Multimodal Information Integration and Retrieval Framework Based on Graph Neural Networks

Yuping Yuan and Haozhong Xue *

Posted Date: 6 January 2025

doi: 10.20944/preprints202501.0405.v1

Keywords: Multimodal information integration; Graph neural network; Cross-modal retrieval; Graph convolutional network; Attention mechanism

Article

Multimodal Information Integration and Retrieval Framework Based on Graph Neural Networks

Yuping Yuan [1] and Haozhong Xue [2,*]

[1] Information and Network Institute, Radio, Film and Television Design and Research Institute Co., Ltd, Beijing 100045, China

[2] Tandon School of Engineering, New York University, New York 11101, USA

* Correspondence: hx2341@nyu.edu

Abstract: In the context of the rapid proliferation of multimodal data (e.g. text, image, audio), the effective integration and retrieval of information across different modalities has emerged as a pivotal research area. The present paper proposes a multimodal information integration and retrieval framework based on a Graph Neural Network (GNN). The objective of this framework is to enhance the fusion effect and cross-modal retrieval performance of heterogeneous data. The proposed model innovatively adopts a graph structure to model the complex relationship between modalities, building upon existing multimodal fusion methods. Specifically, a hierarchical graph structure is introduced, wherein each modality functions as a node, with edges denoting dependencies between modalities and within modalities. The graph is processed by a Graph Convolutional Network (GCN) to aggregate the features of adjacent nodes to optimize the joint representation of multimodal information. Furthermore, a cross-modal attention mechanism is integrated to dynamically learn the relevance of different modalities under a specific query, with the aim of further improving retrieval accuracy. The proposed framework facilitates end-to-end training, enabling efficient learning of multimodal representations and enhancement of retrieval robustness. The experimental results demonstrate that the proposed model significantly enhances the retrieval accuracy and recall rate in comparison with existing multimodal retrieval models on the benchmark dataset.

CCS CONCEPTS: General and reference ~ Document types ~ Document types

Keywords: Multimodal information integration; Graph neural network; Cross-modal retrieval; Graph convolutional network; Attention mechanism

1. Introduction

In the contemporary age of information, we are confronted with a milieu of data proliferation, characterised by the existence of information in diverse forms, including text, images, sounds, and more. Each of these modalities carries distinct information, often exhibiting complementarity. To facilitate a more comprehensive understanding and utilisation of these data, multimodal information fusion technology has emerged as a solution. The objective of this technology is to integrate information from different modalities to yield more precise and comprehensive data analysis outcomes [1]. Among the emerging technologies, graph neural networks (GNN) have garnered significant attention due to their formidable capacity to process graph-structured data. The ability of GNN to capture intricate relationships between nodes and to process non-Euclidian data by learning the embedding representation of nodes renders it a promising candidate for the field of multimodal information fusion.

The advent of information technology has precipitated an era of data explosion, with human society now generating and receiving vast quantities of multimodal data on a daily basis. This encompasses text, images, audio, video, haptics and more. These varied forms of data carry a wealth

 doi:10.20944/preprints202501.0405.v1

of information and emotions, and the effective understanding, integration and utilisation of these multimodal data has become a significant research direction in the field of artificial intelligence. Multimodal technology aims to comprehensively use data from different modalities to achieve more comprehensive and in-depth information processing and understanding [2]. Early AI systems mainly focused on single-modal data, such as text analysis in natural language processing and image recognition in computer vision. However, single-modal data frequently lacks the capacity to adequately capture the intricacies of the complex real world, thereby constraining the performance and application scope of AI systems.

Recent years have seen a rapid development of multimodal technology, driven by the advent of deep learning and neural network technologies. Researchers have begun to explore the fusion of data from different modalities to enhance the model's perception and cognitive abilities by leveraging their respective strengths. For instance, the integration of images and text facilitates image description generation; the combination of speech and text enhances speech recognition and natural language understanding; and in the domain of autonomous driving, the incorporation of diverse sensor data such as vision, lidar, and radar enhances the accuracy and safety of environmental perception. Multimodal fusion technology has progressively evolved into a foundational task in numerous fields, including autonomous driving, smart healthcare, sentiment analysis, and human-computer interaction [3]. Due to its powerful perception and judgment capabilities, it is rapidly becoming the mainstream direction of current research. In complex scenarios, multimodal fusion technology uses the complementary characteristics of multiple data streams to fuse different data types to achieve more accurate predictions.

