计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于多模态融合的图神经网络推荐算法研究
作者：吴志强，解庆，李琳，刘永坚
DOI：10.19678/j.issn.1000-3428.0066929
网络首发日期：2023-04-20
引用格式：吴志强，解庆，李琳，刘永坚．基于多模态融合的图神经网络推荐算法研究[J/OL]．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0066929

网络首发时间：2023-04-20 08:45:05
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.tp.20230419.1324.006.html

基于多模态融合的图神经网络推荐算法研究

吴志强[1,3]，解 庆[1,2,3]，李琳[1,2]，刘永坚[1,2,3]

(1. 计算机与人工智能学院，武汉理工大学，武汉 430070；2. 数字出版智能服务技术教育部工程研究中心，武汉 430070；
3. 武汉理工大学重庆研究院，重庆 401135)

摘 要：已有的 GNN 推荐算法大多利用用户-项目交互图的节点编号信息进行训练，学习用户-项目节点的高阶联系去丰富节点表示。但是忽略了用户对不同模态信息的偏好，没有利用项目的图片、文本等模态信息，或对于不同模态特征的融合简单相加，没有区分用户对不同模态信息的偏好。针对这一情况，提出多模态融合的 GNN 推荐模型。首先针对单个模态，结合用户-项目交互二部图构建单模态图网络，在单模态图中学习用户对此模态信息的偏好；利用 GAT 聚合邻居信息，丰富本节点表示；同时利用 GRU 决定是否聚合邻居信息，达到去噪效果；最后将各个模态图学习到的用户、项目表示通过注意力机制融合得到最终表示，然后送入预测模块。在 MovieLens-20M、H&M 两个数据集上的实验结果表明，多模态信息、注意力融合机制能有效提升推荐的准确度，算法模型在 Precision@K、Recall@K 和 NDCG@K 三个指标上相较于基线最优算法均有显著提升。评估指标 K 值选取 10 时，其指标 Precision@10 在两个数据集上分别提升 4.67%、2.42%，Recall@10 在两个数据集上分别提升 2.03%、2.49%，NDCG@10 在两个数据集上分别提升 5.24%、2.05%。

关键词：多模态推荐；多模态融合；注意力机制；图神经网络；推荐系统；门控图神经网络

开放科学（资源服务）标志码（OSID）：

Recommendation Algorithm Based on Multimodal Fusion Graph Neural Network

WU Zhiqiang[1,3]，XIE Qing[1,2,3]，LI Lin[1,2]，LIU Yongjian[1,2,3]

(1. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China;
2. Engineering Research Center of Intelligent Service Technology for Digital Publishing, Ministry of Education, Wuhan 430070, China;
3. Chongqing Research Institute of Wuhan University of Technology, Chongqing 401135, China)

【Abstract】Most of the existing GNN recommendation algorithms use the node number information of the user-item interaction graph for training, and learn the high-order connectivity among user and item nodes to enrich their representations. However, the user preferences for different modal information are ignored, and the modal information such as images and text of items are not utilized, and the fusion of different modal features is simply summed without distinguishing the user preferences for different modal information. To address this situation, a multimodal fusion GNN recommendation model is proposed. Firstly, for a single modality, a unimodal graph network is constructed by combining the user-item interaction bipartite graph, and the user's preference for this modal information is learned in the unimodal graph; GAT is used to aggregate the neighbor information and enrich the local node representation. At the same time, GRU is used to decide whether to aggregate the neighbor information or not to achieve the denoising effect. Finally, the user and item representations learned from each modal graph are fused by the attention mechanism to get the final representation, and then sent to the prediction module. The experimental results on two public datasets, which are MovieLens-20M and H&M, show that the multimodal information and attention fusion mechanism can effectively improve the accuracy of recommendation. The algorithm in Precision@K, Recall@K and NDCG@K metrics are significantly improved compared with the baseline optimal algorithm. When the K value of the evaluation index is 10, its Precision @10 is improved by 4.67% and 2.42% respectively, Recall@10 is improved by 2.03% and 2.49% respectively and NDCG@10 is improved by 5.24% and 2.05% respectively.

【Key words】multimodal recommendation; multimodal fusion; attention mechanism; graph neural network; recommendation system; gated graph neural network

DOI: 10.19678/j.issn.1000-3428.0066929

基金项目：国家自然科学基金（62276196），重庆市自然科学基金（cstc2021jcyj-msxmX1013），湖北省重点研发计划项目（2021BAA030）

作者简介：吴志强（1997-），男，硕士研究生，主研方向为推荐算法；解庆（通信作者），博士，副教授；李琳，博士，教授；刘永坚，学士，教授。**E-mail：**zhiqiang16@whut.edu.cn

0 概述

如今信息量快速增长，个性化推荐系统已经是大多数面向用户服务平台中的关键部分。推荐系统目标在于给用户提供可能感兴趣的信息，例如一些电影、书籍、商品等。目前有一系列的服务平台，如电商平台（亚马逊、淘宝和京东）、社交平台（微信、推特）、短视频平台（抖音、快手）。因此能够准确地预测用户对哪些信息感兴趣至关重要。

早期的推荐系统主要通过学习用户-项目历史交互记录进行预测。这些模型大部分由表示学习模块和预测模块组成：表示学习模块将用户-项目对以及其辅助信息转换为适当的向量表示；预测模块基于上一模块学习到的向量进行预测。传统的协同过滤推荐算法如矩阵分解算法(MF)[1]、相似性模型(FISM)[2]和 NAIS 模型[3]等，虽然从只考虑项目侧信息到考虑用户侧信息，再到加入注意力机制，效果都有提升，但是无法处理复杂的用户关系。为了解决这些问题，图神经网络（GNN）聚合用户-项目的高阶邻居信息可以增强表示学习的效果。相关算法模型将用户-项目交互记录关联起来组成一个二部图，在二部图中一阶关系表示真实的用户-项目交互记录，而更高阶的关联可以反映出用户行为的相似性，项目内容的相似性以及一些协同过滤信号。受到 GNN 信息传播的启发，推荐系统也采用类似的方法去聚合用户、项目的高阶邻居信息作为新的表示。例如文献[4]提出 NGCF 算法模型去编码协同过滤（CF）信号到向量表示去增强推荐效果。文献[5]提出 LightGCN 简化了 NGCF 的网络设计，得到更好的表示效果。

