引用格式：李泽宇，王紫欣．一种结合ViLBERT和多模态知识图谱注意力网络的新闻推荐方法[J]．中国传媒大学学报（自然科学版），2023，30(05)：15-25.
文章编号：1673-4793(2023)05-0015-11

一种结合ViLBERT和多模态知识图谱注意力网络的新闻推荐方法

李泽宇[1]，王紫欣[2*]
（1.北京邮电大学，北京 100876；2.中国传媒大学，北京 100024）

摘要：推荐系统在解决新闻准确呈现的问题上显示出巨大的潜力。现有的新闻推荐系统大多只考虑新闻文本，忽略了新闻图片与用户之间的关系。但新闻图片也是用户决定点击新闻的重要因素。本文将ViLBERT与多模态知识图注意力网络相结合，利用多模态知识提高新闻推荐系统的准确率，使用多模态图关注技术在多模态知识图关注网络上传播信息，将生成的图像和文本聚合嵌入推荐的表示，以有效地表征目标，缓解推荐系统中用户行为稀疏和冷启动的问题。通过在两个不同的真实中英文新闻数据集上进行了实验，结果表明本模型可以有效地提高新闻推荐的准确率。

关键词：新闻推荐；多模态；图卷积网络；ViLBERT
中图分类号：TP391 文献标识码：A

VMKGAT：ViLBERT combined with multi-modal knowledge graphs attention network for news recommendation

LI Zeyu[1]，WANG Zixin[2*]
(1. Beijing University of Posts and Telecommunication, Beijing 100876, China；
2. Communication University of China, Beijing 100024, China)

Abstract: Recommender systems have shown great potential to solve the problem of accurate presentation of news. Most of the existing news recommender systems only consider the news texts but ignore the relationship between the news picture and user. However, news images are also a significant factor in users′ decision to click on news. In this paper, we proposed ViLBERT combined with Multi-modal Knowledge Graphs Attention Network (VMKGAT) to better enhance the accuracy of the news recommender system by using multi-modal knowledge. We used a multi-modal graph attention technique to disseminate information on the multi-modal knowledge graph attention network, and then used the generated images and text aggregation to embed the representation for the recommendation. It could effectively characterize the target item and alleviate the problems of sparse user behavior and cold start in recommendation system. We conducted a large number of experiments on two different real English and Chinese news datasets, and the experimental results show that our model VMKGAT can effectively improve news recommendation.

Keywords: news recommendation; multimodality; graph convolutional networks; ViLBERT

作者简介（*为通讯作者）：李泽宇（1998-），男，博士研究生，主要从事媒体大数据等方面的研究。Email：lizeyu@bupt.edu.cn；王紫欣（2001-），女，硕士研究生，主要从事媒体大数据等方向的研究。Email：1667286922@qq.com

1 引言

在网络媒体快速发展的现代社会,如何高效地为用户推荐其可能感兴趣的新闻,已成为推荐系统领域越来越受关注的研究课题。

现有的新闻推荐系统常用基于协同过滤(Collaborative Filtering,CF)[1]的方法。但早期直接利用与目标用户兴趣相似的用户为其进行推荐的基于领域的协同过滤,以及后期出现的以奇异值分解(Singular Value Decomposition,SVD)为代表的基于潜在因子模型的协同过滤方法,都面临着数据稀疏问题和冷启动问题[2]。

随着深度学习的发展,Salakhutdinov 等[3]将受限玻尔兹曼机与协同过滤结合,将深度学习的方法应用于推荐系统领域。随后循环神经网络(Recurrent Neural Networks,RNN)和卷积神经网络(Convolutional Neural Networks,CNN)等深度学习网络被引入推荐系统领域[4-5]。基于深度学习的推荐算法从用户和项目的历史交互数据中进行特征学习,将数据映射到另一维度的空间中,获得用户和项目的深层次特征表示。该方法能够有效提升推荐系统的准确度,但数据稀疏所带来的一系列问题依然存在。加入更多的辅助信息能够有效解决数据稀疏问题和冷启动问题。