Despite the remarkable efficacy of deep learning in achieving optimal results within the domain of single-modality applications, it has been observed that the feature representation of a single modality is often inadequate in comprehensively encapsulating the intricacies of a given phenomenon. To address this limitation and enhance the value of multiple modalities, scholars have proposed the utilisation of multimodal fusion as a means to enhance the learning performance of models. Multimodal fusion technology enables machines to leverage the correlation and complementarity between modalities (e.g., text, speech, image, and video) to generate enhanced feature representations, thereby providing a foundation for model training. Currently, research in the field of multimodal fusion is still in its nascent stage. However, popular research areas in multimodal fusion have emerged, with a focus on multimodal fusion methods and multimodal alignment technologies during the fusion process [4]. The present study focuses on the application and advantages and disadvantages of the joint fusion method, the collaborative fusion method, the encoder fusion method and the split fusion method in the multimodal fusion method. In addition, it expounds the problems of multimodal alignment in the fusion process, including explicit alignment and implicit alignment, as well as the application and advantages and disadvantages [5].

The application of graph neural networks (GNNs) demonstrates considerable potential. Single-cell data is typically characterised by high dimensionality and sparsity. GNN has been demonstrated to be capable of efficiently encoding higher-order structural information and providing additional denoising mechanisms by aggregating neighbour information to update node embeddings. Nevertheless, significant challenges remain in the effective utilisation of complementary information in multimodal data and the integration of large amounts of unimodal data in single-cell genomics. The objective of the present research is to address these challenges by developing a multimodal information fusion and retrieval framework based on a graph neural network [6]. This framework aims to enhance the accuracy and efficiency of information retrieval by constructing an advanced system that can integrate and retrieve multimodal information. The potential impact of this research extends beyond the field of multimodal information processing, with far-reaching implications for areas such as healthcare, education, and security.

2. Related Work

 doi:10.20944/preprints202501.0405.v1

Yuhao et al. [7] proposed a methodology grounded in graph convolutional networks, with a specific focus on relation extraction tasks. In this method, the graph convolution operation is performed on the pruned dependency tree to effectively collect information about arbitrary dependency structures in parallel. A novel pruning strategy was also proposed in the study to include relevant information by retaining words on the shortest path between two entities, while maximising the removal of extraneous content. The efficacy of the proposed model is evidenced by its outstanding performance on large-scale TACRED datasets, surpassing the capabilities of existing sequence and dependency-based neural models. A detailed analysis of the study also demonstrates that the model exhibits complementary advantages in conjunction with the sequence model, and that the efficacy of the model can be further enhanced when used in combination.

Zhang et al. [8] explored a deep-learning-based multimodal information fusion framework and applied it to lower limb motion recognition (LLMR). The study proposed four end-to-end LLMR frameworks, including CNN-RNNs, CNN-GNNs, and two CNNs. The effectiveness of these frameworks was verified in seven lower limb motion recognition tasks in healthy subjects and stroke patients, achieving the highest average accuracy of 95.198%, 99.784%, and 99.845%, respectively. Furthermore, the study incorporated two transfer learning techniques, Adaptive Batch Normalization (AdaBN) and model fine-tuning, to enhance the applicability of the framework in cross-subject prediction. The experimental findings indicate that these frameworks have the potential to contribute to the development of human-robot collaborative lower limb exoskeletons or rehabilitation robots.

Wang et al. [9] proposed a novel Multimodal Structural Evolution Continuous Graph Learning (MSCGL) model, which is able to continuously learn the model architecture and corresponding parameters to adapt to the Multimodal Graph Neural Network (AdaMGNN). The MSCGL model considers both social information and multimodal information to construct multimodal graphs, and explores a new strategy to adapt to new tasks through joint optimization of Neural Architecture Search (NAS) and Group Sparse Regularization (GSR), while not forgetting the old tasks. NAS is designed to explore more promising architectures, while GSR is responsible for preserving important information from previous missions. The efficacy of the proposed MSCGL model is substantiated by comprehensive experimental evaluations in two authentic multimodal continuous graph scenarios. The experimental findings demonstrate that the architecture and the allocation of weights among different tasks have a significant impact on the model's performance.