虽然上述方法都取得了很大的成功，但是这些工作没有充分地利用项目的多模态信息，例如抖音短视频有视频、音频等信息；淘宝商品有海报、文字描述等信息[6]。将多种模态信息和用户交互记录结合，可以更深层次的捕获用户的偏好：此外，同一项目的不同模态信息可以存在语义歧义，如两个不同主题电影的海报相似，但是通过文本描述可以区分；用户可能对不同模态信息偏好度不同，如选择商品时，相比较于文字描述用户更想通过图片去了解衣服视觉效果；最后通过将不同模态信息结合起来，可以从多方面去对用户偏好建模。

已存在多模态推荐算法模型主要将多模态信息作为一个整体并将他们融合到协同过滤框架中。具体采用的方法是将多模态特征训练整合到一个特征表示，再将这个表示附加到 CF 框架（如文献[7]提出的 MF 模型）中的用户和项目表示上。文献[8]提出的 VBPR 模型，利用项目的视觉特征去丰富其嵌入向量（利用 ID 生成）。文献[9]提出 ACF 模型，在多模态的基础上利用注意力机制在用户-项目历史交互图上聚合邻居的两跳信息。虽然这些模型融入了多模态信息，取得很好的效果，但对用户细粒度的模态偏好无法捕获。

为了解决这些问题，最近的工作主要将不同模态信息细化并引入 GNN 框架模型中。前提是将多模态信息加入到图中，构建多模态知识图谱，在利用 GNN 去传播和聚合邻居信息的同时，去捕获用户模态偏好。文献[10]提出 MKGAT 模型，在 KGAT 模型[11]的基础上引入多模态节点，将多模态特征作为项目的节点加入到知识图谱中，利用图注意力网络（GAT）[12]去聚合高阶邻居信息，得到更好的特征表示。但是这种构建整体多模态知识图谱可能引入噪声信息，导致推荐效果提升受限。同时如果构建文本-图像的多模态知识图谱进行训练需要为每一个项目节点增加文本和图片节点信息，这对信息要求较为苛刻。文献[13]提出 MGAT，针对单个模态构建用户-项目历史交互图，然后通过 GNN 学习到用户对单模态的偏好，最后融合到的最终偏好表示向量中，推荐效果进一步提升。但是其对不同模态特征的融合仅利用均值、拼接等方法，没有很好地学习用户的偏好。

基于上述分析，本文提出了一种基于轻注意力机制的多模态融合推荐算法。首先针对单个模态信息，通过用户-项目交互记录和模态预训练特征构建单模态图，再利用 GAT 和 GRU 学习图中高阶邻居信息，捕获局部语义信息，最后通过轻注意力模块融合各个模态得到的用户和项目特征，捕获用户对不同模态信息的细粒度偏好。

1 相关工作

1.1 特征融合

特征融合这个主题在很多方向上有广泛地研究，如语义视频分析、多视图图片分类以及 VQA（visual question answering）。目前在多模态推荐领域研究较少，因为多模态推荐任务中涉及各种模态信息：如视频，图片，文本，音频等信息，可以从文本-视频检索相关工作中借鉴经验。

对于视频方向的特征融合，早期工作常使用向量拼接来合并多个特征[14]。文献[15]探索基于学习

的方法，其中特征由七个不同的专家网络组成。文献[16]提出多模态 Transformers 模型去聚合不同视频帧的特征并组合不同特征空间的相似性。

对于文本端的特征融合，文献[17]提出W2VV++模型，采用早期融合方法，融合三种文本预训练模型输出特征，如词袋模型、word2vec、和GRU。相反，文献[18]提出 SEA 模型采用晚融合方法，先构造统一的文本特征空间，然后对单个空间学习的特征进行平均得到最终特征。

文献[19]提出 LAFF（轻注意力特征融合）联合使用特征转换层和注意力层，在模型早期和晚期都进行融合，有效地解决了文本-视频特征检索问题。

1.2 多模态推荐

个性化推荐系统在很多应用上取得很大的成功，如电商、短视频和新闻等平台。早期的多模态推荐算法主要基于 CF 模型[20]，CF 相关模型主要利用用户的显示反馈和隐式反馈去预测用户-项目之间的交互。即使 CF 模型取得了很好的效果，但是对于数据稀疏场景推荐效果有待提升。

最近，随着深度神经网络（DNN）在计算机视觉、自然语言处理和语音方向上发展，DNN 也被引入多模态领域。一些工作尝试用预训练模型提取多模态特征，再将这些特征融合到融合到 CF 模型中增强推荐效果，如文献[21]提出 CITING 模型，挖掘并融合文本特征，对社交媒体图像语义建模，进行图文推荐。文献[22]利用动态递归神经网络融合视频语义和用户兴趣对用户动态偏好进行建模。文献[10]提出 MKGAT 模型，在知识图谱上加入多模态节点构建协同知识图谱，利用 GAT 机制聚合高阶邻居信息，但是这样直接在整个图上加多模态节点并进行学习可能会引入噪声信息。文献[23]利用预训练模型和 GAT 机制学习模态表示并进行三模态融合，在数据稀疏下能很好地捕获用户兴趣偏好。不同上述方法，文献[24]提出 MMGCN 模型以及文献[13]提出 MGAT 模型，对单个模态构建图，利用 GAT 和 GRU 学习用户单模态偏好再进行融合，很好地解耦用户多模态偏好，更细粒度捕获用户兴趣。

1.3 图神经网络

由于 GCN 有效且便利，在最近的工作中被大量使用。文献[25]提出第一个基于 GCN 的模型 PinSAGE，通过采样和聚合邻居信息得到节点表示，得到很好的效果。目前大部分工作都是利用图神经网络进行学习，同时加上 GAT、GRU 等机制，更细粒度对用户偏好建模。文献[26]利用图神经网络处理图中复杂的结构信息，考虑到用户重复点击行为和引入 GAT 机制提升了序列推荐的效果。NGCF 模型在用户-项目交互二部图中将协同过滤信号加入向量嵌入过程，得到高阶交互特征表达建模。