知识图谱(Knowledge Graph,KG)包含内容丰富的辅助数据,可以精确地描述各类用户和项目的属性[6]。Wang 等[7]通过实验证明,基于知识图谱的推荐系统能有效缓解数据稀疏导致的一系列问题。因此,将知识图谱作为辅助数据源引入推荐系统,已成为近年来推荐系统领域的研究热点[8-9]。然而,目前基于知识图的新闻推荐均忽略了新闻中的多模态信息,但新闻的视觉和文字特征在推荐系统中同样重要。因此,有必要将多模态信息引入知识图,将图像或文本当做实体或者实体相关属性。图1为具体的多模态知识图谱(Multi-Modal Knowledge Graph,MKG)示例。本文

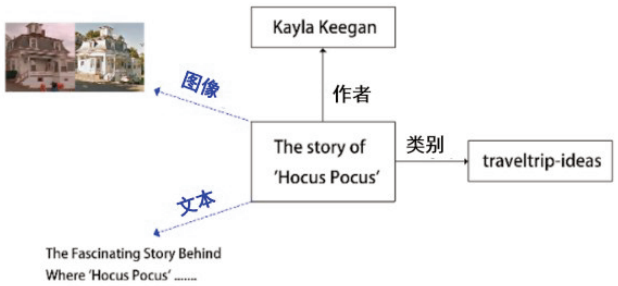

图1 多模态知识图谱示例图

实验也证明,引入MKG能够有效提高新闻推荐的质量。

基于知识图的推荐中最关键的部分为知识图表示学习,基于知识图的推荐模型通常使用知识图表示模型来学习知识图实体的嵌入,将其反馈到具体的推荐任务中。目前的知识图表示学习分为两类:基于特征的方法[10-11]和基于实体的方法[12]。前者将各类别的模态信息当做对应实体的辅助特征,通过实体对应的标题和图像中分别提取文本信息和视觉表示,可以拓展翻译模型(the Translational Models)[13],其中某个映射在关系空间中的三元组的似然性得分(或能量得分)是根据KGs的结构及该三元组中对于实体的向量表示决定的。但是基于特征的方法构建知识图的数据源要保证知识图中的每个实体都有相同种类的多模态信息,这样的要求在实际数据集中很难被满足。因此,研究人员提出了基于实体的方法,这种方法不再将文本和图像等不同类型的信息视为实体的辅助信息,而是将其都当做知识图谱中的实体,进而参与构建不同的三元组。基于实体的方法可以充分且直观地利用知识图谱的网络结构,通过引入新的关系来引入视觉和文本信息。引入新的信息后可以通过翻译模型学习知识图的嵌入。虽然基于实体的方法能降低对MKG数据源的要求,但其只能关注到实体之间的推理关系,容易忽略多模态信息之间的融合。实际上,多模态信息在融合后,本身就可以作为辅助信息丰富其他实体的信息。因此,在建模实体之间的推理关系之前,需要一种直接的交互方式,将多模态信息显式地融合到对应的实体中。为此,本文提出了一个能充分利用多模态信息的MKG表示模型,这一模型采用基于实体的方法构建多模态知识图,结合ViLBERT的多模态知识图注意网络(ViLBERT combined with Multi-modal Knowledge Graphs Attention Network for News Recommendation,VMKGAT)。模型首先通过预先训练的Mask R-CNN模型[14]提取新闻图像的感兴趣区域(Region of Interest,ROI),用于目标检测。然后使用预先训练的视觉语言模型[15]对新闻文本和新闻图像的ROI进行编码,并通过注意力Transformer网络(Co-attentional Transformer Network)对其固有的跨模态相关性进行建模,学习准确的多模态新闻表示。融合了多模态信息后,VMKGAT通过实体信息聚合实体的邻居节点信息,再进行实体关系的推理,利用三元组的打分函数(如TransR)构建推理关系。VMKGAT模型不是对每个知识图进行三元组

