Dynamic Node Graph Neural Network for Multimodal Music Recommendation

Ganghua Bai[1], Tianping Zhang[2]*

[1] School of Economics and Management, Hebi Polytechnic, China
[2] School of Mathematics and Computer Science, Hanjiang Normal University, China
gangh0322@163.com, ztp_2004@163.com

Abstract

Music recommendation systems are becoming increasingly popular among users. With the explosive growth of songs on the web, most music streaming platforms have launched online music listening services, providing millions of music choices for users. However, how to accurately recommend songs for users that match their preferences has become a challenging challenge, which we call music cold-start matching. In this paper, we delve into the multimodal information of music and take full advantage of the unique strengths of graph neural networks in capturing the collaborative filtering relationship between users and music. However, due to the inherent characteristics of graph neural networks, it is difficult to easily add new nodes to perform subsequent tasks in the inference phase. Therefore, we creatively propose a novel neural network architecture, the dynamic node graph neural network. In the training phase, we adopt a knowledge distillation strategy, using the graph neural network as the teacher model and the dynamic node graph neural network as the student model, thus enabling the student model to comprehensively learn and master the collaborative filtering relationship between users and music. In the inference phase, we use the trained dynamic node graph neural network to match new music accurately. After extensive experimental validation on the MSD public dataset, our approach demonstrates excellent effectiveness and efficiency, bringing users a more accurate music recommendation experience. Experimental results show that the state-of-the-art method improves our model by an average of 11.8% on the complete dataset and 7.1% on the cold-start problem compared to the best method, proving the effectiveness of our model.

Keywords: Multimodal, Music recommendation, Graph neural network, Recommendation system

1 Introduction

Personalized recommendation systems have become a powerful tool for users to find relevant information in the vast amount of Internet content [1], especially in the music domain [2], where personalized recommendation systems play a crucial role in discovering potentially great tracks and accurately delivering them to the right listeners. However, when faced with cold-start songs that lack interactive information, how to effectively incorporate them into a user's playlist is a challenge.

Content-based recommendation is a well-solved approach that recommends content with similar attributes by identifying feature similarities of items. For example, Bogdanov et al. [3] decomposed audio into 62 semantic descriptors as a way to extract deep features of music, while Chen et al. [4] clustered by extracting the Mel Frequency Cepstrum Coefficients (MFCCs) to accurately portray music characteristics. Considering that each song contains rich multimodal information, such as images, text, and audio, the reasonable utilization of such information can undoubtedly alleviate the cold start problem to a certain extent. However, content-based recommendation methods, while unique, neglect the collaborative filtering (CF) relationship between users and music. Such relationships can be largely presented through graph structures, and graph neural networks (GNNs) show great potential in capturing non-linear and non-trivial user-item relationships and can easily incorporate multiple data sources. For example, Gori et al. [5] used a PageRank-like algorithm for recommendation ranking; Kabbur et al. [6] effectively dealt with the problem of sparse datasets by decomposing the item-item similarity matrix into a product of low-rank latent factor matrices; and Xie et al. [7] proposed the SCCF method by combining local and global information in similar users, which significantly improves the recommendation performance.

GNNs recommendations still appear to be impotent when they face cold-start songs that lack interaction information. In order to remedy this shortcoming, researchers have started to explore new approaches. Jain et al. [8-9] attempted to combine recurrent neural networks with GNNs and introduced self-encoders to deal with dynamic graph structures [10]. In addition, Yan et al. [11] even cleverly transformed temporal relationships into temporal connections and proposed spatiotemporal graph neural networks, which provided a new idea for the solution to the cold-start problem.

We know that mining similarities between music and exploring their collaborative filtering relationships with users is the key to achieving accurate recommendations on songs without interaction information. Therefore, we

*Corresponding Author: Tianping Zhang; E-mail: ztp_2004@163.com
DOI: https://doi.org/10.70003/160792642025052603013

propose a novel model, Dynamic Node Graph Neural Network for Multimodal Music Recommendation (DNGNN), which aims to fully extract the multimodal information of music and deeply mine the multimodal CF relationship between users and music. Our main contributions include:

- Aiming at the limitations of existing approaches to the music cold-start matching problem, we innovatively propose a new model that leverages knowledge distillation to learn the collaborative filtering capability of graph neural networks. The model also can join dynamic nodes, thus effectively alleviating the music cold-start matching problem.
- To make full use of the semantic information in the feature fusion process, we carefully designed a multimodal feature fuser, which takes into account both the semantic divide of multimodal features and the complementarity between multiple modalities.
- We employ collaborative filtering-based feature fusion loss to facilitate knowledge transfer from the teacher model to the student model, which further enhances recommendation accuracy.
- We conduct exhaustive experimental validation on the real-world music dataset MSD, and the results fully demonstrate the superiority of our approach in solving the music cold-start problem.

