sustainability

MDPI

Article

A Deep Learning Framework for Multimodal Course Recommendation Based on LSTM+Attention

Xinwei Ren [1,2], Wei Yang [1,*], Xianliang Jiang [2], Guang Jin [2] and Yan Yu [2]

[1] College of Science & Technology, Ningbo University, Ningbo 315300, China

[2] College of Electrical Engineering and Computer Sciences, Ningbo University, Ningbo 315211, China; akon_ren@163.com (X.R.); jiangxianliang@nbu.edu.cn (X.J.); jinguang@nbu.edu.cn (G.J.); 1911082064@nbu.edu.cn (Y.Y.)

* Correspondence: yangwei1@nbu.edu.cn

Abstract: With the impact of COVID-19 on education, online education is booming, enabling learners to access various courses. However, due to the overload of courses and redundant information, it is challenging for users to quickly locate courses they are interested in when faced with a massive number of courses. To solve this problem, we propose a deep course recommendation model with multimodal feature extraction based on the Long- and Short-Term Memory network (LSTM) and Attention mechanism. The model uses course video, audio, and title and introduction for multimodal fusion. To build a complete learner portrait, user demographic information, explicit and implicit feedback data were added. We conducted extensive and exhaustive experiments based on real datasets, and the results show that the AUC obtained a score of 79.89%, which is significantly higher than similar algorithms and can provide users with more accurate recommendation results in course recommendation scenarios.

Keywords: multimodal; deep learning; course recommendation; multimodal recommendation

check for updates

Citation: Ren, X.; Yang, W.; Jiang, X.; Jin, G.; Yu, Y. A Deep Learning Framework for Multimodal Course Recommendation Based on LSTM+Attention. *Sustainability* **2022**, *14*, 2907. https://doi.org/10.3390/su14052907

Academic Editor: Miltiadis D. Lytras and Andreea Claudia Serban

Received: 13 January 2022

Accepted: 18 February 2022

Published: 2 March 2022

1. Introduction

With the rapid development of the Internet and online education, traditional education methods are constantly being transformed, and students are no longer satisfied with the original rigid curriculum and tend to choose courses they are interested in through online education platforms. Platforms can provide high-quality, targeted teaching courses, a complete curriculum, and supporting assignments, while having the advantages of free courses, rich content, and flexible choices. However, the exponential growth of platforms and courses leads to "information overload", and users are easily confused and have difficulty in choosing courses when facing many classes. Therefore, it is crucial to help users quickly choose the right course that interests them. When users use the online education platform to learn new courses, although the courses are classified, in the face of many courses, quickly locating exciting and personalized valuable courses is of great significance. To recommend courses to users, a recommendation model based on course content was proposed [1]. At present, to recommend appropriate courses, it is necessary to shorten the selection distance between users' learning ability and a large number of courses.

Currently, there are many problems with course recommendations. Only considering the user's nearest neighbor recommendation method and integrating basic personal information and social preferences for the recommendation, the user's implicit feedback data is ignored [2]. However, the multimodal graph convolution model is used to generate the specific modal representation of users and videos to capture better the implicit feedback data of users [3]. In addition, Gong et al. [4] used the knowledge concept of graph neural networks for course recommendation, considering the fusion of heterogeneous information to grasp the differences in the interests of different students. However, the implicit and explicit feedback information of learners should be fully regarded in the extraction of course features; moreover, course-related information can further enhance the recommendation efficiency. At the same time, because the traditional course recommendation system and algorithm mainly focus on the experiment and test of single-mode information and lack effective utilization and research of multi-mode information, the system cannot obtain the deep-seated preferences of users' construct and complete the user portrait. Therefore, this paper proposes to use LSTM and Attention deep learning to process multimodal data for efficient course recommendations.

When users use the platform to learn a new course, the first search for the course name and read the course description to understand the course details. Users also consider the total review scores and the number of reviews about the course on the platform. Additionally, demographic information [5,6] represents the user's characteristic data, so that course text and numerical information can be used as recommendation data. When viewing a course, users mark the course according to favorites, collections, comments, and so forth, and such behaviors can also be used as explicit feedback data; users' browsing behaviors just for the course can be used as implicit feedback data. In addition, image and acoustic information can be extracted from the course videos by using deep learning models. Therefore, digital, textual, video, and audio information of the course can be used when building the course recommendation model, and users' explicit and implicit feedback data can be used as inputs. Short video recommendations have developed rapidly in the industry and academia in recent years. Videos contain rich modal features, including video, audio, text, tags, and other multimodal data. Through these multimodal features, users' deep interests can be mined. Therefore, it is essential to propose a multimodal in-depth learning recommendation system, which can be applied to the online education platform and improve the accuracy of personalized course recommendations.

The main contributions of this paper are as follows: first, multimodal data were used as model input, mainly including course text, video, and audio information; meanwhile, user demographic information, explicit and implicit feedback are also considered. Second, this paper designs a new multimodal deep learning course recommendation model for extracting the modal features of learners and courses; the model improves and optimizes the traditional multimodal model to enhance its effectiveness in the field of course recommendation. The practical significance of this research is to improve recommendation accuracy and increase the users' click-through rate of courses by merging various modal information.

The rest of this paper is organized as follows. Section 2 is about related work, summarizing the current literature related to online course platforms and multimodal course recommendation research. Section 3, the algorithm framework, describes the extraction of course features and the overall framework of the algorithm. The experimental section in Section 4 compares the impact on course recommendation results by different modal course information and learners. Finally, the conclusion and future work section gives possible research directions in multimodal course recommendations.