独立处理，而是对实体的邻居信息进行聚合，可以更好地学习融合其他模态信息的实体嵌入。

本文的贡献可以总结为：

（1）将多模态知识图引入到新闻推荐领域。

（2）开发了一种新的VMKGAT模型，在图神经网络框架下以显式信息融合方式实现了高阶关系建模。

（3）在两个大规模真实英文和中文新闻数据集上进行大量实验证明了模型的合理性和有效性。

2 研究现状

本文相关的背景知识和现有工作包括多模态知识图谱、多模态预训练和基于知识图的推荐。

2.1 多模态知识图谱

知识图谱技术已经被广泛用于处理结构化数据和文本数据，但对非结构化的视觉数据关注度较低，缺乏有效的技术手段从中提取结构化知识。近年来，一些多模态视觉技术被提出，但这些技术主要用于提升图像分类、图像生成、图像问答，不能支撑多模态知识图谱的构建。视觉数据库通常是图像或视频数据，提供关于知识图谱中实体的视觉信息。如果在更大范围内进行链接预测和实体对齐，并进行实体关系抽取，现有的模型在综合考虑文本和视觉特征时应能获得更好的性能。

传统知识图谱主要集中研究文本和数据库的实体和关系，多模态知识图谱在传统知识图谱的基础上构建了多种模态（例如视觉模态）下的实体语义关系[24]，如图1所示。多模态知识图谱的应用场景十分广泛，一个完备的多模态知识图谱对跨领域的信息融合研究有极大帮助。多模态结构数据在底层表征上是异构的，但相同实体的不同模态数据在高层语义上是统一的，因此多种模态数据的融合有利于推进语言表示等模型的发展，为在语义层级构建多种模态下统一的语言表示模型提供数据支持。其次多模态知识图谱技术可以服务于各种下游领域，例如多模态实体链接技术可以融合多种模态下的相同实体。

2.2 多模态预训练

多模态新闻推荐任务中，多模态数据的处理十分重要，数据的预处理及预训练的质量会直接影响推荐系统的准确度。近年来，自然语言处理领域较热门的BERT模型[16]是基于Transformer模型的双向语言模型，其使用大量的无标注数据预训练，明显提高了多种自然语言处理任务的效果。本文使用的ViLBERT模型[15]是最早将BERT扩展到多模态的模型之一。

目前，图像-文本多模态预训练模型主要采用Transformer结构，训练数据来自图像标注数据集的图像-文本对，其中的文本是对应图像的自然语言描述。对各下游任务，模型的使用方法可以参照纯文本Transformer模型的使用方法：对分类任务，提取<CLS>或所需位置对应的表示传入分类器；对序列任务，通过对Transformer模型输出的表示序列进行后续操作。ViLBERT处理多模态数据的方式采用双流结构，分别对每种模态进行建模，通过一组基于注意力的交互将模态进行融合。这种方法允许对每种模态使用可变的网络深度，并支持不同深度的跨模态连接。图像和文本两种模态分为两条路径进行处理，图像和文本的表示只在模型尾段发生交互。在ViLBERT图像流（图2上半部分）中，图像首先通过特征抽取模型提取出一系列ROI和每个ROI的向量表示，传入随机初始化的层Transformer Encoder中；为了编码ROI的位置信息，每个ROI的表示都加上了被投影到与其表示相同维度的5维位置信息。文本流（图2下半部分）采用预训练好的层BERT，对文本的处理与BERT一致。

图像和文本分别含有分类标记。图像分类标记<IMG>在被拼接于ROI序列前传入图像流，文本分类标记<CLS>在被拼接于token序列前传入文本流，通过计算对应输出表示hv与hw的点积并学习一个线性层，可以判断给定的图像和文本是否匹配。

