representation and weight matrix of its neighboring nodes, which has the characteristics of parameter sharing and efficiency, so it has achieved good results in many applications, as shown in Figure 4 below.

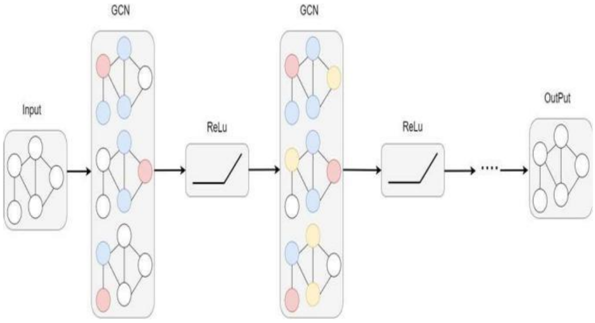

Fig. 4. GCN diagram

The core idea of GCN is neighbor-based aggregation, where the information of each node and its neighboring nodes is aggregated to obtain global information [13]. Suppose we have a graph G= (V, E), where V denotes the set of nodes, E denotes the set of edges, and each node i has an edge e_{ij} with weights between it and its neighbor node j. The GCN model can be used to update the feature representation of node i with the following equation, as in:

$$h_i^{(l+1)} f(\sum_{j\in N_i} \frac{1}{c_{ij}} h_j^l W^l) \quad (13)$$

In the formula (13), h_i^l denotes the feature representation of node i at layer l, h_i^0 denotes the initial feature, N_i denotes the set of neighboring nodes of node i, W^l denotes the weight matrix at layer l, c_{ij} denotes the normalization constant, and the function f is usually a nonlinear activation function, such as ReLU.

BIL-MA-GCN ALGORITHM

In short video recommendation, users have diverse interests and personalized preference differences, and the diversity of topics, styles and languages of short videos also make the differences between short videos. The interaction between users and short videos also intensifies the complex relationship between users and short videos, so the traditional recommendation algorithm is more difficult to handle. To this end, this paper proposes the BIL-MA-GCN model, which is able to capture the complex relationship between users and short videos from different perspectives and at different levels by using a multilayer neural network structure.

Specifically, by using a text embedding layer and a BiLSTM layer, the short video titles and description contents are converted into vector representations, and the short video sequences watched by users are converted into vector representations to better learn the short video and user representations. With the multi-headed attention mechanism, the relationship between short videos and users can be effectively captured and the relationship representation between short videos and users can be obtained to better understand the interaction between short videos and users. By using the GCN model to model the relationship representation between short videos and users, the updated vector representation of short videos and users is obtained, which can better capture the relationship between short videos and users and improve the accuracy of recommendations, as shown in Figure 5 below.

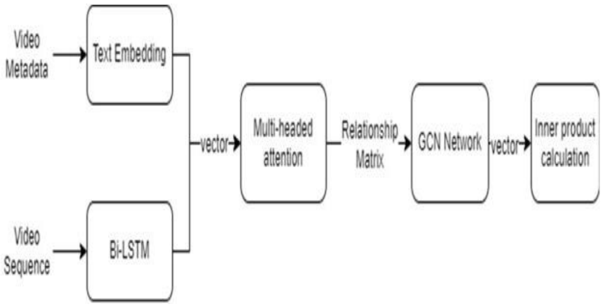

Fig. 5. Structure diagram of BIL-MA-GCN model

The algorithm flows roughly as follows:

1) Text embedding layer: converting text information of short videos into vector representation.
Input Data: Title and content description of the short video
Output Data: Vector representation of short videos
2) Bi-LSTM layer: converting short video sequences watched by users into vector representations
Input Data: Short video sequences that users have watched
Output Data: Vector representation of the user
3) Multi-headed attention layer: capturing the relationship between short videos and users
Input Data: Vector representation of short videos and users
Output Data: The relationship matrix between short videos and users
4) GCN layer: modeling the relationship between short videos and users
Input Data: The relationship matrix between short videos and users
Output Data: Updated short video and user vector representation
5) Recommendation score calculation layer: Short video recommendation scores are calculated using inner product.
Input Data: Updated short video and user vector representation
Output Data: Short Video Recommendation Score

EXPERIMENT AND RESULTS

The short video recommendation model in this paper uses text embedding and Bi-LSTM to extract short video text vectors and user viewing video history sequences, combines a multi-headed attention mechanism to model the relationship between users and short videos, uses a graph convolutional neural network to update the vector representation of users and short videos, and uses a recommendation score calculation layer to compute recommendation scores. We want to validate the performance of this model in practice and compare it with other recommendation algorithms. Therefore, we propose several experiments as follows.