基于上述相关工作的分析，现有基于图神经网络的多模态推荐算法没有很好地区分用户对模态的偏好。本文在图神经网络推荐中引入多模态信息，并且在最后特征融合阶段引入注意力机制以建模用户对不同模态的偏好，从而更细粒度地捕获用户兴趣，同时注意力模块可扩展性强，便于加入其它模态信息。

2 模型框架

2.1 公式化描述

在介绍模型之前，首先详细介绍文中涉及的一些基本概念和符号。

用户-项目二部图由用户和项目交互记录构成，其边代表用户和项目有交互，记用户-项目二部图为 $G=(V,E)$，G 是无向图，其中 $\mathrm{V}=U\bigcup I$ 代表节点集合，U 表示用户节点，I 表示项目节点，$\mathrm{E}=\{(u,i)\,|\,u\in U,i\in I\}$ 代表边集合，每条边表示一条用户-项目交互记录。

多模态交互图（见图 1）表示在用户-项目交互二部图 G 基础上，除了用户和项目的交互记录外，一个项目还有其他模态信息（如文本，图片等）。为每个模态，都单独构建一个交互图，用户与项目之间边表示交互记录，项目与多模态节点的边表示项目拥有这个模态信息，记多模态图集合为 $\{G_k\}$，$k\in\{1,2\}$ 表示图片和文本模态。

模型输入：用户-项目交互记录，图片、文本预训练特征。

模型输出：一个可以预测用户 u 和项目 i 交互可能性的推荐函数。

2.2 总体框架

我们提出一个新的多模态模型方法 MFGAT，算法模型框架如图 2 所示，MFGAT 主要由 5 个部分组成：（1）数据集模块表示在用户-项目历史交互数据基础上，为每个项目都引入图像和文本信息，然后构建单模态图分别进行训练。（2）嵌入层利用用户、项目的 ID 信息初始化向量表示，同时项目的图片、

文本向量由预训练模型得到。（3）信息传播层中在单个模态图上利用信息传播机制捕获用户在单模态下的偏好。（4）特征融合层中利用轻注意力机制将用户对不同模态信息的偏好融合到最终表示向量中。（5）预测层将基于最终融合得到的向量去评估用户的交互兴趣。

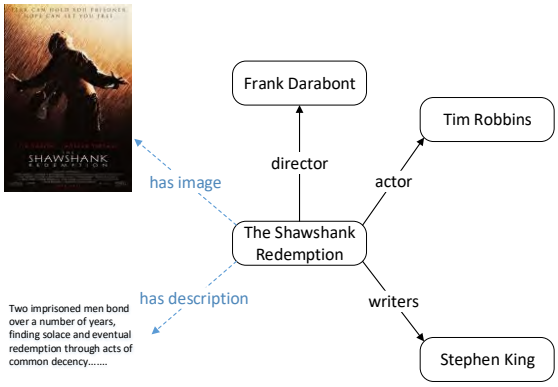


图 1 多模态项目示例

Fig.1 Example of a multi-modal item

2.3 嵌入层

用户和项目都有自己唯一的 ID 信息。利用用户和项目 ID 信息初始化用户和项目的表示向量，分别记为 e_u 和 e_i。

在第 k 个模态的交互图中，每一个项目 i 都有一个预训练模型得到的模态特征向量 $e_{k,i}$，用于表示该项目在此模态的特征含义，如利用 BERT[27]模型对项目的描述文本进行训练得到文本特征向量。所有的嵌入表示向量如下：

$$E=\left\{e_i,e_u,e_{k,i}\mid u\in U,i\in I,k\in K\right\} \tag{1}$$

其中 $e_u\in R^{|U|\times d}$，$e_i,e_{k,i}\in R^{|I|\times d}$，d 代表嵌入向量的长度。$e_i,e_u$ 在模型开始时随机初始化，然后在计算过程中不断被训练，$e_{k,i}$ 由预训练网络模型得到。

2.4 信息传播层

传统推荐模型一般直接将用户-项目对的表示向量送入预测模块进行训练，而基于 GNN 的模型会从用户-项目交互图中学习新表示。可以在多模态图上捕获用户在模态粒度上的偏好。

在交互图中，信息传播机制是节点利用图注意力网络聚合邻居信息并递归进行此步骤，以此来捕获用户-项目和项目-项目的关系。对于一个节点 h，其相邻的一跳邻居节点集合可以定义为 $N_h=\{t\mid(h,t)\in E\}$。对于信息从邻居节点 N_h 传播到节点 h 的公式可以表示为：

$$\mathbf{e}_{k,N_h}=\text{LeakyReLU}\left(\sum_{t\in N_h}f_g(h,t)\,f_a(h,t)W_{k,1}\mathbf{e}_{k,t}\right) \tag{2}$$

其中 k 表示第 k 个模态特征；$f_g(h,t)$ 表示门控模块的注意力机制，门控注意力机制决定邻居信息是否被传播到节点 h；$f_a(h,t)$ 表示 GAT 模块的注意力机制，决定邻居信息对节点 h 的影响大小；$W_{k,1}$ 是一个可训练的权重矩阵。

对于门控组件，受到之前的工作如 MKGAT 和 MGAT 启发，有四种不同的门控组合类型：

- 求和门控机制先将 $e_{k,h}$ 和 $e_{k,t}$ 相加，再乘上$\frac{1}{\sqrt{o}}$以处理邻居信息。

$$f_{ga}(h,t)=\delta\left(\frac{\mathbf{e}_{k,h}+\mathbf{e}_{k,t}}{\sqrt{d}}\right) \tag{3}$$

$\delta(.)$ 是激活函数，d 是节点 h 的度。所以门控参数依赖于节点 h 和节点 t。

- 内积门控机制将 $e_{k,h}$ 和 $e_{k,t}$ 相乘。

$$f_{gi}(h,t)=\delta\left(\frac{\mathbf{e}_{k,h}^{T}\mathbf{e}_{k,t}}{\sqrt{d}}\right) \tag{4}$$

- 连接门控机制将两个向量直接拼接。

$$f_{g_c}(h,t)=\delta\left(\frac{W_c\left(\mathbf{e}_{k,h}\|\,\mathbf{e}_{k,t}\right)}{\sqrt{d}}\right) \tag{5}$$