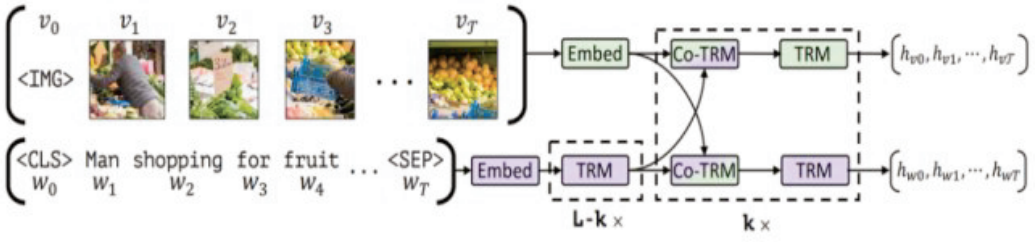

图2 ViLBERT示例图

2.3 基于知识图的推荐

近年来,已有研究在尝试利用KGs结构进行推荐,现有的方法大体上可以分为基于嵌入的方法、基于路径的方法和混合方法三种类型。

基于嵌入的方法[8, 17]通常用知识图嵌入(Knowledge Graph Embedding,KGE)[18]算法对知识图谱进行预处理,将知识图谱中的所有实体和关系统一表示为嵌入向量,进而扩充原有物品和用户表征的语义信息,再将学习到的实体和关系嵌入做为推荐部分的输入。Zhang等[19]结合知识图谱表示学习方法提出了协同知识库嵌入(Collaborative Knowledge base Embedding,CKE),将CF模块与知识嵌入、文本嵌入和项目图像嵌入结合在一个统一的贝叶斯框架中。Wang等[8]提出的深度知识网络(Deep Knowledge-Aware Network, DKN)将实体嵌入和单词嵌入作为不同的通道,使用CNN框架将其结合进行新闻推荐。之后Wang等[20]又提出了知识图谱增强推荐的多任务特征学习(Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation, MKR),通过多任务学习框架,交替进行知识图谱表示学习和推荐模块训练,可抑制过拟合并提高泛化能力。从实际的推荐效果来看,基于嵌入的方法在利用知识图辅助推荐系统方面表现出了很高的灵活性,但由于这些方法中采用了KGE算法和平移模型,所以基于嵌入的方法仅利用了知识图谱中的语义信息,而未能很好地利用知识图中的结构信息,导致该方法会忽略掉多模态信息之间的融合,不能很好地处理多模态知识图。

基于路径的方法[21-22] 通常会将知识图谱视为一个异构信息网络(Heterogeneous Information Network,HIN),通过探索知识图中项目之间的各种连接模式,为推荐提供额外的指导。例如,在个性化实体推荐(Personalized Entity Recommendation,PER)[21]和基于元图的推荐[22]中提取基于元路径/元图的潜在特征,用这些特征表示用户和项目之间不同类型的关系路径/图的连通性。基于路径的方法通常能够充分且直观地利用知识图谱的网络结构,但需要手工构建元路径,且构建的元路径的质量对推荐结果影响较大。因此不能方便有效地挖掘图谱中所有信息。基于路径的方法在实践中很难进行优化且不能适用于所有场景。

混合方法是由基于嵌入的方法和基于路径的方法集成而来的。基于嵌入的方法利用KGs中实体的语义表示进行推荐,基于路径的方法使用KGs中实体之间的连接模式进行推荐,两者都只利用了KGs中信息的一个方面。为了充分利用KGs中的信息进行更好的推荐,研究人员提出了混合的方法,该类方法目前还没有统一的权威定义,但大多集成了实体和关系的语义表示以及连接信息的模式,同时依赖于KGE。例如,Li等[23]提出的注意增强的知识感知用户偏好模型(Attention-Enhanced Knowledge-Aware User Preference Model for Recommend,AKUPM)和Wang等[9]提出的知识图注意网络(Knowledge Graph Attention Network for Recommendation,KGAT)。基于混合方法的模型大多采用了KGE,所以与基于嵌入的方法类似,混合方法也会独立地处理每一个三元组,而不考虑多模态信息融合。

3 模型框架

3.1 基本概念

本节介绍本文模型中涉及到的一些概念,并对基于多模态知识图谱的推荐任务建模。

3.1.1 用户-项目二部图

在推荐场景中,通常有历史的用户-项目交互(例如单击)。这里把用户和新闻的交互数据表示成用户-项目二部图 $\mathcal{G}_1 = \{(u,y_{ui},i)|u \in U,i \in I\}$,其中 U 和 I 分别表示用户集和项目集,函数 $y_{ui} = 1$ 表示用户 u 与项目 i 之间存在观察到的交互行为,否则 $y_{ui} = 0$。