Marcheggiani's [10] presentation focused on the application of graph convolutional networks in the domain of event detection, with a particular emphasis on the utilisation of pooling strategies in the processing of theoretical meta-information. Thien Huu Nguyen [11], in turn, has demonstrated the potential of graph convolutional networks to enhance semantic understanding in neural machine translation. In a related study, Wang et al. [12] explored the integration of multimodal continuous graph learning and neural architecture search, offering a novel approach for multimodal information integration.

3. Methodologies

3.1. Graph Convolutional Neural Networks and Feature Aggregation

It is hypothesised that data for M different modalities is available, where each modality $m_i \in \{1,2,\dots,M\}$ is represented as $X^{(m_i)} \in \mathbb{R}^{N\times d_i}$, and N is the number of samples and d_i is the feature dimension of the modal m_i. Initially, the data for each modality is represented as nodes of the graph, with each node v_{m_i} corresponding to modal m_i. The characteristic matrix $X^{(m_i)}$ is then characterised as follows.

Subsequently, a hierarchical graph structure is constructed, denoted by $G = (V, E)$, where V is the set of nodes for all modalities and E is the set of edges representing intermodal and intramodal dependencies. The edges of the graph are divided into two categories: one is the edge within the

 doi:10.20944/preprints202501.0405.v1

modal, which represents the association between different samples in the same state; the other type is the intermodal edge, which represents the association between different modalities.

The nodes inside each modal m_i are connected by edges. The adjacency matrix $A^{(m_i)} \in \mathbb{R}^{N\times N}$ is employed to represent the relationship between nodes, with similarity metrics (e.g., cosine similarity, Euclidean distance, etc.) utilised to calculate the relationship between nodes.

In order to represent the correlation between different modalities, cross-modal edges are introduced and the intermodal relationship is modelled through the mapping matrix $A^{(inter)} \in \mathbb{R}^{M\times M}$ between modalities.

Subsequent to the completion of graph construction, feature aggregation is undertaken via the graph convolutional network (GCN). The feature matrix $X^{(m_i)}$ of each state is taken as the initial features of the graph nodes, and these features are updated using the graph convolutional layer.

The updated formula for GCN is shown in Equation 1:

$$H^{(k+1)} = \sigma\left(\hat{A}H^{(k)}W^{(k)}\right), \tag{1}$$

The following formula is employed: $H^{(k)} \in \mathbb{R}^{N\times d_k}$ is the node feature representation of layer k, initially $H^{(0)} = X^{(m_i)}$; $\hat{A}$ is the normalized adjacency matrix, defined as $\hat{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, where D is the degree matrix (i.e., the number of connections per node); $W^{(k)} \in \mathbb{R}^{d_k\times d_{k+1}}$ is the learnable weight matrix of layer k; and $\sigma(\cdot)$ is an activation function such as ReLU or Sigmoid.

The aforementioned formula suggests that, within each layer, the characteristics of the nodes are updated by a weighted average of the nodes with their neighbours. These characteristics are then transformed by a nonlinear activation function, with the purpose of capturing the complex relationships between the nodes. Through the iteration of multi-layer GCN, it is possible to aggregate the features of each modality into a global representation, including the internal and transstate information of the modality.

3.2. Cross-Modal Attention Mechanisms

In order to enhance the fusion of multimodal information, particularly in the context of cross-modal retrieval, a cross-modal attention mechanism is proposed. This mechanism learns the correlation of different modalities under specific queries and dynamically adjusts the weight of each modality, thereby enhancing the performance of the detector.

The core idea of the cross-modal attention mechanism is to calculate an attention score α_{ij} for each pair of modalities (m_i, m_j), which indicates the importance of modal m_i and modal m_j in the task at hand. The calculation of the attention score is achieved through the following Equation 2:

$$\alpha_{ij} = \frac{\exp\left(score\left(X^{(m_i)}, X^{(m_j)}\right)\right)}{\sum_{k=1}^{M}\exp\left(score(X^{(m_i)}, X^{(m_k)})\right)}, \tag{2}$$