2. Related Works

2.1. Traditional Course Recommendation

Recommendation systems have been widely used in music, movies, videos, and other fields in recent years. Traditional course recommendations mainly include content-based recommendations, collaborative filtering recommendations, and hybrid recommendations combining both advantages. Content-based recommendation mainly considers the demographic information of learners, such as gender, age, major, and the course text description, overall rating, and number of reviews. Khribi et al. [7] proposed an online learning platform in which the offline module is loaded to pre-process data to build a model with a learner–content relationship, and the online module dynamically identifies learner needs and goals and predicts the recommendation list. However, the cold start of the system is not considered [8]. For the cold start problem, Sengottuvelan et al. [9] proposed to use multiple attributes of learners to model their preferences, use data mining techniques to model learners' history records, and finally sort them according to the similarity threshold. The K-means clustering algorithm effectively reduces data sparsity and cold start problems and increases the diversity of ecological annotation lists. In addition, demographic information, learning emotion, motivation, and learners' style can also be studied as modal information (Fu et al. [10]). The advantage of content-based recommendation is understanding user preferences fully and not considering data sparsity. At present, the main advantages of content-based course recommendation methods are: do not consider the sparse data, and the recommended content depends on user preferences. However, its main disadvantage is that it requires the feature content to have a good structure and only considers the user's preferences, ignoring the situation of other users.

Collaborative filtering algorithm recommendations can be divided into user-based, and item-based depending on the object. Firstly, based on the user's collaborative filtering algorithm, the user behavior logs are collected for analysis, and the user interest model vector is constructed; secondly, a label is created for each resource in the system, and the user score of the resource is collected in the process of user use; finally, according to the user interest model and resource characteristics, combined with the personalized recommendation algorithm, the resources that meet the needs of users are recommended to the target users [11]. In order to effectively alleviate the problems of cold start and sparse data, a nearest-neighbor collaborative filtering algorithm based on weight optimization according to user history score is proposed [12]. In item-based collaborative filtering, the similarity matrix among courses is first calculated, and then the similar courses are further ranked and recommended based on the user's positive feedback history of similar courses. Pang et al. [13] proposed a multilayer Bucketing Model (MLBR) for MOOC course recommendation, which first transforms the learner vector into the same dimension and disperses it into buckets containing similar learners that have more common courses. Collaborative filtering can fully consider similar learner or course attributes without concerning about the content attributes of the courses themselves compared to content-based recommendation algorithms. In short, compared with the content-based recommendation algorithm, the advantage of collaborative filtering is that it can fully consider the attributes of similar learners or courses without considering the content attributes of the course itself. However, the actual situation is that users have little evaluation of the course, which leads to the problem of sparse data, and there is no user using data in the initial stage of the platform. Such problems will also be encountered after the new course is uploaded. At the same time, traditional collaborative filtering is effective in dealing with small data sets, but in the face of massive data sets, the accuracy of the recommendation system will decline.

However, all the above recommendation algorithms have corresponding disadvantages, such as cold start problems in content-based recommendations and sparse data and cold start for new users in collaborative filtering recommendations. Therefore, deployments often combine the advantages of different algorithms and models for hybrid recommendations. For example, Burke et al. [14] classified hybrid recommendation combination strategies into weighting, switching, partitioning, hierarchy, waterfall, feature blending, and feature enhancement, and introduced hybrid recommendation systems based on knowledge and collaborative filtering. Jannach et al. [15] classified the hybrid recommendation approach into holistic, parallel, and pipelined, where firstly, artificial neural networks are used to classify learners and users can get course recommendations based on learners' opinions; then when relevant interest groups are established, data mining techniques can be used to elicit the best learning paths. To sum up, the current traditional recommendation algorithms mostly use the learners' preferences and the unimodal information of demographic information in the representation of the curriculum model, ignoring the text information and video information of the curriculum itself. In hybrid recommendations, collaborative filtering and content-based recommendation model fusion are mostly used, but the audio mode of the course and the explicit feedback and implicit feedback of users are not considered.

2.2. Course Recommendation in the Field of Deep Learning

With the application of deep learning in the field of recommendation, it is gradually applied to the field of course recommendation. The deep course recommendation system can be divided into personalized learning recommendations based on convolutional neural networks and recurrent neural networks. Among them is the personalized learning recommendation based on the convolutional neural network: firstly, the learning behavior and learning history of learners are represented as feature vectors; then the correlation is estimated based on the difference between the estimated value and the actual value by using the attention mechanism; finally, the course is recommended to learners by training the recommendation model.

Personalized learning recommendation based on recurrent neural networks: Zhou et al. [16] clustered learners and used the LSTM model to predict learning paths and achievement, and finally recommended personalized and complete learning paths for learners. To extract the emotional factors expressed in the text, Ange et al. used the user-sensitive deep multimodal structure and extracted the rich potential data representation of users, which improved the effect of text classification [17]. Wang et al. [18] proposed to extract features by using learners' behavior and history, combining attention mechanisms and the difference between predicted and actual values of neural networks to improve recommendation performance. At the same time, to improve the attention mechanism, Zhu et al. introduced the double-layer attention mechanism into the summary of the parallel mental network recommendation model through data preprocessing and standardization, which effectively improved the ability of the model to mine the characteristics of users and courses [19]. Liu et al. [20] predicted the list of courses that students are good and bad at in the next semester by a hybrid model of deep learning and collaborative filtering. On the other hand, Ni et al. [21] first constructed a graph convolution network using the bipartite graph of user project interaction, and then integrated multi-task learning into the convolution neural network learning framework using multimodal auxiliary information, improving the classification effect multi-task.

At present, in the process of improving the model structure, the course recommendation in the field of in-depth learning continuously excavates the hidden meaning between the course and the user data, which increases the network depth, increases the calculation time of the model, and reduces the timeliness of recommendations while increasing the recommendation accuracy. Therefore, in mining user characteristics, we need to fully consider the computational performance and the complete modal characteristics of users.