2 Related Work

2.1 Multimodal Recommendation

Various types of auxiliary information [12-13], such as attributes [14-15], comments [16], and images [17-18], play an increasingly important role. However, existing models tend to process this information separately rather than fusing it when utilizing it. While multimodal features can provide complementary information to each other, effectively combining different forms of information remains a challenge. To this end, researchers have proposed several innovative approaches such as JRL [19-20] which uses deep neural networks to extract user and item features from multiple information sources and connect them to form a final representation. Recommendation models based on multimodal graph neural networks combine the advantages of graph neural networks and multimodal data to effectively model and exploit the relationships between multiple types of data. By representing data of different modalities as graph structures and jointly modeling them using graph neural networks, these models enable interaction and fusion between modalities. For example, GCN-PHR [21] attempted to introduce multimodal information into a user-item graph and extracted the features of the nodes through a graph neural network algorithm, which improved the accuracy of recommendations. In addition, multimodal graph neural network recommendation models also introduce advanced techniques such as graph enhancement techniques and knowledge graphs to further improve the accuracy of recommendations. MMGCL [22] introduces two graph enhancement techniques, by randomly deleting some edges between modalities and masking part of the information in the modalities, these models can efficiently learn correlations between modalities. Meanwhile, introducing the knowledge graph as a separate graph into the model can also provide richer semantic and structural information, which can help to better understand and predict the user's behaviors. MKGAT [23] introduces the knowledge graph into a multimodal graph neural network recommendation system for the first time, and proposes a multimodal graph attention mechanism to model multimodal knowledge graphs; while MMKGV [24] proposes a graph attention network for message propagation over knowledge graphs.

2.2 Knowledge Distillation

Knowledge distillation has made significant progress in the field of graph neural networks, and numerous studies have attempted to apply this technique to the field. The Graph Markov Neural Network (GMNN) [25] is one of the leaders, which cleverly uses the EM algorithm optimization framework to iteratively optimize two graph convolutional networks, which can be regarded as a teacher and a student, respectively. Sun et al. [26], on the other hand, integrate multiple graph convolutional neural networks with the same structure by training them to act as a student network, and, by integrating them, these students outperform the teacher network. TinyGNN [27], on the other hand, transfers the knowledge from a deep GNN to a smaller GNN through knowledge distillation, allowing it to gain a strong inferred node representation in a short period while maintaining the capture of local neighborhood information. Yang [28] et al. by designing a specific student model and combining it with a knowledge distillation framework, succeeded in making this model outperform traditional graph neural networks. SGDD [29], on the other hand, distill knowledge from GNNs to multilayer perceptual machines (MLPs), enabling MLPs to exhibit robust performance without relying on graph topology.

3 Problem Formulation

The recommendation task studied in this paper focuses on accurately predicting the probability of a target user's preference for a song in a context where that user is known. Based on this probability, we will rank the uninteracted songs in descending order and thus generate a Top-N song recommendation list. In this study, let $U = \{u_1, u_2, \ldots, u_N\}$ represent the user, $M = \{m_1, m_2, \ldots, m_P\}$ represent the song, and *lyr* and *fre* denote the lyrics and audio part of the song, respectively. For the processing of the lyrics features, we use the pre-trained language model BERT [30] for encoding and selecting the transformed representation of the last layer of [CLS] markers as the representation of the lyrics, denoted as $e_{lyr} \in R^d$. As for the audio features, we extract them using YAMNet to generate the representation of the audio $e_{fre} \in R^d$.

4 Our Method

4.1 Multimodal Recommendation

BERT is a powerful pre-trained language model that captures deep contextual information in text. BERT is based on the Transformer architecture and learns word-word relationships through the self-attention mechanism to generate word embeddings that contain rich semantic information. YAMNet is a pre-trained audio classification model that extracts a variety of features in audio. YAMNet can identify different sound events in audio, such as instrumental sounds, human voices, environmental sounds, etc., and generate corresponding feature representations. By using YAMNet to extract audio features, we can obtain an audio representation that matches the lyrics representation, which facilitates subsequent analysis and fusion. The overall framework diagram is shown in Figure 1. After obtaining the lyrics embedding and audio embedding codes respectively, this paper uses the cross-coding method of dual-stream structure, i.e., LXMERT [31], to perform the feature extraction of the cross as a way to achieve modal interaction.