3.1.2 知识图谱

除了交互之外,通常考虑知识图谱中项目边(side)信息。这些辅助数据一般由真实世界的实体和它们之间的关系组成,用以分析一个项目。例如,一篇新闻可以由作者、标题和类型等来描述。

知识图谱 $\mathcal{G}_2 = (V,E)$ 为有向图,其中 V 为节点集,E 为边集。节点是实体,边是主体-属性-对象(subject-property-object triple facts)三重事实。每条边都属于一个关系类型 $r \in R$,其中 R 是一组关系类型。(head entity, relation, tail entity)形式的每条边记为 (h,r,t),其中 $(h, t \in V, r \in R)$ 表示 r 从 h 到 t 的关系。

3.1.3 协同知识图谱

协同知识图谱(Collaborative Knowledge Graph,CKG)将用户行为和项目知识编码为统一的关系图,本文用 $\mathcal{G}$ 表示。CKG首先定义一个用户-项目二部图,基于item-entity对齐集,可以将用户-项目二部图与知识图谱无缝集成为统一的图。如图3所示,i_{e1}、i_{e2} 和 i_{e3} 同时出现在知识图谱和用户-项目二部图中,CKG的对齐依赖于它们。

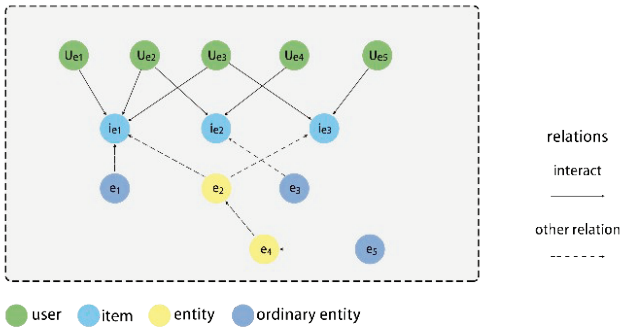

图3 协同知识图谱示例图

3.1.4 任务描述

本文设计了基于多模式 KGs 的推荐任务，即：输入协作知识图谱，包括用户-项目二部图和多模态知识图谱；输出一个预测函数 $\hat{y}_{ui}$，用于预测用户 u 选择某个新闻 i 的概率。

3.2 架构概述

VMKGAT 模型框架如图 4 所示，主要包括三部分：

（1）嵌入层：通过多模态知识图谱实体编码器，在保留 CKG 结构的同时，将每个节点作为一个向量进行参数化。

（2）有关注的嵌入传播层：递归地将嵌入信息从一个节点的邻居中传播，将每个实体的邻居实体信息聚合到每个实体本身，学习新的实体嵌入。

（3）预测层：将用户和项目在所有传播层的表示进行聚合，输出预测的匹配分数。

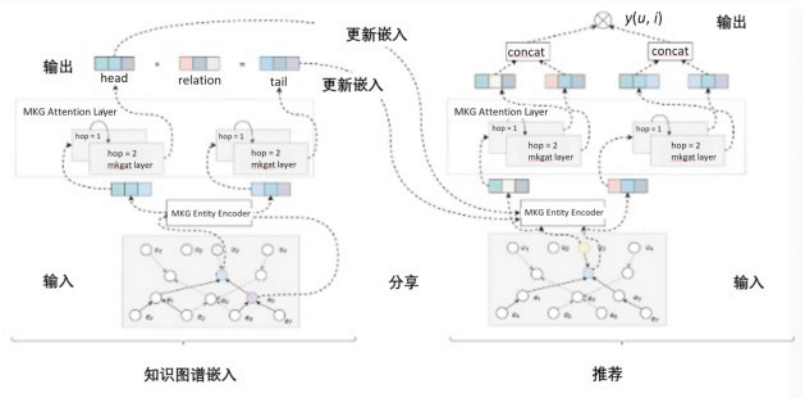

图4 VMKAGT 模型框架图

VMKGAT 模型中多模态知识图谱实体编码器（Multi-Modal Knowledge Graph Entity Encoder）能将多模态实体合并到模型中，并利用深度学习为实体构建编码器并表示，为所有实体提供嵌入。如图 5 所示，使用不同的编码器嵌入特定的数据类型。