2.3. Multimodal-Based Course Recommendation in Deep Learning Domain

With a large number of applications of deep learning in the fields of image recognition and sentiment analysis, researchers have gradually fused multi-domain data to achieve complementarity between heterogeneous information for more comprehensive information. For example, three kinds of information, such as image, video, and text, are fused in cross-modal embedding [22]. Tamura et al. [23] suggests a multimodal response synchronization measurement system with an eye tracker and electroencephalography signals, through which the eye tracker can obtain information from the learner's attention and the brain signal can provide clues to estimate the mental state in learning. Wang et al. [24] used an online learning system to collect multimodal behavioral data from three dimensions: psychological, physiological, and behavioral, thus providing a more comprehensive evaluation of the overall situation. Xu et al. [25] proposed a multimodal deep learning framework to extract multimodal course information, such as course video, text, and audio, as well as explicit and implicit feedback from users. Chango et al. [26], on the other hand, focused on improving the recommendation performance using different fusion algorithms for four types of multimodal information about students' theory classes, practical classes, online courses, and final grades. However, the multi-modal data of the course need to capture representative modal features in the acquisition process, and the implicit feedback data of users is difficult to obtain. The hidden behavior of users watching videos has essential research significance in the platform. Users' implicit feedback data can be effectively obtained through users' pause times, playback times, and playback time nodes.

Multimodal fusion can provide more information for model decision-making, which can improve the accuracy and precision of the overall decision result. The difficulty lies in handling heterogeneous information, selecting the fusion method, and adjusting the modal alignment.

3. Methods

We design and implement a course recommendation system based on a multimodal deep learning framework, which uses different modal information in the course for fusion recommendation. The user's operation while watching the course video involves explicit and implicit feedback, and therefore, it needs to be differentiated according to the user's behavior. This system mainly includes six key components: data collection, data processing, feature extraction, profiling, deep learning course recommendation and recommended course result presentation. The overall process of this multimodal course recommendation framework is shown in Figure 1.

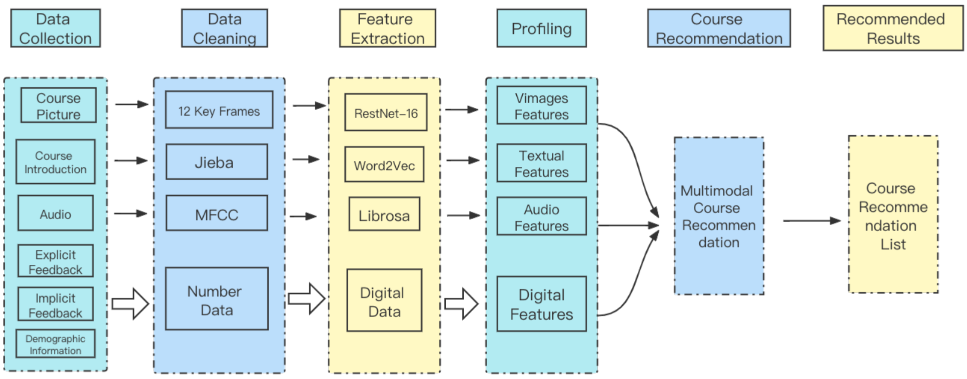

Figure 1. Multimodal course recommendation framework.

3.1. Data Collection Module

The online platform collects data from two aspects: learners and course data. The learner data mainly includes explicit and implicit feedback data. The explicit feedback data is reflected in the comments, collections, and thumbs of the course; the implicit feedback is reflected in the data of only browsing the course without marking. The course data mainly includes the course name and introduction, video data, and audio data extracted from the course video. Figure 2 depicts the main process of acquiring data.

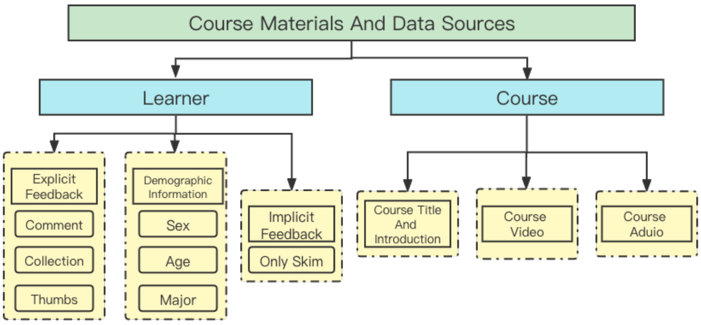

Figure 2. Data collection module.

3.2. Data Processing

The course information collected directly has interference information, so the data needs to be cleaned before extracting the course features. The video information of the course is described by video keyframes. However, due to the long duration of course videos (usually more than 10 min), acquiring video frames frame by frame will cause a sudden increase in data volume. Therefore, In this paper, video description is performed by capturing keyframes. We extracted the audio from the course video to obtain the course audio information and analyzed the obtained audio information. Additionally, the title, introduction, and comments of the course need to be divided into words to clean up the deactivated parts, such as "of" and "the", as well as the markers.

3.3. Feature Extraction Module

3.3.1. Video Feature Extraction

The course video contains visual modality information of the course. To obtain the course video keyframes, Mahasseni et al. [27] implemented the selection of keyframes without keyframe annotation through a per-frame score. Inspired by the above idea, in this paper, the video stream is considered as a series of images, and keyframes are extracted from the video at intervals by the FFmpeg tool. To obtain the high-level features of video frames, for extracting the content of each frame a ResNe-16 model is used to extract visual features, which can provide stable initialization in ensuring semantic recognition. The complete video is transformed into 12 keyframes by initializing the ImageNet dataset with pre-training weights, where each keyframe contains 4096-dimensional features.

3.3.2. Audio Feature Extraction

Due to diverse contents and some missing information in the video part, audio modal information is extracted to supplement the information. The audio modal information includes the tone, intonation, and pauses of the lecturer. For audio modality extraction, we use the Librosa module to process the audio content by executing a 10 s window and 80% overlapping spectrograms, and for obtaining 512-dimensional features of audio modality.