- Experiment 1: Comparing the performance of using different numbers of attention heads
- Experiment 2: Comparing the performance of using different numbers of GCN layers.
- Experiment 3: Overall model performance comparison experiment

Experimental environment

For the experimental environment of the BiL-MA-GCN algorithm proposed in this paper, as in:

Table I. Experimental Software/Hardware Environment Configuration

Physical Environment	
Hardware	Configuration
CPU	Intel Core i9-11900K 3.5GHz
GPU	NVIDIA GeForce RTX 3080
RAM Memory	32GB DDR4-3200MHz
Storage Hard Drives	1TB M.2 NVMe SSD
Software Environment	
System / Software	Configuration
Operating System	Ubuntu 18.04
Deep Learning Framework	PyTorch 1.8.1
Python Versions	3.8.5
CUDA Versions	11.1

Experimental data

Three publicly available video datasets are mainly used for the experimental data, as in:

Table Ii. Short Video (Movie) Dataset

Dataset	Users	Videos	Interactions	Visual	Textual
MovieLens	71567	10681	10000054	128	128
KuaiShou	37819	115145	1622592	2048	128
TikTok	65531	768546	4100167	128	128

1) Movie Lens

MovieLens is commonly used for performance evaluation of recommendation models. This dataset contains the user's interaction behavior with the movie and the corresponding textual information about the movie (title, details).

2) KuaiShou

The KuaiShou short video dataset is a relatively mainstream video set among short video platforms in China. The video set contains information about the user's interaction with the short video (excluding sensitive information), short text messages.

3) TikTok

TikTok is a foreign version of China's hottest short video platform. Its short video collection contains user interaction behavior (like or not, follow, etc.), as well as text information of short videos.

Evaluation indicators

1) Precision

The proportion of users clicking to watch a video in the short video recommendation list can be called the accuracy rate, which can evaluate the proportion of relevant strengths in the retrieved instances. U is the set of all users, R(u) is a list of short videos recommended by users, and T(u) is a list of short videos actually watched by users, as in:

$$Precision = \frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \quad (14)$$

2) Recall

In the short video recommendation list, the ratio of the number of videos watched by users stands the number of all videos watched by users is the recall rate. It is actually the ratio of retrieved relevant instances to all relevant instances, as in:

$$Recall = \frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \quad (15)$$

3) Normalized Discounted Cumulative Gain(NDCG)

The short video recommendation list can be rated using NDCG's score. Compare the videos in the recommendation list with the videos watched by users, the higher the position of the common video in the recommendation list, the better the recommendation effect. DCG is a cumulative value that needs to be normalized, and IDCG is the maximum DCG value under ideal conditions. |REL| is the top p result sets in descending order of relevance, as in:

$$NDCG = \frac{DCG_p}{IDCG_p} = \frac{\sum_{i=1}^{p}\frac{2^{rel_{i-1}}}{log_2(i+1)}}{\sum_{i=1}^{|REL^p|}\frac{2^{rel_{i-1}}}{log_2(i+1)}}, rel \in \{0,1\} \quad (16)$$

Experimental design and analysis of results

1) Experiment 1: Comparing the performance of using different numbers of attention heads.

It is very common to use attention mechanisms in short video recommendation models, which allow the model to focus more on video features relevant to the user. However, for different application scenarios and datasets, choosing different numbers of attention heads may have different effects on model performance, as shown in Figure 6 below.

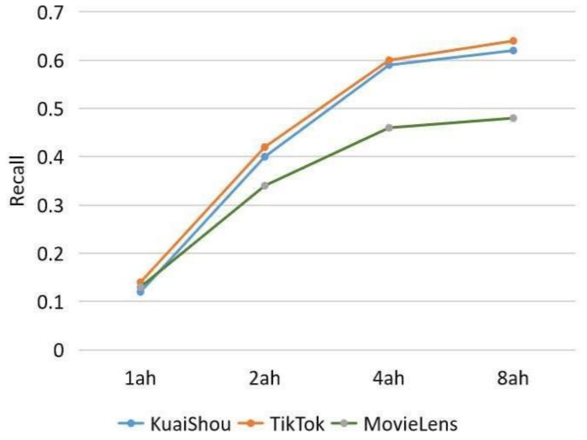

Fig. 6. Comparison diagram of different number of attention heads

Therefore, our experiments aim to compare the effects of different numbers of attention heads on the performance of the short video recommendation model in order to further optimize the model. Comparisons of metrics for Recall on different data sets were considered for 1, 2, 4 and 8 attention heads, respectively. From the experiment, it can be seen that the best effect is achieved when the 8-headed attention mechanism is selected, but considering that the effect of selecting 4 and 8 heads is close, the 4-headed attention mechanism is selected for this experiment.

2) Experiment 2: Comparing the performance of using different numbers of GCN layers.

Evaluate the impact of different numbers of GCN layers on the performance of the short video recommendation model to determine the optimal number of GCN layers. An initial model is constructed using the optimal number of attentional heads in Experiment 1 as a benchmark. Set up 1, 2, 3 and 4 layer GCN models for comparison, as shown in Figure 7 below.