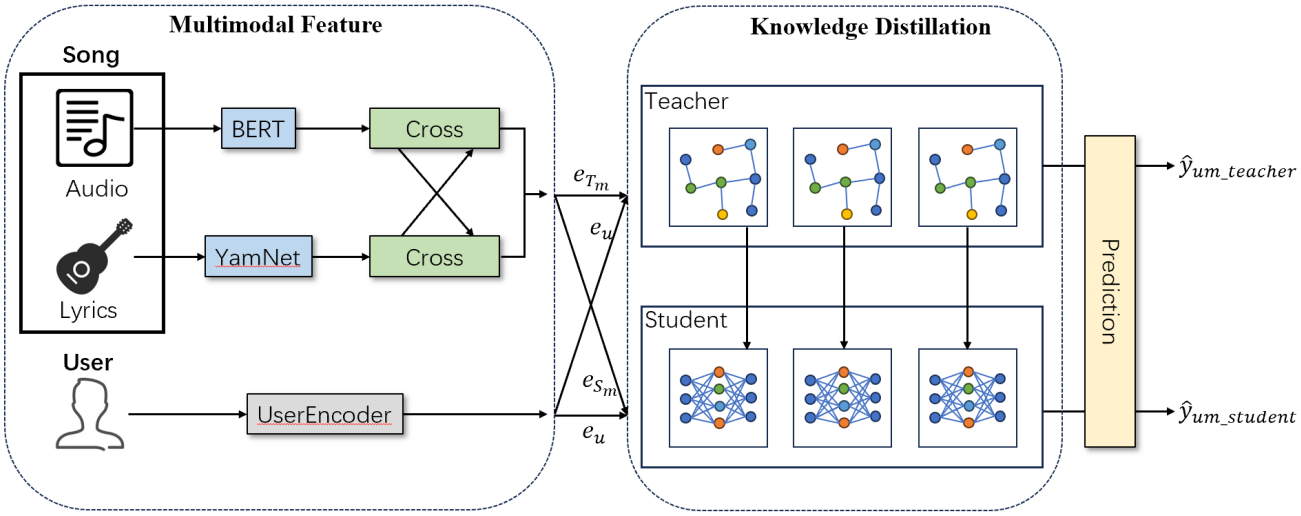

Figure 1. The overall architecture of the DNGNN model

(It consists of two parts: multimodal characterization and knowledge distillation, and each step is explained in detail in the following subsections.)

4.1.1 Multimodal Feature Extraction

We encode the lyrics by a pre-trained language model BERT and use the transformed representation labeled in the last layer [CLS] as the lyrics representation. At the same time, we use YAMNet to extract the audio representation. We fine-tune BERT and YAMNet during the training phase. The relevant formulas are as follows:

$$e_{lyr} = BERT(lyr) \tag{1}$$

$$e_{fre} = YAMNet(fre) \tag{2}$$

4.1.2 Multimodal Feature Fusion

After obtaining the lyrics feature vectors and audio feature vectors, to effectively fuse the lyrics feature vectors and audio feature vectors, there exist two mainstream interaction methods, single-stream and dual-stream. Generally speaking, the dual-stream structure performs better in feature learning and can capture the correlation between different features more comprehensively. Therefore, in this paper, we mainly rely on the LXMERT model with the dual-stream structure for cross-feature learning, which is an excellent way to achieve the interactive learning of lyrics and audio features. Specifically, the related calculation formula is described as follows:

$$e_m = \sum_l e_{lyr}^l \oplus e_{fre}^l \tag{3}$$

$$e_{lyr}^l = CrossAtt\left(e_{lyr}^{l-1}, \left(e_{fre1}^{l-1}, e_{fre2}^{l-1}, \ldots, e_{fren}^{l-1}\right)\right) \tag{4}$$

$$e_{fre}^l = CrossAtt\left(e_{fre}^{l-1}, \left(e_{lyr1}^{l-1}, e_{lyr2}^{l-1}, \ldots, e_{lyrn}^{l-1}\right)\right) \tag{5}$$

where l is the number of layers in the LXMERT model, through the interactive attention mechanism between lyrics and audio, we can learn implicit interaction features that are reflected in the lyrics and audio features after the interaction. This learning of implicit features allows the model to better understand the multidimensional information of the song.