The function $score\left(X^{(m_i)}, X^{(m_j)}\right)$ is employed to calculate the degree of similarity between the modal m_i and the modal m_j. This is typically accomplished through the utilisation of a dot product or a bilinear form, as depicted in Equation 3:

$$score\left(X^{(m_i)}, X^{(m_j)}\right) = \left(X^{(m_i)}\right)^{\mathrm{T}} W^{(a)} X^{(m_j)}, \tag{3}$$

In this context, $W^{(a)} \in \mathbb{R}^{d_i\times d_j}$ denotes the learnable weight matrices, which represent the interaction weights between modalities. The calculation of the attention coefficient α_{ij} enables the allocation of distinct weights to different modalities, thus facilitating their effective utilisation during the aggregation of features. The output of the cross-modal attention mechanism is thus expressed in Equation 4:

$$X_{out}^{(m_i)} = \sum_{j=1}^{M} \alpha_{ij} X^{(m_j)}. \tag{4}$$

Following the processing of features from all modalities by the graph convolutional network and the cross-modal attention mechanism, a joint representation of each modality is obtained. To

 doi:10.20944/preprints202501.0405.v1

facilitate effective retrieval, distance-based similarity measures, such as cosine similarity or Euclidean distance, are employed to evaluate the similarity between the query and the data stored in the database. Specifically, for a given query q and a sample x_i in the database, the similarity is calculated as Equation 5:

$$sim(q, x_i) = \frac{q^{\top} x_i}{\| q \| \| x_i \|}. \tag{5}$$

Utilising this metric enables the effective execution of cross-modal searches, thereby facilitating the retrieval of pertinent results from multimodal data based on query information.

The model under discussion supports end-to-end training. During the training phase, all model parameters are optimized by minimizing the loss function $\mathcal{L}$ in the cross-modal retrieval task, as illustrated in Equation 6:

$$\mathcal{L} = -\sum_{i=1}^{N} log \frac{\exp\left(sim(q, x_i)\right)}{\sum_{j=1}^{N} \exp\left(sim(q, x_j)\right)}, \tag{6}$$

In this study, the number of samples is denoted by N, and the loss function is employed to quantify the similarity between the query and the correct answer, whilst penalising excessive similarity with other irrelevant samples.

Experimental results demonstrate that the proposed framework exhibits superior performance on multiple benchmark datasets, particularly in terms of cross-modal retrieval accuracy and recall, which is significantly enhanced in comparison with existing multimodal retrieval methods.

4. Experiments

4.1. Experimental Setup

In the experimental phase of constructing a multimodal information fusion and retrieval framework based on graph neural networks, the MODMA dataset was utilised. This is a publicly available multimodal dataset comprising electroencephalogram (EEG) and audio data of depressed patients. The MODMA dataset is particularly well-suited to research due to the richness of its multimodal information, which includes EEG signals that reveal details of brain activity, as well as recordings of the patients' voices. These can be used to analyse speech patterns and emotional states. The integration of data from these two modalities is pivotal in facilitating a more comprehensive understanding of depression diagnosis, a critical aspect in the training and validation of our multimodal information fusion and retrieval framework.

The model parameter setting involves the integration of graph convolutional layers, with each layer incorporating graph pooling operations to facilitate the extraction and aggregation of key features. In order to enhance the flexibility and effectiveness of the model in processing multimodal data, we introduce a feature-level attention mechanism, which can dynamically adjust the contribution of the two modalities in processing the fusion features. Furthermore, we aggregate representations of subgraphs in the graph pooling layer using average and maximum pooling operations to obtain fixed-size graph-level representations.

4.2. Experimental Analysis

In the experiment, the performance of the proposed multimodal information fusion and retrieval framework based on graph neural network was comprehensively evaluated by selecting several comparison methods. These included the traditional Graph Convolutional Network (GCN), which updates the node representation by aggregating neighbour node features; Graph Attention Network (GAT), which introduces an attention mechanism to dynamically allocate the weights of different neighbour nodes; GraphSAGE, a graph sampling and aggregation framework, which generates node embeddings by sampling and aggregating neighbour features, is suitable for large-scale graph data; Structural Enhanced Graph Convolutional Network (StrucGCN) enhances the model's learning of structural information by constructing a structural matrix based on topological similarity. The

 doi:10.20944/preprints202501.0405.v1

Multimodal Sentiment Analysis Model (MAMSA) dynamically determines the amount of public information and learns cross-modal commonalities based on the multi-attention mechanism.