3.3.3. Text Feature Extraction

The text modal information mainly includes the course name and introduction. In this paper, the course name and introduction are processed using the Jieba word separation model, and the information that does not contain specific content is eliminated according to the deactivation word rule. The collection dataset is obtained on a MOOC platform with complex data dimensions. Therefore, a large Chinese corpus is used to capture the semantic information between different words [28]. Besides, we use the Word2Vec tool to convert the text corpus as input and convert the title and course profile content into a 300-dimensional feature vector.

3.3.4. Digital Feature Extraction

To obtain learners' comments, collections, and thumbs data of the course, we marked them by one-hot coding. At the same time, learners' browsing of the detailed interface of the course was marked by the platform script.

Based on the above four kinds of feature data extraction, the course multimodal information extracted in this paper is summarized as follows, which contains 4096-dimensional visual modal information, 512-dimensional audio modal information, 300-dimensional text modal information, and 4-dimensional digital features. The course multimodal information is shown in Table 1.

Table 1. Summary of course multimodal information.

Model Information	Dimension
Course Video	4096
Course Audio	512
Course Text	300
Digital Features	4

3.4. Profiling Module

After obtaining the multimodal features of the course through the feature extraction module, the different modal features need to be reasonably fused. Given that there is an interaction between the demographic information of learners and explicit and implicit feedback data, this paper integrates the above three data into digital modal features. The specified features are first extracted from the sequential units of each modality, and then these features are input into the parallel LSTM separately, and finally, the sequential structures in the visual, audio, text, and digital modal features are captured.

Since the output of the LSTM is a vector sequence model, the variable-length vectors need to be transformed into fixed-length vectors to facilitate the later model training. This paper uses an attention-based pooling operation to achieve dynamic adjustment of weights according to the importance of time steps based on the weighted sum sequences of all time step vector representations and the content relevance of the weights to each time step [29]. The output of the fully connected layer is used as the result of course recommendation. Figure 3 illustrates the schematic diagram of the multimodal course recommendation algorithm model.

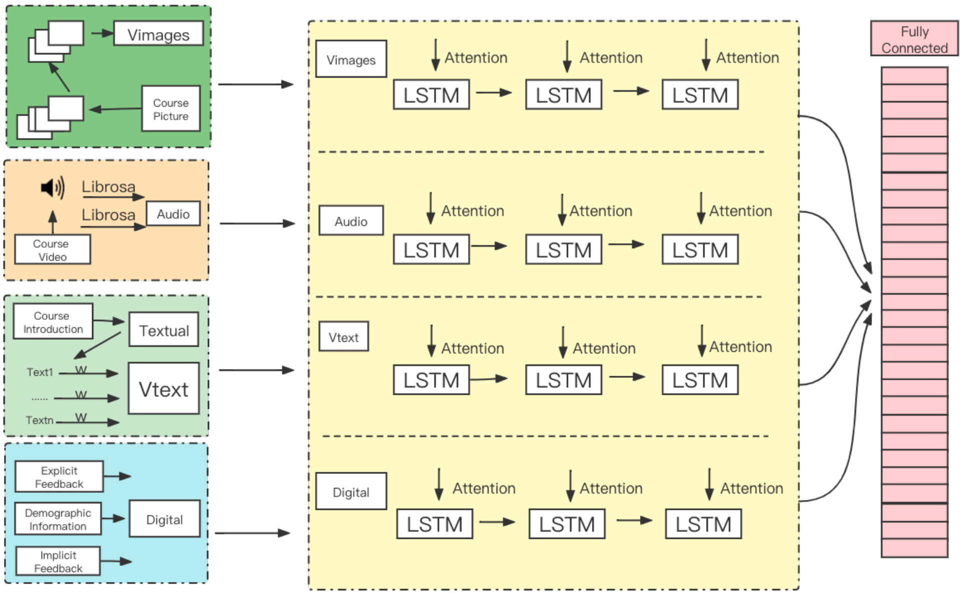

Figure 3. Multimodal course recommendation algorithm model.

3.5. Course Recommendation and Recommendation Results

By extracting and processing the features of different modalities, the output of the fully connected layer is obtained. In this paper, multiple modal information is processed by LSTM and Attention, and visual, audio, text, and digital modal feature vectors are connected vertically to finally obtain a 4912-dimensional course vector, which is used as the input of the recommendation model. The platform will discriminate whether the learner has completed the course, and if the output target is 1 then the course is completed, otherwise, the learning is not completed.

In the deep learning-based course recommendation method, Salampasis et al. [30] presented the use of Recurrent Neural Networks (RNN) to consider the similarity between the user's recently viewed items and precomputed items. Pradhan et al. [31] proposed a bidirectional Long- and Short-Term Memory model (LSTM) and an integrated framework of a recommender system based on an attention mechanism for recommending suitable academic papers. Li et al. [32] suggested a neural network model of a Bayes Personalized Ranking Network (BPRN), which learns paired course preferences based on each user's history of course registration. Current deep learning applications in the field of course recommendation focus on improving and enhancing the model, but usually consider only one kind of modal information without mining data features from the text, image, and video information of the course itself as the representation of course recommendation modeling.

Therefore, by obtaining the video, audio, text, and digital modal features of the course, this paper improves the accuracy of recommendation from the perspectives of course information and learners and increases the click rate of users. Keyframes are extracted from a series of video frames in the course video, and video modal features are extracted for each frame by a pre-trained ResNet-16 model; audio modal features are extracted in the course video by using Librosa; text modal features are extracted by Word2Vec in the course name and introduction text information; and finally, digital modal features are extracted by statistical learners' demographic information, explicit feedback, and implicit feedback data.

The course recommendation results will be based on learners' collections, favorites, and comments on the platform, and the multimodal course features will be integrated to improve the recommendation accuracy and learners' course click rate.