4.2 Knowledge Distillation based on GNNs

4.2.1 Teacher Model

We adopt 4-layer LightGCN [32] as the teacher model with encoded user features and song multimodal features as input. After extensive experiments, it is shown that

4-layer LightGCN can achieve the best performance. The relevant formulas are as follows:

$$e_u^{(l+1)} = \sum_{T_m \in \mathcal{N}_u} AGG\left(e_u^{(l)}, e_{T_m}^{(l)}\right) \tag{6}$$

$$e_{T_m}^{(l+1)} = \sum_{u \in \mathcal{N}_{T_m}} AGG\left(e_{T_m}^{(l)}, e_u^{(l)}\right) \tag{7}$$

where l denotes the number of LightGCN layers, *AGG* denotes the aggregation function, e_u denotes the user feature vector, and e_{T_m} denotes the song feature vector in the teacher model.

4.2.2 Student Model

In order to extend the generalization ability of the model, we designed a new network as the student model and adopted a 4-layer structure to make the student model fit the structure of the teacher model more closely. The relevant formulas are as follows:

$$e_{S_m}^{(l+1)} = e_{S_m}^{(l)} \cdot W + b \tag{8}$$

where W denotes the weight matrix and b denotes the bias coefficients, when l is 0, $e_m = e_{T_m}^{(l)} = e_{S_m}^{(l)}$.

4.3 Prediction

To preserve the semantic information of each hop, we used an average weighting to get the final user features and song features, and the relevant equations are shown below:

$$e_u = \sum_{l=0}^{L} \frac{1}{L+1} e_u^{(l)} \tag{9}$$

$$e_{T_m} = \sum_{l=0}^{L} \frac{1}{L+1} e_{T_m}^{(l)} \tag{10}$$

$$e_{S_m} = \sum_{l=0}^{L} \frac{1}{L+1} e_{S_m}^{(l)} \tag{11}$$

Model predictions are defined as the inner product of the final representations of users and items:

$$\hat{y}_{um\_train} = e_u^T \cdot e_{T_m} \tag{12}$$

$$\hat{y}_{um\_test} = e_u^T \cdot e_{S_m} \tag{13}$$

In the training phase, we use $\hat{y}_{um\_train}$ as the prediction score and participate in the loss calculation. In the inference phase $\hat{y}_{um\_test}$ is used as the final prediction score.

We use Bayesian Personalized Ranking (BPR) loss [33], as the predictive loss for LightGCN, with the loss function shown below:

$$\mathcal{L}_{BPR} = -\sum_{u}^{U} \sum_{i \in \mathcal{N}_u} \sum_{j \notin \mathcal{N}_u} ln\sigma\left(\hat{y}_{ui_{train}} - y_{uj_{train}}\right) \tag{14}$$

where $\sigma()$ denotes the sigmoid function.

To make the student model learn better and make the student model simulate the similarity between the samples in the teacher model, the student model can fully exploit the structured feature information between the samples in the teacher network, and realize to extract generic, rich, and sufficient knowledge from the teacher model to guide the student model. The relevant formula is shown below:

$$\mathcal{L}_G = \sum_{l}^{L} \sum_{(e,'e) \in e} MSE\left(D_G\left(e_{S_m}^l, {}'e_{S_m}^l\right), D_G\left(e_{T_m}^l, {}'e_{T_m}^l\right)\right) \tag{15}$$

where *MSE*() denotes the distance metric function that represents the constructed graph in minimizing students and teachers, and D_G denotes the similarity constructor between nodes.

For the Dynamic Node Graph Neural Network, we propose an objective function defined as follows:

$$\mathcal{L}(\Theta) = \mathcal{L}_{BPR} + \mathcal{L}_G + \lambda \left\|\Theta\right\|_2^2 \tag{16}$$

where λ denotes the parameter of L2 regularisation to prevent overfitting, and we jointly implement the optimization of the Dynamic Node Graph Neural Network by combining the BPR loss, the distillation loss, and the L2 regularization.

5 Experiments

5.1 Dataset and Metrics

We used the MSD-A dataset, as shown in Table 1 which is related to the Million Song Dataset (MSD). We used 7 to 30-second audio previews retrieved from 7digital.com. After removing ambiguous artists and missing tracks, the final dataset consists of 328,821 tracks by 24,043 artists, each with at least 15 seconds of audio and 50 characters in length, with 5.2 M interactions. To validate the cold-start problem, we retained 10% of the songs for testing. We used Recall@50 and NDCG@50 as evaluation metrics and evaluated the ranking results on the entire set of songs.