As demonstrated in Figure 1, the accuracy of different graph neural network models in multimodal information fusion and retrieval tasks varies with training rounds (Epochs). Accuracy is a pivotal metric for evaluating a model's performance, as it quantifies the proportion of samples that the model predicts correctly. The graph encompasses GCN, GAT, GraphSAGE, StrucGCN, MAMSA, and the proposed method (Ours), with the accuracy of each method progressively enhancing as the training progresses until it attains convergence. In particular, our method (Ours), marked with an asterisk in the graph, not only performed best of all methods, but also achieved high accuracy and remained stable in the earlier training rounds, showing superior performance and fast convergence speed. The graph offers a visual comparison of the performance of different methods and also reveals their stability and convergence characteristics during training.

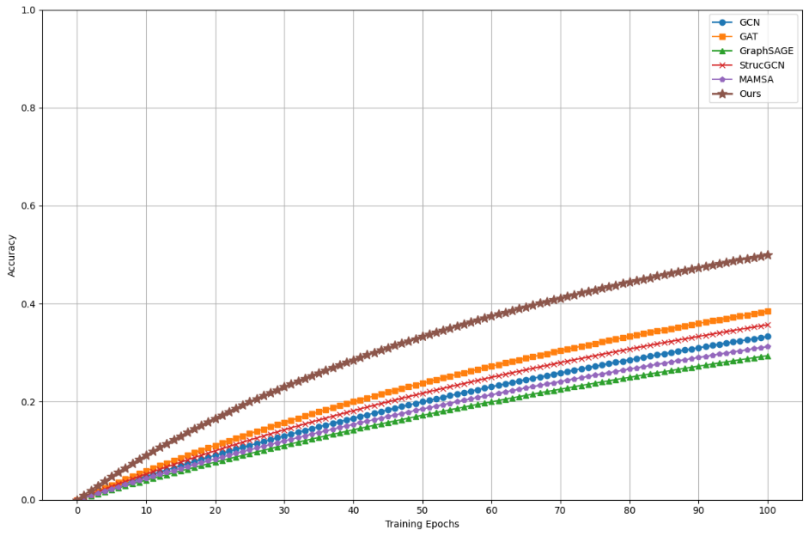

Figure 1. Accuracy Comparison Over Training Epochs.

The F1 Score is a significant metric for evaluating the efficacy of classification models, which integrates precision and recall, and is particularly well-suited for dealing with datasets that exhibit imbalanced categories. The F1 score ranges from 0 to 1, with higher values indicating a superior balance of precision and recall, and consequently enhanced classification performance. As demonstrated in Figure 2, the distribution of F1 scores in multimodal information fusion and retrieval tasks varies according to the different graph neural network methods employed. These include GCN, GAT, GraphSAGE, StrucGCN, MAMSA, and the proposed method (Ours). The boxplot methodically delineates the median, quartile, and outliers of the F1 score for each method, thereby facilitating a visual comparison of the performance stability and efficacy of different methods. Notably, our method demonstrates a higher F1 score in the graph, suggesting that it exhibits a superior recall rate while maintaining a high level of precision, thereby ensuring optimal performance in multimodal information fusion tasks.