4. Experimental Settings

4.1. Datasets

The course data and learner data in this paper were obtained from a real online learning MOOC platform [33]. We mainly collect course data including video information, audio information, the course name, and introduction of the course. The duration of the course video was about 10 min. The learner data mainly included the learners' comments, collections, and thumbs on the course in the explicit feedback data; the behavioral data of the learners who only browsed the course content but did not operate were mainly recorded in the implicit feedback data. This dataset was collected from March 2020, with a total of 32,413 learners, 813 courses, 63,936 course-explicit feedback data, and 81,537 course-implicit feedback data.

In the experimental process, the data set was further divided into three disjoint sets in this paper, which were 80%, 10%, and 10% of the course data randomly selected as the training set, validation set, and test set, respectively. The validation set was used for model optimization. Finally, the final model performance was verified on the test set.

4.2. Evaluation Indicators

The Area Under the Curve (*AUC*) [34] represents the area under the ROC curve, and its practical meaning is to correctly predict the ratio of learner and course pairs to all learner and course pairs in the recommender system, that is, the probability that a positive sample ranks ahead of a negative sample based on its rating. Suppose the list of recommended courses contains a_0 positive samples and a_1 negative samples, where n positive samples are predicted to be larger than the negative samples; then, the *AUC* calculation process can be expressed as:

$$AUC = \frac{n}{a_0 \cdot a_1} \tag{1}$$

The Hit Radio (*HR*) [35] represents the hit rate, which is the proportion of learners who have K correct recommendations in the recommended course list. The denominator *GT* represents all test sets, and the numerator *NumberHits@K* represents the sum of the number of test sets in each learner's Top-N list. *HR@K* can be expressed as:

$$HR@K = \frac{NumberHits@K}{GT} \tag{2}$$

Recall that the Ref. [36] refers to how many of the positive examples in the recommended course were predicted correctly.

Normalized Discounted Cumulative Gain (*NDCG*) [37] means the normalized discounted cumulative gain, which is often used as an evaluation of the recommended ranking results. *NDCG* can be expressed as:

$$A_{NDCG_{\mathrm{u}}@K} = \frac{B_{DCG_u@K}}{B_{IDCG_u}} \quad (3)$$

$$C_{NDCG@K} = \frac{B_{NDCG_u@K}}{|u|} \quad (4)$$

$A_{NDCG_{\mathrm{u}}@K}$ represents the discounted cumulative of the user's true list; $B_{DCG_{\mathrm{u}}@K}$ represents the score of the user's true list, and $C_{NDCG@K}$ represents the average of each user.

4.3. Experimental Parameter Settings

We processed the experimental data by Python and built the course recommendation algorithm framework by PyTorch. In the initialized embedding layer mixed with hidden layer parameters, a Gaussian distribution was used to set randomly (mean value of 0 and standard deviation of 0.1). Relu was used as the activation function in the feature extraction of visual, audio, and text modalities, and the Sigmoid activation function was used in the fully connected layer. The three modalities were extracted by LSTM, and their hidden units were set to 300, 200, and 50, respectively. The loss function was "binary cross-entropy" and the evaluation was "accuracy". At last, the batch size was set to 12, the number of iterations is 200, and the learning rate was 0.001.

4.4. Experimental Results

4.4.1. Experimental Results in Different Recommendation Models

This section evaluates the impact of the feature extraction module on the results of the LSTM+ATTENTION network model using the LSTM and the Attention mechanism networks, and the combination of both. The LSTM model was used to extract multimodal information, which can mine the potential temporal details of the data but has disadvantages in parallel processing. When the Attention model is used to extract the modal information, it can adjust the weights dynamically according to the extracted information; however, it is unable to learn the temporal relationships in the sequence. The LSTM+ATTENTION model can combine the advantages of the two, mine the potential timing information of the data, and dynamically adjust the weight according to the extracted information. The above analysis is also demonstrated by the experimental results, which show that the LSTM+ATTENTION model can combine the advantageous parts of the LSTM and Attention mechanisms with improving the extraction of modal features. The LSTM+ATTENTION model proposed in this paper outperforms the baseline model in both the AUC and the four click-through evaluation metrics. The effects of different extraction methods on the model are shown in Table 2.

Table 2. Effects of different extraction methods on the model.

	LSTM	**ATTENTION**	**LSTM+ATTENTION**
AUC	75.93	78.07	79.89
HR@5	0.68	0.71	0.78
HR@10	1.53	1.54	1.62
HR@15	2.41	2.48	2.53
HR@20	3.03	3.14	3.21

Table 3 compares the experimental results of the baseline model comparison. From the table, it is shown that the model in this paper achieves better results in both recall rate and NDCG, which is significantly better than other methods. The main reason is that LSTM+ATTENTION can fully exploit the temporal features and important dimensional features in multimodality. Meanwhile, the improved performance of AUC further supports the effectiveness of the attention mechanism to effectively distinguish important features in each modality. As shown in Table 3, a comparison of the performance under different recommended methods is presented.

Table 3. Performance comparison of different methods.

Methods	k = 5		k = 10		k = 15		k = 20	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
LSTM	0.5145	0.4176	0.6723	0.4623	0.7321	0.4832	0.7541	0.5215
ATTENTIONN	0.5132	0.4259	0.6671	0.4672	0.7143	0.4875	0.7583	0.5145
LSTM+ATTENTION	0.5573	0.4531	0.6801	0.4881	0.7412	0.5102	0.7591	0.5257

4.4.2. Experimental Results in Different Course Modalities

To verify the effectiveness of the multimodal features proposed above, we compare and test the effects of different course feature combinations on the results. In this paper, the different course modal combinations are as follows: C1 (digital modality), C2 (digital modality and text modality), C3 (digital modality, text modality, and audio modality), C4 (digital modality, text modality, audio modality, and visual modality), and C5 (all modal information of the course). Figure 4 represents the comparison of information hit rates under different combinations of course modalities.