Table 1. Dataset details

Dataset	# songs	# interactions
MSD-A	328K	5.2M

5.2 Baseline

To verify the effectiveness of our method, we will compare it with the following method:

- Dropoutnet [34] imposes random feature dropping on the input to satisfy the condition of missing preference patterns.
- Global Orthogonal Regularization (GOR) [35] applies global orthogonal regularization to the input to maximize the "unfolding" feature in the descriptor space;
- Correlated Feature Masking (CFM) [36] applies correlated feature masking to the input to learn better potential relationships between item features.
- Bootstrapping Contrastive Learning (BCL) [37] improves the quality of learned representations by applying contrast regularization.

5.3 Implementation Details

The method described in this paper is implemented using PyTorch and deployed on an NVIDIA Titan XP GPU with 24G of memory. To ensure a fair comparison, we adhere to the optimal parameters or publicly available source code provided by the baseline model as a basis for obtaining experimental results. In optimizing the objective function, we employ the stochastic gradient descent approach. Given the remarkable efficiency demonstrated by Adaptive Moment Estimation (Adam) in non-convex optimization problems, we utilize the Adam optimizer for model optimization and parameter updates. To determine the optimal hyperparameter values, we primarily rely on a grid search strategy, setting the batch size to 1024 and the learning rate to 0.001. the features obtained from BERT and YAMNet are set to be 300 dimensional for ease of feature interaction. Finally, we select Recall@50 and NDCG@50 as evaluation metrics to comprehensively assess the ranking results across the entire song dataset.

Table 2. Overall results for different models

(They are trained on the cold start and full training data, choosing 10% of the songs as a cold start problem. The best results are shown in bold.)

	MSA-A		Cold-start	
Method	Recall@50	NDCG@50	Recall@50	NDCG@50
Dropoutnet	0.0481	0.0281	0.0629	0.0389
GOR	0.0504	0.0299	0.0652	0.0421
CFM	0.0551	0.0313	0.0693	0.0430
BCL	0.0619	0.0351	0.0734	0.0475
DNGNN	**0.0682**	**0.0398**	**0.0785**	**0.0510**

5.4 Performance Comparison

The performance comparison results are shown in Table 2. Recall@50 and NDCG@50 are used as evaluation metrics. Our DNGNN model achieves the best performance on both the MSA-A complete dataset and the cold-start problem. Compared to the baseline model BCL, our model improves on average by 11.8% on the full dataset and 7.1% on the cold-start problem, proving the effectiveness of our model. Meanwhile, we performed cluster analysis on songs and users, and the song clustering results are shown in Figure 2, and the correlation analysis between user clusters and song clusters is shown in Figure 3. This significant improvement can be attributed to three reasons:

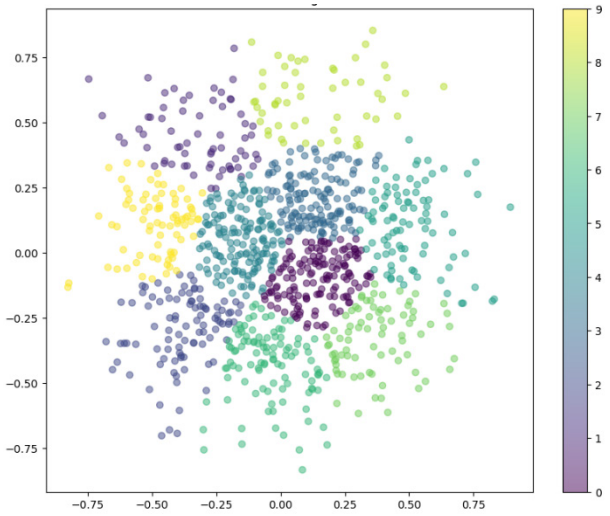

Figure 2. Song clustering analysis

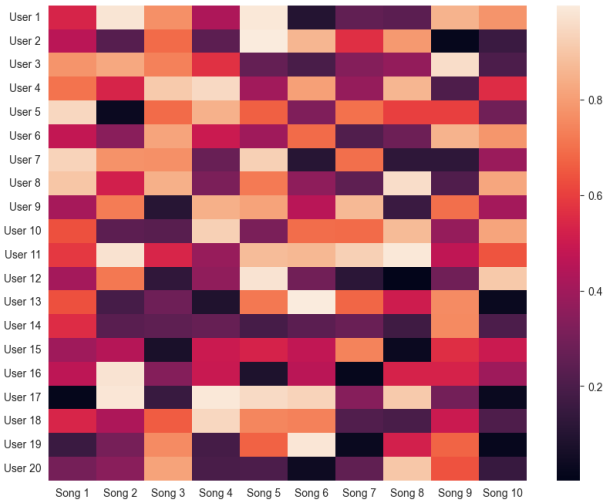

Figure 3. User-song heat map

Multimodal information fusion: music, as a multimodal information carrier, contains multiple dimensions such as melody, rhythm, lyrics, and emotion. Traditional recommendation methods often only consider a single dimension, such as user behavior data or labels of music, which makes it difficult to fully capture the characteristics of music. Dynamic node graph neural networks can fully fuse the multimodal information of music, and more accurately capture the collaborative filtering relationship between users and music by constructing a complex relationship graph between users and music. This fusion of multimodal information enables the model to understand the user's preferences more comprehensively, thus improving the accuracy of recommendations.

Dynamic Node Processing: Traditional graph neural networks have difficulty in easily adding new nodes to perform subsequent tasks during the inference phase, which limits their application in music recommendation systems. Dynamic node graph neural networks can solve this problem. By adopting a knowledge distillation

strategy in the training phase, the dynamic node graph neural network can comprehensively learn and master the collaborative filtering relationship between users and music. In the inference phase, the model can easily add new music nodes and perform fast and accurate analysis and matching of new nodes, which makes the model more flexible to cope with continuously updated music libraries and ensures the timeliness of recommendations.

Efficient and accurate recommendation: the dynamic node graph neural network model continuously optimizes its node representation and relationship learning capabilities during training, enabling it to accurately capture the complex relationship between users and music. By combining the user's historical behavior, the music's attributes, and their interactions, the model can generate high-quality recommendations. In addition, the model has high computational efficiency and can process a large amount of user and music data in a short period to meet the real-time requirements of music recommendation systems.

5.5 Ablation Experiment

We conducted ablation experiments to assess the contribution of different components of our proposed DNGNN model. Our goal was to understand the impact of key design choices and to examine factors that contribute to model improvement. We explored the impact of the multimodal cross-fusion module as well as the teacher module on the DNGNN, and the results of the ablation experiments are shown in Table 3, (*w/o Cro*) denotes the removal of the multimodal cross-fertilization module, the component focuses on the fusion of features from different modalities such as lyrics and audio in the recommending process, by ablating the module this study aims to understand the importance of integrating information from multiple modalities. (*w/o Dis*) denotes removal removes the teacher module, which is responsible for providing guidance and knowledge transfer to the student model during the knowledge distillation process ablating the teacher module helps to assess the impact of knowledge distillation on the learning process of the student model, by removing this module, it is possible to analyze to what extent the distillation of the teacher model contributes to improving the performance of the student model. (*w/o C&D*) denotes the removal of both. For all ablation setups, we replicated the experimental setup and evaluation metrics used in the original experiments. Based on the results of our ablation experiments, we observed the following:

Table 3. Results of ablation experiment

	MSA-A		Cold-start	
Method	Recall@50	NDCG@50	Recall@50	NDCG@50
w/o Cro	0.0532	0.0310	0.0600	0.0388
w/o Dis	0.0496	0.0290	0.0550	0.0360
w/o C & D	0.0441	0.0212	0.0484	0.0324
DNGNN	**0.0682**	**0.0398**	**0.0785**	**0.0510**

Firstly, the removal of the multimodal cross-fusion module showed a significant decrease in the performance of the DNGNN model. This result fully demonstrates the important role of the multimodal cross-fusion module in the model. By fusing information from different modalities, the model can understand the input data more comprehensively, thus improving the accuracy of the prediction.

Second, we examined the effect of the teacher module on the model performance. In the ablation experiment, we remove the teacher module and observe the performance change of the student model. The experimental results show that the performance of the model also decreases after removing the teacher module. This suggests that the teacher module plays a key role in the knowledge distillation process and helps to improve the performance of the student model by transferring knowledge and experience from the teacher model.

Finally, we also compared the simultaneous removal of the multimodal cross-fusion module and the teacher module. The experimental results show that the most significant decrease in model performance is found in this dual absence scenario. This finding further reinforces the centrality of the multimodal cross-fusion module and the teacher module in the DNGNN model.