$\|$ 表示拼接操作，W_c 表示可训练权重矩阵。

- 还可以组合(4)和(5)的形式，得到更好的表示。公式可以表示为

$$f_{g_b}(h,t)=\delta\left(\frac{W_b\left(\mathbf{e}_{k,h}\|\,\mathbf{e}_{k,t}\right)+\mathbf{e}_{k,h}\odot\mathbf{e}_{k,t}}{\sqrt{d}}\right) \tag{6}$$

$\odot$ 表示元素相乘，W_b 表示可训练权重矩阵。

对于聚合邻居操作，我们引入注意力机制去学习不同邻居的重要性，如公式(7)所示：

$$f_a(h,t)=\left(W_{k,h}\mathbf{e}_{k,h}\right)^{\top}\tanh\left(W_{k,t}\mathbf{e}_{k,t}\right) \tag{7}$$

其中 tanh 是一个非线性激活函数；$W_{k,h}$ 和 $W_{k,t}$ 是可训练权重矩阵。此后我们采用 softmax 函数对所有邻居节点的注意力权重参数归一化，如公式(8)所示：

$$f_a(h,t)=\frac{\exp f_a(h,t)}{\sum_{t'\in N_h}\exp f_a\left(h,t'\right)} \tag{8}$$

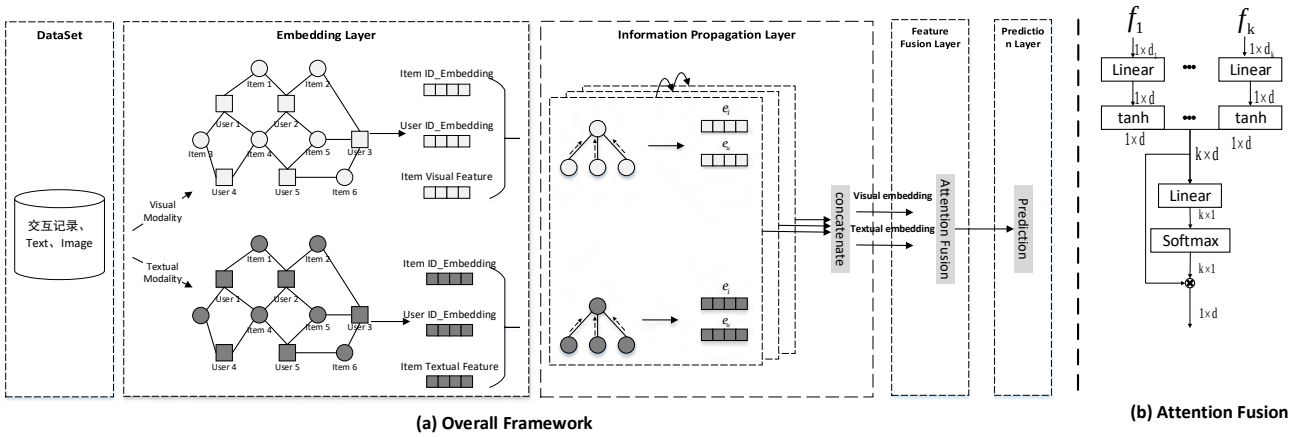


图 2 MFGAT 的总体框架

Fig.2 The overall framework of MFGAT

获得 $f_g(h,t)$ 和 $f_a(h,t)$ 后，对其进行内积，可以在模态粒度上捕获用户兴趣偏好。即 $f_g(h,t)$ 决定是否将这个模态信息传递给节点 h,而 $f_a(h,t)$ 决定这个模态信息对节点 h 的贡献程度。

然后我们利用从邻居传播来的信息 e_{k,N_h} 去更新节点 h 的表示。其中我们保留节点 h 的 ID 信息。如公式(9)所示：

$$\tilde{\mathbf{e}}_{k,h} = \text{LeakyReLU}\left(\mathbf{W}_{k,2}\mathbf{e}_{k,h} + \mathbf{e}_h\right) \tag{9}$$

其中 $W_{k,2}$ 是可训练权重矩阵；e_h 是节点 h 的 ID 表示向量。然后将 $\tilde{\mathbf{e}}_{k,h}$ 和 e_{k,N_h} 组合，如公式(10)表示：

$$\mathbf{e}_{k,h}^{(1)} = \text{LeakyReLU}\left(\mathbf{W}_{k,3}\mathbf{e}_{k,N_h} + \tilde{\mathbf{e}}_{k,h}\right) \tag{10}$$

其中 $\mathbf{e}_{k,h}^{(1)}$ 表示节点 h 在一阶连接邻居聚合后的向量表示；$W_{k,3}$ 表示可训练权重矩阵。

获得节点 h 一阶连接邻居聚合表示后，我们可以堆叠连接层去探索高阶连接表示，来增强效果。通过递归进行之前的聚合操作，节点 h 的高阶表示如公式(11)所示：

$$\mathbf{e}_{k,h}^{(l)} = \text{LeakyReLU}\left(\mathbf{W}_{k,3}^{(l-1)}\mathbf{e}_{k,N_h}^{(l-1)} + \tilde{\mathbf{e}}_{k,h}^{(l-1)}\right) \tag{11}$$

其中 $\mathbf{e}_{k,N_h}^{(l-1)}$ 表示第 $(l-1)$ 跳传播的表示；$\mathbf{e}_{k,h}^{(0)}$ 即初始向量 $e_{k,h}$。

2.5 特征融合

之前的多模态推荐工作如 MGCN 和 MGAT 对特征融合操作仅仅将多个模态向量相加再求平均，在融合层面上无法更好地区分用户对不同模态信息的偏好。受到 LAFF 工作的启发，利用注意力机制去区分用户的偏好程度。如图 2(b),同时我们的注意力模块扩展性更好，本文数据集只有文本描述和图片海报，当引入音频、视频等模态信息时，可以很方便的加入模型中，并且此注意力模块参数量较多头注意力模块要少很多，是一个轻量级模块。

各个模态传播层将训练得到的中间向量 $\left\{\mathbf{e}_{1,h}^{(l)},\ldots,\mathbf{e}_{k,h}^{(l)}\right\}$,向量长度为 $\{d_1,\ldots,d_k\}$。我们使用转换层将 K 个特征向量的维度统一到 d 维,如公式(12)所示：

$$f_k^{'} = \sigma\left(\text{ Linear }_{d_k\times d}\left(f_k\right)\right) \tag{12}$$