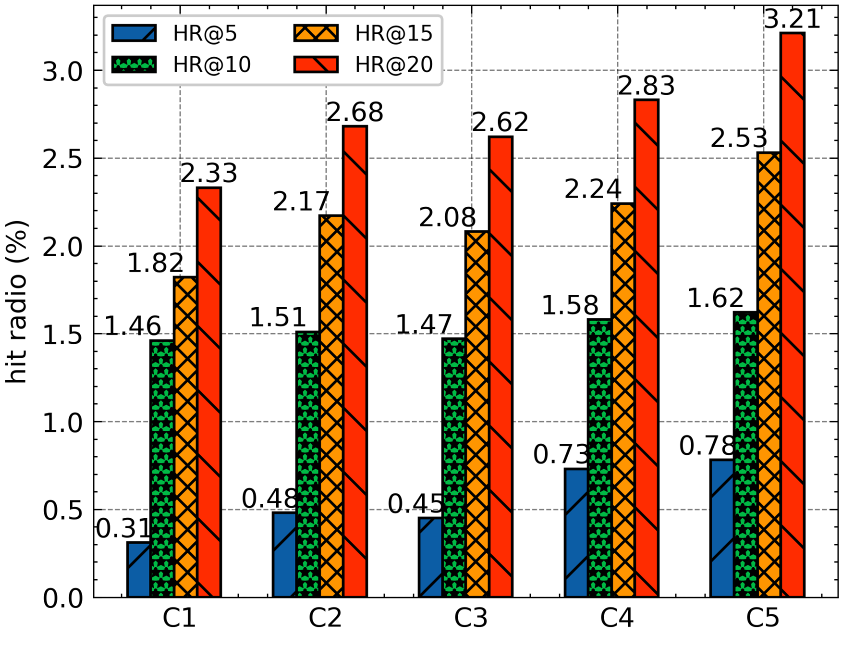

Figure 4. Information hit rate of different course modalities.

The experimental results show that the hit rate of C1 (numerical modality) is the lowest, which means that it represents the least information of the course. The highest hit rate of the user for the course is reached in C5 (all modal information of the course), which improves the hit rate by 0.47% compared with HR@5. In comparing the hit rate of C3 (digital modality, text modality, and audio modality) and C4 (digital modality, text modality, audio modality, and visual modality), it can be found that visual modality information of the course effectively improves the accuracy of the recommendation, which further indicates that the video modality contains more information of the course. Thus, it is shown that multimodal course information has a positive impact on recommendation accuracy.

4.4.3. Experimental Results with Different Learner Modal Information

In this section, we verify the impact of four information sets (L1, L2, L3, and L4) on the results, which represent demographic information (gender, age, major, etc.), learner-explicit feedback information, learner-implicit feedback information, and the sum information of all modal characteristics. These information sets mainly describe the users' operational behavior data and their attributes. Figure 5 depicts the hit rate of different learner modalities.

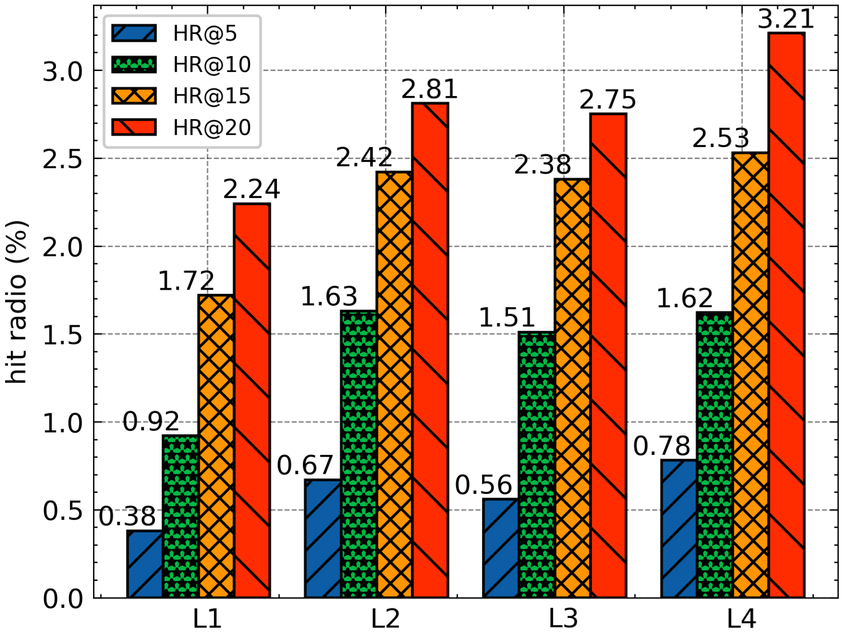

Figure 5. Hit radio of different learner modalities.

The relatively small amount of information represented by demographics leads to a low hit rate on courses. Adding learners' explicit feedback data (comments, collections, and thumbs) and implicit feedback data (browsing operations on the course) can effectively improve the user hit rate. L4 achieves better performance in the hit rate in all results, with hit rates of 2.24%, 2.81%, 2.75%, and 3.21% for four different K values, mainly because L4 incorporates all modalities of this paper, and can better obtain more modal information which is beneficial to improve the recommendation performance. The above experimental results show that the fusion of multiple modal user information effectively increases the course hit rate and can improve the course recommendation effect.

4.5. Discussion

When learners browse courses on the platform, they can click on the courses they are interested in through the course name and recommended courses, and they will comment, collect and thumb the courses they are interested in. Otherwise, there will be no operation behavior for the courses they are not interested in. The platform will record various operation behaviors of learners, and can collect explicit and implicit feedback data of learners to further explore users' viewing habits. When users watch the course video, the video modality, audio modality, course name, and introduction of the course can effectively build the user portrait. Therefore, by fusing the above modal information, we can effectively improve the performance of the course recommendation model, and LSTM and Attention can effectively explore the important contents in different modal information, as well as explore the temporal information between modalities to effectively improve the recommendation performance. The experimental results reveal that the framework of this paper can achieve better course recommendation results, in which the AUC score reaches 79.89% and the hit rate reaches 0.78%, 1.62%, 2.53%, and 3.21% at Top-N values from 5 to 20, respectively, indicating that the framework can effectively improve the accuracy of course recommendation.