Through the ablation experiments, we gained insights into the impact of the multimodal cross-fusion module and the teacher module in the DNGNN model on the model performance. These results provide important guidance for us to further optimize the model. In future work, we will continue to explore other possible modules and design choices to further improve the performance of the DNGNN model.

6 Conclusion

In this paper, we addressed the challenging task of multimodal music recommendation, particularly focusing on the cold-start problem. Leveraging the unique strengths of graph neural networks in capturing collaborative filtering relationships between users and music, we proposed a novel neural network architecture known as the Dynamic Node Graph Neural Network (DNGNN). This architecture overcame the limitations of traditional graph neural networks in incorporating new nodes during the inference phase.

Our approach fused knowledge distillation, using a GNN teacher to impart knowledge to the DNGNN student. The student model mastered intricate user-music relationships. During inference, DNGNN accurately matched music to preferences, refining recommendation accuracy.

MSA-A dataset tests proved our method's superior efficacy and efficiency, addressing music cold-start issues with tailored recommendations. Future graph neural network-based multimodal music recommendation research will emphasize: 1) integrating more modalities beyond lyrics and audio, 2) adapting to evolving user preferences and trends, 3) developing interpretable models for transparency, and 4) addressing scalability and

efficiency for large-scale systems. Goals include enhancing personalization and music listening experience.

References

[1] M. Kaur, S. Rani, Recommender System: Towards Identification of Shilling Attacks in Rating System Using Machine Learning Algorithms, *International Journal of Performability Engineering*, Vol. 19, No. 7, pp. 443-451, July, 2023.

[2] S. P. Deore, SongRec: A Facial Expression Recognition System for Song Recommendation using CNN, *International Journal of Performability Engineering*, Vol. 19, No. 2, pp. 115-121, February, 2023.

[3] D. Bogdanov, M. Haro, F. Fuhrmann, A. Xambó, E. Gómez, P. Herrera, Semantic audio content-based music recommendation and visualization based on user preference examples, *Information Processing & Management*, Vol. 49, No. 1, pp 13-33, January, 2013.

[4] R. Chen, B. Tang, A music recommendation system based on acoustic features and user personalities, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Auckland, New Zealand, 2016, pp. 203-213.

[5] M. Gori, A. Pucci, ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines, *International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007, pp. 2766-2771.

[6] S. Kabbur, X. Ning, G. Karypis, FISM: factored item similarity models for top-N recommender systems, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago, Illinois, USA, 2013, pp. 659-667.

[7] X. Xie, F. Sun, X. Yang, Z. Yang, Y. Xiaoyong, Y. Zhao, G. Jinyang, O. Wenwu, C. Bin, Explore user neighborhood for real-time e-commerce recommendation, *International Conference on Data Engineering*, Chania, Greece, 2021, pp. 2464-2475.

[8] A. Jain, A. R. Zamir, S. Savarese, A. Saxena, Structural-RNN: Deep Learning on Spatio-Temporal Graphs, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 5308-5317.

[9] M. Cao, V. O. K. Li, V. W. S. Chan, A CNN-LSTM Model for Traffic Speed Prediction, *2020 IEEE 91st Vehicular Technology Conference*, Antwerp, Belgium, 2020, pp. 1-5.

[10] A. García-Durán, S. Dumančić, M. Niepert, Learning sequence encoders for temporal knowledge graph completion, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 4816-4821.

[11] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 7444-7452.

[12] J. Chen, F. Zhuang, X. Hong, X. Ao, X. Xie, Q. He, Attention-driven Factor Model for Explainable Personalized Recommendation, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor, MI, USA, 2018, pp. 909-912.

[13] C. Du, F. Yu, M. Jiang, A. Hua, X. Wei, T. Peng, X. Hu, Vton-scfa: A virtual try-on network based on the semantic constraints and flow alignment, *IEEE Transactions on Multimedia*, Vol. 25, pp. 777-791, February, 2023.

[14] L. Fan, Z. Cheng, L. Zhu, C. Liu, L. Nie, An Attribute-Aware Attentive GCN Model for Attribute Missing in Recommendation, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 34, No. 9, pp. 4077-4088, September, 2022.

[15] J. McAuley, J. Leskovec, Hidden factors and hidden topics: understanding rating dimensions with review text, *Proceedings of the 7th ACM conference on Recommender systems*, Hong Kong, China, 2013, pp. 165-172.