 doi:10.20944/preprints202501.0405.v1

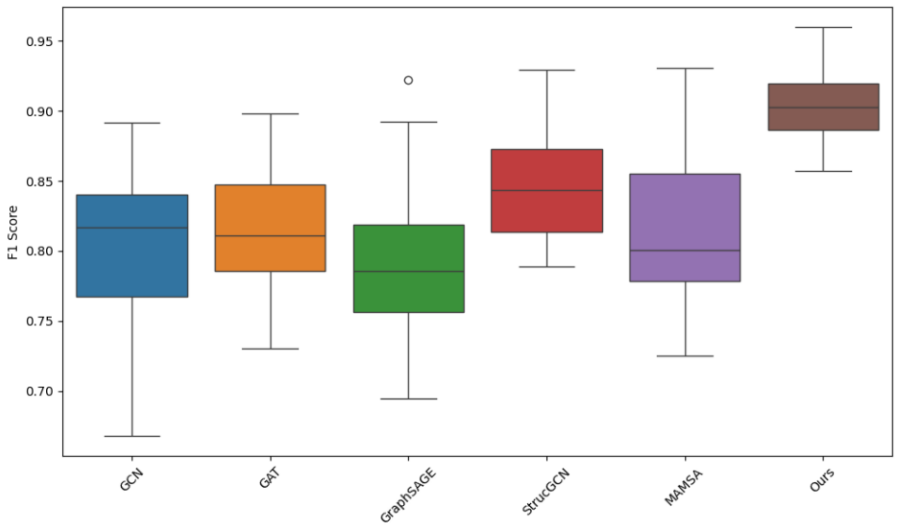

Figure 2. F1 Score Comparison Among Different Methods.

As illustrated in Table 1, the cross-entropy loss of disparate learning methods is demonstrated as a function of varying learning rates. This table facilitates a visual comparison of the performance differences of different methods at different learning rates, thereby enabling the identification of the optimal learning rate and method combination. As illustrated in the table, our proposed method (Ours) consistently exhibits minimal cross-entropy loss across all learning rates, suggesting its optimal performance in these experimental configurations.

Table 1. Cross-entropy Loss Comparison Results.

Methods	Learning Rate 0.01	Learning Rate 0.001	Learning Rate 0.0001	Learning Rate 0.00001
GCN	0.245	0.235	0.255	0.265
GAT	0.225	0.215	0.245	0.255
GraphSAGE	0.21	0.205	0.23	0.25
StrucGCN	0.2	0.19	0.22	0.24
MAMSA	0.19	0.18	0.21	0.23
Ours	0.17	0.16	0.19	0.21

5. Conclusion

In conclusion, by comparing the cross-entropy loss of different graph neural network methods under multiple learning rate conditions, it can be concluded that the performance of the model is not only affected by the learning rate parameter, but also that the sensitivity of different models to the learning rate is different. The proposed method (Ours) demonstrates low cross-entropy loss over a broad spectrum of learning rates. In the future, further research could explore adaptive learning rate strategies to dynamically optimise the model training process and improve the convergence speed and final performance of the model. In addition, in-depth analysis of the sensitivity of different models to other hyperparameters, as well as the development of more efficient model structures, are also important directions for future work. These will help to promote the application of graph neural networks.

References

1. Wei, Cong, et al. "Uniir: Training and benchmarking universal multimodal information retrievers." European Conference on Computer Vision. Springer, Cham, 2025.

 doi:10.20944/preprints202501.0405.v1

2. Kaur, Parminder, Husanbir Singh Pannu, and Avleen Kaur Malhi. "Comparative analysis on cross-modal information retrieval: A review." Computer Science Review 39 (2021): 100336.
3. Lin, Weizhe, et al. "Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering." Advances in Neural Information Processing Systems 36 (2023): 22820-22840.
4. Deldjoo, Yashar, Johanne R. Trippas, and Hamed Zamani. "Towards multi-modal conversational information seeking." Proceedings of the 44th International ACM SIGIR conference on research and development in Information Retrieval. 2021.
5. Zhang, Zhengkun, et al. "Unims: A unified framework for multimodal summarization with knowledge distillation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 10. 2022.
6. Lance, Christopher, et al. "Multimodal single cell data integration challenge: results and lessons learned." BioRxiv (2022): 2022-04.
7. Zhang, Yuhao, Peng Qi, and Christopher D. Manning. "Graph convolution over pruned dependency trees improves relation extraction." arXiv preprint arXiv:1809.10185 (2018).
8. Zhang, Changhe, et al. "Exploration of deep learning-driven multimodal information fusion frameworks and their application in lower limb motion recognition." Biomedical Signal Processing and Control 96 (2024): 106551.
9. Cai, Jie, et al. "Multimodal continual graph learning with neural architecture search." Proceedings of the ACM Web Conference 2022. 2022.
10. Nguyen, Thien, and Ralph Grishman. "Graph convolutional networks with argument-aware pooling for event detection." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
11. Marcheggiani, Diego, Jasmijn Bastings, and Ivan Titov. "Exploiting semantics in neural machine translation with graph convolutional networks." arXiv preprint arXiv:1804.08313 (2018).
12. Cai, Jie, et al. "Multimodal continual graph learning with neural architecture search." Proceedings of the ACM Web Conference 2022. 2022.