In summary, the use of multimodal data in course recommendations can effectively improve the accuracy rate. The LSTM+ATTENTION model used in this paper is more effective in multimodal feature extraction compared with traditional neural networks, and the visual modality contains more effective information than the audio modality.

5. Conclusions and Future Work

With the development of the Internet and cloud computing, online education platforms have proliferated, and it has become especially important for learners to find suitable courses among the massive courses. Previous studies only consider unimodal features and build or improve deep learning models for a recommendation; however, less attention is being paid to multimodal features in courses. Moreover, most of the current course recommendations focus on digital modal features, and there is less research on video mode, audio mode, and text mode. When constructing user portraits, researchers pay more attention to the explicit feedback data of users and less capture the implicit feedback data of users in the operating platform. In addition, the effective combination of user demographic information, explicit feedback data, and implicit feedback data can improve the construction of the user model.

The research contributions of this paper are as follows: first, a deep learning framework for course recommendation based on multimodality is proposed, which can effectively improve the accuracy of course recommendation. Among them, the multimodal features involved in this paper mainly considered the video modality, audio modality, text modality, feedback, and implicit feedback of users and demographic information. Secondly, the LSTM+Attention model was constructed to build the course recommendation system modeling, and the temporal features and important dimensional features in the multimodal features were effectively mined. Finally, the validity of the model was verified by real data.

The practical significance of this paper lies in the following. The multimodal recommendation system based on deep learning realizes the course recommendation of an online education platform, effectively reducing the time for users to choose courses and realizing the Personalized Course recommendation service. In fully acquiring the modal features of users and the three modal features of courses, the in-depth learning framework of this paper effectively considers the integration of multiple modal features, so the recommended personalized course content can meet the needs of users. Because this paper uses the operation of users browsing the course to obtain the implicit feedback data, most of the implicit feedback data of users in daily life are difficult to capture, which needs to be improved in the next step of this paper.

In future work, based on obtaining learners' implicit feedback data, we will further delve into modeling learners' operations, such as the number of pauses, the number of plays, and the number of repeated views of videos while watching videos. Given that the degree of preference for the course is represented in the user's evaluation of the course, we will subsequently research information on the learner's evaluation of the course and conduct sentiment analysis on the learners' evaluation, modeling and verifying the relationship between learners' course preferences and ratings to improve the effectiveness of course recommendation.

Author Contributions: Data curation, X.R. and Y.Y.; Formal analysis, X.J.; Funding acquisition, W.Y; Project administration, G.J.; Software, X.R.; Visualization, Y.Y.; Writing—original draft, X.R.; Writing—review & editing, X.R., X.J. and W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Talent Project of College of Science and Technology (RC202016), the Natural Science Foundation of Zhejiang Province grant number LY20F020008, and the Ningbo Natural Science Foundation grant number 202003N4085.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Huang, R.; Lu, R. Research on Content-based MOOC Recommender Model. In Proceedings of the 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, China, 10–12 November 2018; pp. 676–681.
2. da S Dias, A.; Wives, L.K. Recommender system for learning objects based in the fusion of social signals, interests, and preferences of learner users in ubiquitous e-learning systems. *Pers. Ubiquitous Comput.* **2019**, *23*, 249–268. [CrossRef]
3. Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; Chua, T.S. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1437–1445.
4. Gong, J.; Wang, S.; Wang, J.; Feng, W.; Peng, H.; Tang, J.; Yu, P.S. Attentional graph convolutional networks for knowledge concept recommendation in moocs in a heterogeneous view. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, China, 25–30 July 2020; pp. 79–88.
5. Aljohani, T.; Cristea, A.I. Predicting Learners' Demographics Characteristics: Deep Learning Ensemble Architecture for Learners' Characteristics Prediction in MOOCs. In Proceedings of the 2019 4th International Conference on Information and Education Innovations, Durham, UK, 10–12 July 2019; pp. 23–27.
6. Hsu, C.M.; Efendi, H.; Junedi. Perspectives of Online Learners: Demographic Characteristics on Synchronous Learning Environment in Taiwan. In *International Conference on Educational Sciences and Teacher Profession (ICETeP 2020)*; Atlantis Press: Dordrecht, The Netherlands, 2021; pp. 261–267.
7. Khribi, M.K.; Jemni, M.; Nasraoui, O. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies, Santander, Spain, 1–5 July 2008; pp. 241–245.
8. Viniski, A.D.; Barddal, J.P.; de Souza Britto, A., Jr.; Enembreck, F.; de Campos, H.V.A. A case study of batch and incremental recommender systems in supermarket data under concept drifts and cold start. *Expert Syst. Appl.* **2021**, *176*, 114890. [CrossRef]
9. Sengottuvelan, P.; Gopalakrishnan, T.; Lokesh Kumar, R.; Kavya, M. A recommendation system for personal learning environments based on learner clicks. *Int. J. Appl. Eng. Res.* **2015**, *10*, 15316–15321.
10. Fu, D.; Liu, Q.; Zhang, S.; Wang, J. The undergraduate-oriented framework of MOOCs recommender system. In Proceedings of the 2015 International Symposium on Educational Technology (ISET), Wuhan, China, 27–29 July 2015; pp. 115–119.
11. Zhao, X.; Liu, B. Application of Personalized Recommendation Technology in MOOC System. In Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS), Vientiane, Laos, 11–12 January 2020; pp. 720–723.
12. Salehi, M.; Kamalabadi, I.N.; Ghoushchi, M.B.G. An effective recommendation framework for personal learning environments using a learner preference tree and a GA. *IEEE Trans. Learn. Technol.* **2013**, *6*, 350–363. [CrossRef]
13. Pang, Y.; Jin, Y.; Zhang, Y.; Zhu, T. Collaborative filtering recommendation for MOOC application. *Comput. Appl. Eng. Educ.* **2017**, *25*, 120–128. [CrossRef]
14. Burke, R. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact.* **2002**, *12*, 331–370. [CrossRef]
15. Jannach, D.; Zanker, M.; Felfernig, A.; Friedrich, G. *Recommender Systems: An Introduction*; Cambridge University Press: Cambridge, UK, 2010.
16. Zhou, Y.; Huang, C.; Hu, Q.; Zhu, J.; Tang, Y. Personalized learning full-path recommendation model based on LSTM neural networks. *Inf. Sci.* **2018**, *444*, 135–152. [CrossRef]
17. Ange, T.; Roger, N.; Aude, D.; Claude, F. Semi-supervised multimodal deep learning model for polarity detection in arguments. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
18. Wang, J.; Xie, H.; Au, O.T.S.; Zou, D.; Wang, F.L. Attention-based CNN for personalized course recommendations for MOOC learners. In Proceedings of the 2020 International Symposium on Educational Technology (ISET), Bangkok, Thailand, 21–23 July 2020; pp. 180–184.
19. Zhu, Q. Network Course Recommendation System Based on Double-Layer Attention Mechanism. *Sci. Program.* **2021**, *2021*, 7613511. [CrossRef]
20. Liu, J.; Yin, C.; Li, Y.; Sun, H.; Zhou, H. Deep Learning and Collaborative Filtering-Based Methods for Students' Performance Prediction and Course Recommendation. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2157343. [CrossRef]
21. Ni, J.; Huang, Z.; Hu, Y.; Lin, C. A two-stage embedding model for recommendation with multimodal auxiliary information. *Inf. Sci.* **2022**, *582*, 22–37. [CrossRef]
22. Pan, Y.; Mei, T.; Yao, T.; Li, H.; Rui, Y. Jointly modeling embedding and translation to bridge video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4594–4602.