[16] Y. Tan, M. Zhang, Y. Liu, S. Ma, Rating-boosted latent topics: Understanding users and items with ratings and reviews, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, NY, USA, 2016, pp. 2640-2646.

[17] J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel, Image-based recommendations on styles and substitutes, *the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Santiago, Chile, 2015, pp. 43-52.

[18] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, H. Liu, What your images reveal: Exploiting visual contents for point-of-interest recommendation, *the 26th International Conference on World Wide Web*, Perth, Australia, 2017, pp. 391-400.

[19] F. Yu, A. Hua, C. Du, M. Jiang, X. Wei, T. Peng, L. Xu, X. Hu, VTON-MP: Multi-Pose Virtual Try-On via Appearance Flow and Feature Filtering, *IEEE Transactions on Consumer Electronics*, Vol. 69, No. 4, pp. 1101-1113, November, 2023.

[20] Y. Zhang, Q. Ai, X. Chen, W. B. Croft, Joint representation learning for top-n recommendation with heterogeneous information sources, *the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, Singapore, 2017, pp. 1449-1458.

[21] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, L. Nie, Personalized Hashtag Recommendation for Micro-videos, *the 27th ACM International Conference on Multimedia*, Nice, France, 2019, pp. 1446-1454.

[22] Z. Yi, X. Wang, I. Ounis, C. Macdonald, Multi-modal graph contrastive learning for micro-video recommendation, *the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, 2022, pp. 1807-1811.

[23] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Virtual Event, Ireland, 2020, pp. 1405-1414.

[24] H. Liu, C. Li, L. Tian, Multi-modal graph attention network for video recommendation, *IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET)*, Beijing, China, 2022, pp. 94-99.

[25] M. Qu, Y. Bengio, J. Tang, Gmnn: Graph markov neural networks, *International conference on machine learning*, Suzhou, Jiangsu, China, 2019, pp. 5241-5250.

[26] Y. Sun, J. Han, Mining heterogeneous information networks: a structural analysis approach, *Acm Sigkdd Explorations Newsletter*, Vol. 14, No, 2, pp. 1931-0145, April, 2013.

[27] B. Yan, C. Wang, G. Guo, Y. Lou, Tinygnn: Learning efficient graph neural networks, *the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, NY, USA, 2021, pp. 1848-1856.

[28] C. Yang, J. Liu, C. Shi, Extract the knowledge of graph

neural networks and go beyond it: An effective knowledge distillation framework, *WWW'21: Proceedings of the Web Conference 2021*, Ljubljana, Slovenia, 2021, pp. 1227-1237.

[29] B. Yang, K. Wang, Q. Sun, C. Ji, X. Fu, H. Tang, Y. You, J. Li, Does graph distillation see like vision dataset counterpart?, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2024, pp. 53201-53226.

[30] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[31] H. Tan, M. Bansal, LXMERT: Learning Cross-Modality Encoder Representations from Transformers, *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 5099-5110.

[32] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *ACM SIGIR: the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Xi'an, China, 2020, pp. 639-648.

[33] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian Personalized Ranking from Implicit Feedback, *UAI: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, Canada, 2009, pp. 452-461.

[34] M. Volkovs, G. Yu, T. Poutanen, Dropoutnet: Addressing cold start in recommender systems, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, 2017, pp. 4964-4973.

[35] X. Zhang, F. X. Yu, S. Kumar, S.-F. Chang, Learning spread-out local feature descriptors, *ICCV: IEEE International Conference on Computer Vision*, Venice, Italy, 2017, pp. 4595-4603.

[36] T. Yao, X. Yi, D. Z. Cheng, F. Yu, T. Chen, A. Menon, L. Hong, Ed H. Chi, S. Tjoa, J. Kang, Evan Ettinger, Self-supervised learning for large-scale item recommendations, *CIKM: Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Queensland, Australia, 2021, pp. 4321-4330.

[37] X. Zhao, Y. Zhang, Q. Xiao, Y. Ren, Y. Yang, Bootstrapping Contrastive Learning Enhanced Music Cold-Start Matching, *WWW: Companion Proceedings of the ACM Web Conference*, 2023, Austin, TX, USA, pp. 351-355.

Biographies

Ganghua Bai works at the School of Economics and Management, Hebi Polytechnic. His research interests include Computer Science and Technology.

Tianping Zhang works at the School of Mathematics and Computer Science, Hanjiang Normal University. Her research interests include artificial intelligence, cloud computing and Internet of Things.