其中 σ 是一个非线性激活函数，本工作采用 tanh 函数以提高转换层的学习能力。$\text{Linear}_{d_k\times d}$ 表示全连接层，输入大小为 d_k，输出大小为 d。每一个输入向量有自己的线性层 Linear。当 $d=d_k$ 时，线性层可以不加。

考虑到不同模态信息重要程度不同，对转换后的特征加权融合，如公式(13)所示：

$$\overline{f} = \sum_{k}^{K} a_k f_k^{'} \tag{13}$$

其中权重参数 $\{a_1,\ldots,a_k\}$ 由轻注意力层计算得出，如公式(14)所示：

$$\{a_1,\ldots,a_k\} = \text{softmax}\left(\text{ Linear }_{d\times 1}\left(\left\{f_1^{'},\ldots,f_k^{'}\right\}\right)\right) \tag{14}$$

通过轻注意力层将不同模态特征融合，再送入预测层。同时本工作将轻注意力用于晚期特征融合，也可以尝试在预训练时对同一个模态信息，用不同

预训练模型训练，然后利用轻注意力层进行前期特征融合。

2.6 生成向量表示及模型预测

假设传播跳数为L，为了保留所有跳的信息，同时考虑一定的有序性，将各跳得到的用户和项目输出向量进行拼接得到最终向量，如公式(15)、(16)所示：

$$\mathbf{e}_u^* = \mathbf{e}_u^{(0)} \| \cdots \| \mathbf{e}_u^{(L)} \tag{15}$$

$$\mathbf{e}_i^* = \mathbf{e}_i^{(0)} \| \cdots \| \mathbf{e}_i^{(L)} \tag{16}$$

其中$\|$表示拼接操作；L 代表传播层数。通过这样做，我们不仅可以通过执行传播操作来丰富初始向量，还可以通过调整 L 来控制传播的强度。

最终我们执行用户和项目向量的内积，得到预测匹配分数，如公式(17)所示：

$$\hat{y}_{ui} = \mathbf{e}_u^{*\top} \mathbf{e}_i^* \tag{17}$$

然后我们使用贝叶斯个性化排序损失函数(BPR)去优化推荐预测损失。具体来说，我们假设用户喜欢之前交互过的项目，应该比没有交互过的项目预测得分更高。如公式(18)所示：

$$\mathcal{L} = \sum_{(u,i,j)\in O} -\ln\left(\delta\left(\hat{y}_{ui} - \hat{y}_{uj}\right)\right) + \lambda \| \theta \|_2^2 \tag{18}$$

其中，$O \in \left\{(u,i,j) | (u,i) \in R^+, (u,j) \in R^-\right\}$是训练集；$R^+$是用户-项目有交互记录的集合；$R^-$是没有交互记录的集合；$\delta(.)$是激活函数；$\lambda$是一个超参数，表示衰减因子；$\theta$是一个超参数。

3 实验和分析

3.1 数据集

为了验证模型的有效性，在公共数据集 MovieLens-20M 和 H&M 上进行了大量实验。

MovieLens 是推荐算法中常用的电影数据集，记录了用户对不同电影的显式评分，评分等级在 1 到 5 之间。本工作使用 MovieLens-20M 作为实验数据集，为了模拟用户-项目二部图，我们将评分转换为隐式反馈，保留评分为 5 的强相关交互记录作为正样本，用户对该项目交互记为 1，其余项目交互记为 0。为了丰富电影数据的多模态信息，我们从 IMDB[1]网站上爬取电影海报以及电影描述，并将电影名称与电影描述合并作为文本描述。过滤掉没有多模态信息的电影记录之后，对电影数据进行多模态特征提取，我们采用预训练模型 ResNet50[28]从海报图像中提取视觉特征，然后采用预训练模型 BERT 从文本描述中提取文本特征。

H&M[2]是 kaggle 多模态算法比赛公开的数据集。数据集本身提供了用户购买商品的记录，商品图片以及商品各种属性描述信息。因为数据量太大，筛选有超过 50 条交互记录的用户，同时筛选有多模态信息的商品；将商品名称、商品描述以及一些属性信息共同作为文本描述。同样采用预训练模型 ResNet50 从商品图片中提取视觉特征。然后采用预训练模型 BERT 从商品描述、名称、属性中提取文本信息。

数据集的统计数据如表 1 所示。对于每个数据集，将预处理之后的数据集的 80%作为训练集，10%作为测试集，10%作为验证集用于调整超参数。

3.2 实验设置

评估指标：

我们采用三种广泛使用的评估指标：Precision@K , Recall@K , NDCG@K。Precision@K 指标量化了排名前 K 的结果中有多少项是相关的。Recall@K 指标表示预测出的结果中相关结果占所有结果的比重。NDCG@ K 是归一化折损累计增益，一个与预测结果位置有关的指标，度量预测结果排名高低，值越大表示推荐系统推荐效果越好。我们设置 K=10 并且取测试集中预测所有用户的平均结果。每个用户的负样本被定义为未交互过的项目。所有实验代码都使用 PyTorch 框架实现

基线：

选择以下方法与我们的模型进行比较：

•NGCF[4]： NGCF 以一种显式的方式将协作信号加到向量中。通过合并来自多跳邻居信息来学习交互图中的高阶特征。本工作中，我们将所有多模态特征向量连接作为项目的向量表示，用于后续的模型训练。

•VBPR[8]：VBPR 将视觉特征注入到项目向量表示中，然后利用矩阵分解来学习基于用户和项目交互矩阵的表示向量。在我们的实验中，我们将数据集的多模态特征拼接成一个特征向量，并与 ID 信息集成，用于后续预测用户和项目之间的交互关系。

•MMGCN[24]：MMGCN 是一个基于图的算法。为了学习用户对不同模态的偏好，它分别对不同模态

[1] https://www.imdb.com/?ref_=nv_home

[2] https://www.kaggle.com/

表 1 数据集统计数据，V 和 T 表示视觉、文本特征向量长度

Table 1 The statistics of datasets,V and A represent the number of features used for the raw visual and textual.