23. Tamura, K.; Lu, M.; Konomi, S.; Hatano, K.; Inaba, M.; Oi, M.; Okamoto, T.; Okubo, F.; Shimada, A.; Wang, J.; et al. Integrating multimodal learning analytics and inclusive learning support systems for people of all ages. In Proceedings of the International Conference on Human-Computer Interaction, Orlando, FL, USA, 26–31 July 2019; pp. 469–481.
24. Wang, L.; He, Y. Online Learning Engagement Assessment Based on Multimodal Behavioral Data. In *Transactions on Edutainment XVI*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 256–265.
25. Xu, W.; Zhou, Y. Course video recommendation with multimodal information in online learning platforms: A deep learning framework. *Br. J. Educ. Technol.* **2020**, *51*, 1734–1747. [CrossRef]
26. Chango, W.; Cerezo, R.; Romero, C. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Comput. Electr. Eng.* **2021**, *89*, 106908. [CrossRef]
27. Mahasseni, B.; Lam, M.; Todorovic, S. Unsupervised video summarization with adversarial lstm networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 202–211.
28. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical reasoning on chinese morphological and semantic relations. *arXiv* **2018**, arXiv:1805.06504.
29. Cao, D.; Miao, L.; Rong, H.; Qin, Z.; Nie, L. Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities. *Knowl.-Based Syst.* **2020**, *203*, 106114. [CrossRef]
30. Salampasis, M.; Siomos, T.; Katsalis, A.; Diamantaras, K.; Christantonis, K.; Delianidi, M.; Karaveli, I. Comparison of RNN and Embeddings Methods for Next-item and Last-basket Session-based Recommendations. In Proceedings of the 2021 13th International Conference on Machine Learning and Computing, Shenzhen, China, 26 February–1 March 2021; pp. 477–484.
31. Pradhan, T.; Kumar, P.; Pal, S. CLAVER: An integrated framework of convolutional layer, bidirectional LSTM with attention mechanism based scholarly venue recommendation. *Inf. Sci.* **2021**, *559*, 212–235. [CrossRef]
32. Li, X.; Li, X.; Tang, J.; Wang, T.; Zhang, Y.; Chen, H. Improving Deep Item-Based Collaborative Filtering with Bayesian Personalized Ranking for MOOC Course Recommendation. International Conference on Knowledge Science, Engineering and Management, Hangzhou, China, 28–30 August 2020; pp. 247–258.
33. Sakboonyarat, S.; Tantatsanawong, P. Massive Open Online Courses (MOOCs) Recommendation Modeling using Deep Learning. In Proceedings of the 2019 23rd International Computer Science and Engineering Conference (ICSEC), Phuket, Thailand, 30 October–1 November 2019; pp. 275–280.
34. Volk, N.A.; Rojas, G.; Vitali, M.V. UniNet: Next Term Course Recommendation using Deep Learning. In Proceedings of the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 17–18 October 2020; pp. 377–380.
35. Lu, W.; Yu, Y.; Chang, Y.; Wang, Z.; Li, C.; Yuan, B. A dual input-aware factorization machine for CTR prediction. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Virtual Reality, 19–26 August 2021; pp. 3139–3145.
36. Liu, J.; Zhang, H.; Liu, Z. Research on Online Learning Resource Recommendation Method Based on Wide Deep and Elmo Model. *J. Phys. Conf. Ser.* **2020**, *1437*, 012015. [CrossRef]
37. Trirat, P.; Noree, S.; Yi, M.Y. IntelliMOOC: Intelligent Online Learning Framework for MOOC Platforms. In Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020), Virtual Conference, 10–13 July 2020; pp. 682–685.