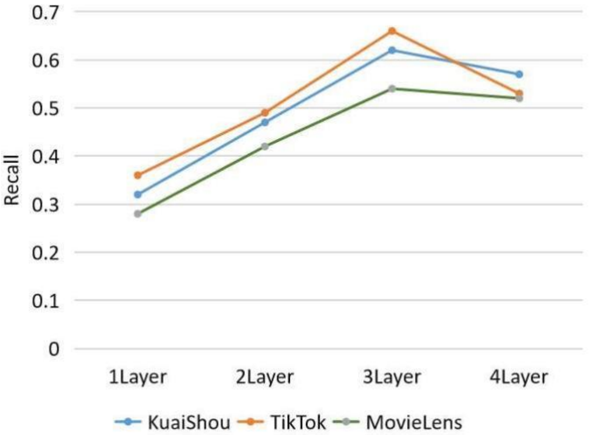

Fig. 7. Comparison diagram of different number of attention heads

3) Experiment 3: Overall model performance comparison experiment.

In order to better test the performance of BiL-MA-GCN model, we selected GraphSAGE [14], GAT [15] model, which is the same method of graph convolutional network, as the comparison model, as in:

Table Iii. Three Models Performance Comparison Table

Dataset	Indicators	Models		
		Graph SAGE	GAT	BiL-MAGCN
KuaiShou	Precision	0.271	0.278	0.341
	Recall	0.302	0.341	0.353
	NDCG	0.186	0.212	0.223
TikTok	Precision	0.184	0.191	0.223
	Recall	0.471	0.472	0.482
	NDCG	0.296	0.311	0.362
MovieLens	Precision	0.122	0.137	0.145
	Recall	0.421	0.431	0.453
	NDCG	0.277	0.286	0.312

We compare the performance of BiL-MA-GCN model, GraphSAGE model and GAT model to cope with different datasets. Comparing multiple dimensions from the three datasets, the models proposed in this paper have certain advantages.

CONCLUSIONS

This paper mainly introduces a GCN-based short video recommendation model. First, the paper introduces the background and significance of short video recommendation, and proposes a short video recommendation model based on graph neural network for the limitations of traditional recommendation algorithms.

In terms of model design, the article incorporates text embedding and Bi-LSTM, and uses attention mechanism techniques to enhance the expressive power of the model. To verify the effectiveness of the model, the article examines the aspects of different attention heads, different GCN layers, and comparison with GraphSAGE and GAT models through three experiments, respectively. The experimental results show that our proposed model outperforms the comparison algorithm in all metrics.

However, the algorithm does not consider video information (content, tags, location), etc. Subsequent work will try to incorporate these parameters to achieve more accurate short video recommendations.

REFERENCES

1. Zhang Shu Han, Kong Chao Peng, Kong Jing Yuan. Research on the influence of short video information dissemination in the new media era[J]. Intelligence Science,2021,39(09):59-66.
2. Sun Y. Research on real-time recommendation method of short video based on knowledge graph [D]. Liaoning University,2022.
3. Wang Lei. Research on recommendation algorithm and system based on deep learning [D]. Beijing University of Posts and Telecommunications, 2021.
4. Guo H, Tang R, Ye Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction [J]. arXiv preprint arXiv: 1703.04247, 2017.
5. Hongliang C, Xiaona Q. The video recommendation system based on DBN[C]//2015 IEEE International Conference on Computer and
6. Information Technology; Ubiquitous Computing and Communications;
7. Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing. IEEE, 2015: 1016-1021.
8. Liu J, Choi W H, Liu J. Personalized movie recommendation method based on deep learning[J]. Mathematical Problems in Engineering, 2021, 2021: 1-12.
9. Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
10. Chen Zhenrui,Ding Zhiming. A word vector improvement method based on GloVe model[J]. Computer System Applications, 2019,28(01):194-199.
11. Wu Z, Huang M, Zhao A. Traffic prediction based on GCN-LSTM model[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1972(1): 012107.

12. CHEN J，ZHANG H，HE X，et al.Attentive collaborative filtering：multimedia recommendation with item-and component-level attention[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval，2017:335-344.
13. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 31st Conference on Neural Information Processing Systems 2017(NIPS 2017). Long Beach, CA, USA: 2017: 5998–6008.
14. D. Lin Gao. Joint Extraction of Entity Relations with Fusion Attention Mechanism and Graph Neural Network [D]. China University of Mining and Technology, 2021.
15. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:1609.02907, 2016.
16. Will Hamilton, Zhitao Ying, Jure Leskovec. Inductive representation learning on large graphs [C]. In Proceedings of International Conference on Neural Information Processing Systems, 2017:10241034.
17. Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio. Graph attention networks [C].In International Conference on Learning Representations (ICLR),2018.