数据集	用户数	项目数	交互数	稀疏度	V	T
MovieLens-20M	58985	8867	102569	99.98%	1000	768
H&M	67015	21272	2196622	99.84%	1000	768

表 2 MFGAT 和其它方法实验结果

Table 2 The experimental results of MFGAT and other methods

数据集	MovieLens-20M			H&M		
Models	Precision@10	Recall@10	NDCG@10	Precision@10	Recall@10	NDCG@10
VBPR	0.1961	0.4254	0.2402	0.1603	0.2469	0.1312
NGCF	0.2009	0.4540	0.2504	0.1804	0.2884	0.1590
MMGCN	0.2010	0.4709	0.2582	0.1674	0.2650	0.1460
MGAT	0.2076	0.4837	0.2693	0.1822	0.2894	0.1561
MFGAT	**0.2173**	**0.4935**	**0.2834**	**0.1866**	**0.2966**	**0.1593**
%Improv.	4.672%	2.026%	5.236%	2.415%	2.488%	2.05%

建立用户-项目交互图，最后再将不同图中学习到的用户-项目表示累加得到最终的用户-项目表示，再送入预测模块。

•MGAT[13]：MGAT 与 MMGCN 类似，也是一个基于图的算法。同时也分别对不同模态建立用户-项目交互二部图，但是与 MMGCN 不同在于在单个模态图训练过程中加入 GRU 门控模块，同时利用注意力机制聚合邻居信息，得到更好的向量表示。

参数设置：

我们采用 Xavier 初始器来初始化所有模型参数，然后使用 Adam 优化器优化模型。利用网格搜索进行超参数调优，其中学习率的值选择分别为 $\{e^{-1},e^{-2},e^{-3},e^{-4},e^{-5}\}$。图神经网络层数 L 范围为 $\{1,2,3\}$。权重衰减因子和注意力 Dropout 比率值选择范围分别为 $\{e^{-1},e^{-2},e^{-3},e^{-4},e^{-5}\}$ 和 $\{0.1,0.2,...,0.8\}$。节点 dropout 和消息 dropout 默认为 0。其他基线模型都使用原论文中超参数。

3.3 结果比较

将模型与基线方法进行比较，结果如表 2 所示。每个指标的最佳结果使用黑体加粗。通过观察表中数据，我们有以下发现：

MFGAT 表现要优于所有基线模型。结果证明了我们模型设计的合理性和有效性。与传统考虑使用用户-项目交互记录的矩阵分解算法（如 VBPR）相比，MFGAT 基于图算法，同时利用高阶连接性聚合邻居信息可以提高表示学习的效果。与基于 GNN 的推荐算法（如 NGCF，NGCF 在用户-项目交互二部图上，仅将不同模态信息的特征统一为一个特征向量）相比，MFGAT 对不同模态分别建立图，学习到不同通道特征向量再进行预测，有更好的表示能力。与 MMGCN 和 MGAT 算法相比，虽然都是分别对不同模态建立图再通过高阶传播特征进行学习，但 MMGCN 没有使用注意力机制区别不同邻居重要性，MGAT 虽然使用门控机制和注意力机制区分邻居的信息的重要性，但是在最后不同模态特征融合时并没有区分不同模态重要性，仅仅取平均值。而 MFGAT 通过在特征融合时加上轻注意力机制，可以区分不同模态信息的重要性，从而识别用户的细粒度偏好。

基于 GNN 的模型在 MovieLens-20M 和 H&M 上优于基于 CF 的模型。这些效果改进归因于图卷积层和注意力机制，这些操作不仅捕获了局部结构信息，同时还学习了每个节点的邻居特征的分布。基于多模态图的模型（MFGAT，MGAT，MMGCN）优于其他基线模型。与直接统一多模态特征相比，分别对不同模态进行学习可以获得更好的性能。这表明关注不同模态信息可以更好地建模用户偏好。其中在 H&M 数据集上 NGCF 模型效果好于 MMGCN,推测由于 H&M 数据集是服装数据集，用户对于图片更加关注，而且 MMGCN 卷积层没有注意力机制来区分不同模态信息重要性，从而导致引

表 3 不同门控机制实验结果

Table3 The experimental results of gate mechanisms

数据集	MovieLens-20M			H&M		
metrics	Precision@10	Recall@10	NDCG@10	Precision@10	Recall@10	NDCG@10
MFGAT_a	0.2100	0.4856	0.2708	0.1659	0.2776	0.1459
MFGAT_bi	0.2116	0.4887	0.2739	0.1775	0.2878	0.1469
MFGAT_c	0.2132	0.4912	0.2730	0.1821	0.2895	0.1546
MFGAT_i	**0.2173**	**0.4935**	**0.2834**	**0.1866**	**0.2966**	**0.1593**
%Improv.	1.875%	0.081%	4.653%	2.412%	2.394%	3.04%

入较多噪声。

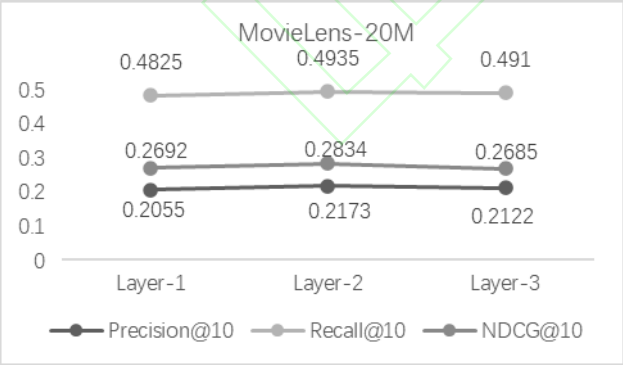
使用注意力机制的模型（如 MFGAT 和 MGAT）效果优于 MMGCN，表示注意力机制有利于捕获用户更细粒度的偏好。

3.4 模型分析与讨论

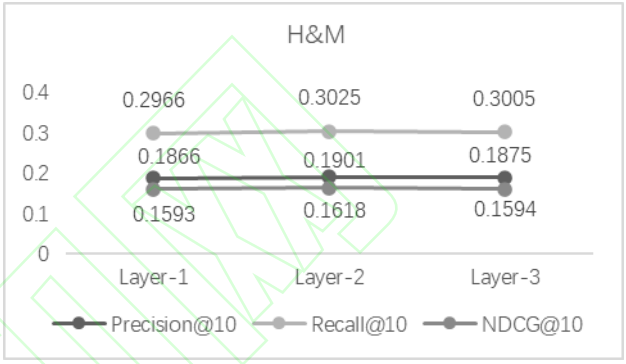
接下来对模型进一步分析和讨论，以分析可能对我们模型的性能产生影响的因素。

图神经网络传播层数的影响：

为了评估传播层数给 MFGAT 性能带来的影响，我们总共设计了三种不同的传播层数实验，传播层数分别是 1,2,3，MFGAT_1 代表聚合一阶邻居。此外，在两个数据集上嵌入向量长度都为 64。如图 3 所示，对不同层数建模的 MFGAT 性能进行比较。根据实验结果可以看到，MFGAT_2 的表现要优于 MFGAT_1 和 MFGAT_3，这与 MMGCN 和 MGAT 论文观察的结果一致。说明了图神经网络过平滑的问题，即堆叠更多的图卷积层或从高阶邻居聚合信息容易引入噪声，导致效果下降。



3.(a) MovieLens-20M 数据集结果



3.(b) H&M 数据集结果

图 3 数据集上图神经网络层数的影响

Fig.3 The influence of depth of GNN on datasets

不同门控机制的影响：

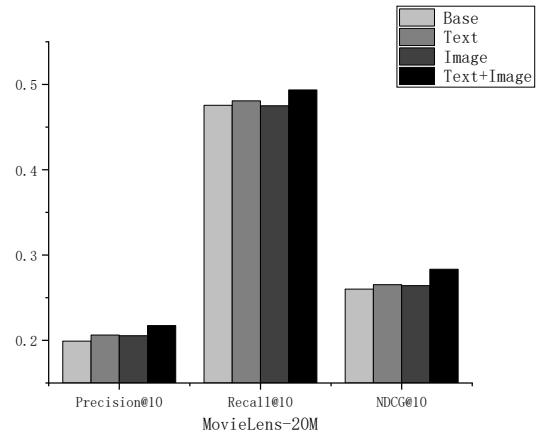
为了评估不同的门控注意力机制引起的影响，我们对门控机制的四种变量进行了实验，包括内积(MFGAT_i)、求和(MFGAT_a)、拼接(MFGAT_c)和 Bi-interaction(MFGAT_bi)。实验结果如表 3 所示，MFGAT_i 的效果优于其他三种方法。这表明内积方法可能更适合基于多模态图模型中的关系建模。相比之下其他方法可能由于参数过多导致过拟合。

模态信息数量的影响：

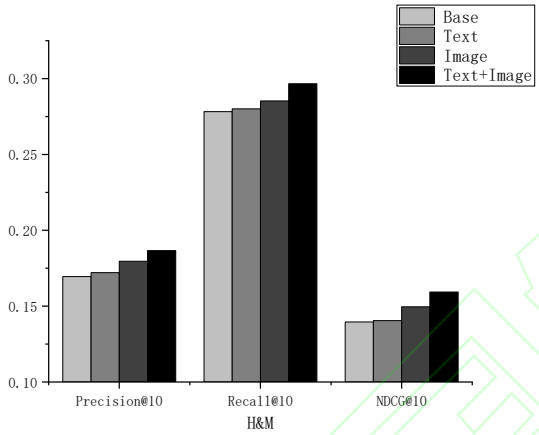
为了评估加入不同模态信息对实验结果的影响，我们设计了三个实验：不添加文本和图像、只添加文本、只添加图像、同时添加文本和图像。实验结果如图 4 所示，我们观察到两点如下：

一是多模态信息的加入可以获得更好的预测效果，并且模态信息种类累加有利于提高推荐性能；

二是不同数据集中不同模态信息重要性不同，如电影数据集中加入文本的效果比图像效果好，可能由于用户更容易受文本描述影响，就像我们找电影会更倾向于看简介，因为我们通过一张海报并不能更多地捕获电影细节，这个实验结果符合预期。



4.(a) MovieLens-20M 数据集结果



4.(b) H&M 数据集结果

图 4 数据集上模态类型的影响

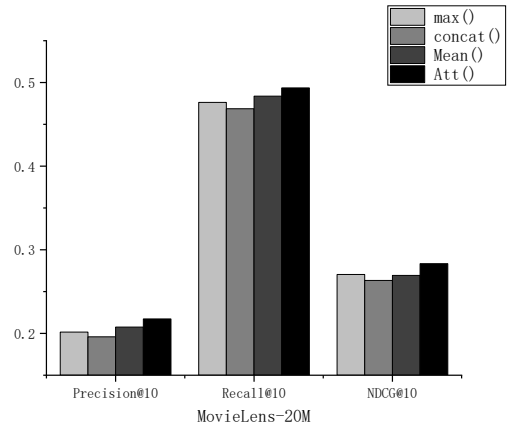
Fig.4 The influence of multimodal type on datasets

模态融合中不同融合方法的影响：

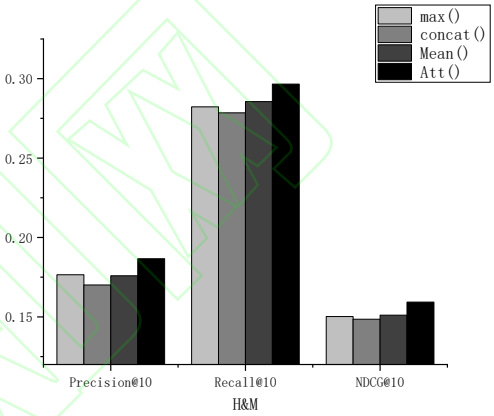
为了探索不用的融合方法对实验结果的影响，我们设计了四种融合方法，包括取最大值(Max)、拼接(Concat)、取均值(Mean)和注意力机制(Att)。实验结果如图 5 所示，我们可以看到加入注意力机制后能获得更好的效果。说明注意力机制可以更好地对用户兴趣建模，更好地捕获用户兴趣。

3.5 用户对不同模态偏好的分析：

为了探索用户对不同模态的偏好程度，并直观的展示用户偏好和轻注意力机制的作用。我们给出了一个案例研究，即从 MovieLens-20M 数据集中随机选择 10 个节点研究其在特征融合阶段对文本和图片模态的注意力权重。得益于轻注意力机制，我们可以计算注意力得分，并在图 6 中可视化相关得



5.(a) MovieLens-20M 数据集结果



5.(b) H&M 数据集结果

图 5 多模态融合方法的影响

Fig.5 The influence of multimodal fusion methods

分。我们可以观察到在 MovieLens-20M 中，文本模态得分普遍偏高，说明用户对于电影的文本描述更加关注。这与图 4 中加入文本模态相对于加入图片模态性能提升更多的结果相符合。也证明了在多模态特征融合阶段加入轻注意力机制的有效性。

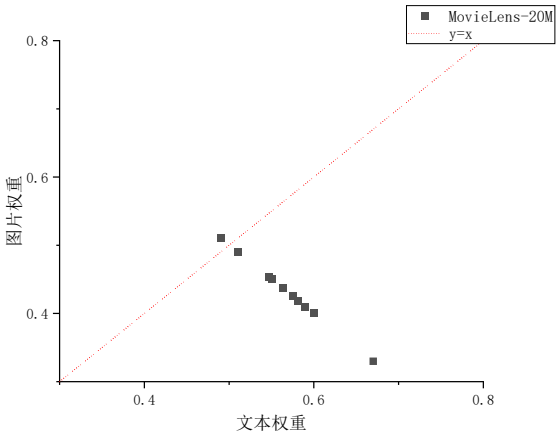


图 6 注意力权重分布散点图

Fig.6 Scatter plot of attention weight distribution

4 结束语

在本文中,我们提出了一种基于图的多模态融合推荐算法 MFGAT。分别对不同模态建立交互图，在单模态图上聚合高阶邻居信息，同时利用注意力机制去除噪声和区分邻居信息重要性，以此进行用户偏好建模。在最后模态特征融合时，设计轻注意力模块去捕获用户对不同模态的偏好，同时模块设计兼顾可扩展性和便利性。基于两个公开数据集上的大量实验证实了我们提出的 MFGAT 方法的有效性和合理性。下一步将对引入更多辅助信息带来的影响进行研究，以达到进一步提升推荐准确性和验证轻注意力模块有效性的目的。

参考文献

[1] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[2] Kabbur S, Ning X, Karypis G. Fism: factored item similarity models for top-n recommender systems[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 659-667.

[3] He X, He Z, Song J, et al. Nais: Neural attentive item similarity model for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(12): 2354-2366.

[4] Wang X, He X, Wang M, et al. Neural graph collaborative filtering[C]//Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval. 2019: 165-174.

[5] He X, Deng K, Wang X, et al. Lightgcn: Simplifying and powering graph convolution network for recommendation[C]//Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020: 639-648.

[6] Liu M, Nie L, Wang X, et al. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning[J]. IEEE Transactions on Image Processing, 2018, 28(3): 1235-1247.

[7] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[J]. arXiv preprint arXiv:1205.2618, 2012.

[8] He R, McAuley J. VBPR: visual bayesian personalized ranking from implicit feedback[C]//Proceedings of the AAAI conference on artificial intelligence. 2016, 30(1).

[9] Chen J, Zhang H, He X, et al. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention[C]//Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017: 335-344.

[10] Sun R, Cao X, Zhao Y, et al. Multi-modal knowledge graphs for recommender systems[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1405-1414.

[11] Wang X, He X, Cao Y, et al. Kgat: Knowledge graph attention network for recommendation[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 950-958.

[12] Wu S, Sun F, Zhang W, et al. Graph neural networks in recommender systems: a survey[J]. ACM Computing Surveys, 2022, 55(5): 1-37.

[13] Tao Z, Wei Y, Wang X, et al. Mgat: Multimodal graph attention network for recommendation[J]. Information Processing & Management, 2020, 57(5): 102277.

[14] Dong J, Li X, Snoek C G M. Predicting visual features from text for image and video caption retrieval[J]. IEEE Transactions on Multimedia, 2018, 20(12): 3377-3388.

[15] Liu Y, Albanie S, Nagrani A, et al. Use what you have: Video retrieval using representations from collaborative experts[J]. arXiv preprint arXiv:1907.13487, 2019.

[16] Gabeur V, Sun C, Alahari K, et al. Multi-modal transformer for video retrieval[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer International Publishing, 2020: 214-229.

[17] Li X, Xu C, Yang G, et al. W2vv++ fully deep learning for ad-hoc video search[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 1786-1794.

[18] Li X, Zhou F, Xu C, et al. Sea: Sentence encoder assembly for video retrieval by textual queries[J]. IEEE Transactions on Multimedia, 2020, 23: 4351-4362.

[19] Hu F, Chen A, Wang Z, et al. Lightweight attentional feature fusion for video retrieval by text[J]. arXiv preprint arXiv:2112.01832, 2021.

[20] Cao Y, Wang X, He X, et al. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences[C]//The world wide web conference. 2019: 151-161.

[21] Chen T, He X, Kan M Y. Context-aware image tweet modelling and recommendation[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 1018-1027.

[22] Gao J, Zhang T, Xu C. A unified personalized video recommendation via dynamic recurrent neural networks[C]//Proceedings of the 25th ACM international

conference on Multimedia. 2017: 127-135.

[23] 潘华莉,谢珺,高婧,等.融合多模态特征的深度强化学习推荐模型[J]. 数据分析与知识发现: 1-18.
PAN H L, XIE J, GAO J, et al. Deep Reinforcement Learning Recommendation Model Fused with Multi-modal Features [J]. Data Analysis and Knowledge Discovery:1-18.

[24] Wei Y, Wang X, Nie L, et al. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video[C]//Proceedings of the 27th ACM international conference on multimedia. 2019: 1437-1445.

[25] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.

[26] 胡承佐,王庆梅,李迪超,等. 基于复杂结构信息的图神经网络序列推荐算法[J]. 计算机工程，2022, 48(5): 82-90+97.
HU C Z, WANG Q M, LI D C, et al. Sequence Recommendation Algorithm of Graph Neural Networks Based on Complex Structure Information[J]. Computer Engineering, 2022, 48(5): 82-90+97.

[27] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.