结构化知识是以 (h,r,t) 形式存在的三元组信息。为了将 head entity h, tail entity t 和 relation r 表示为独立的嵌入向量，通过嵌入层传递实体 id 或关系 id 生成密集向量。

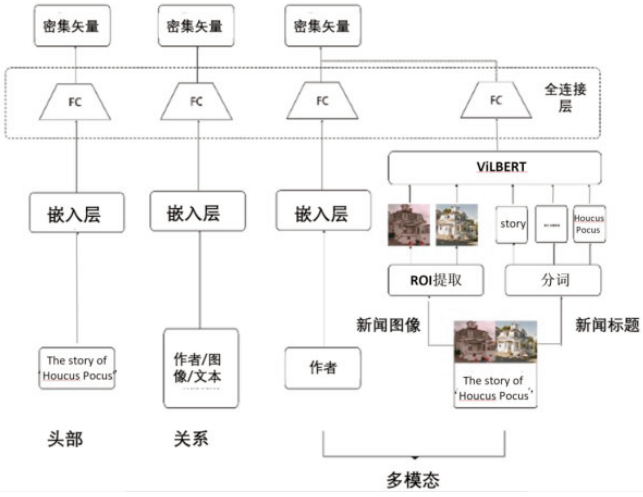

图5 多模态知识图谱实体编码器

通过预先训练的Mask R-CNN模型[14]提取新闻图像的ROI,然后将ROI输入到预先训练的视觉语言模型[15]中,实现对新闻图像的ROI编码。

文本信息与内容高度相关,可捕捉用户的偏好。本文将文本转化为词序列,输入到预先训练的视觉语言模型中,得到对应的句子向量。

最后,如图5所示,使用全连接层将实体的所有模态统一到同一个维度中。

如图6所示,沿着高阶连接性[25]递归地传播嵌入。利用图注意网络(Graph Attention Networks, GATs)[26]思想,生成级联传播的注意权值,以揭示这种连通性的重要性。GATs虽然有效,但由于忽略了KGs之间的关系,不适用于KGs,因此本文对GATs进行修改,考虑了KGs关系的嵌入。此外,注意力机制[27]的引入可以减少噪声的影响,使模型关注有用信息。

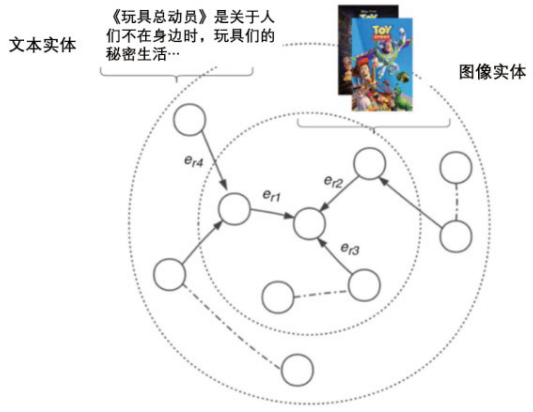

图6 多模态知识图谱注意层示意图

3.3 嵌入层

知识图嵌入是将实体和关系作为向量表示进行参数化的有效方法,保留了图的结构。本文在CKG上使用了TransR方法。具体来说,如果图中存在一个三元组(h, r, t),首先将实体空间内的实体通过Mr矩阵投影到关系r所在空间内,得到hr和tr,然后使$hr+t\approx tr$,学习嵌入各个实体和关系。本文用e_h, $e_t \in \mathbb{R}^d$, $e_r \in \mathbb{R}^k$分别表示h,t和r的嵌入。因此,对给定的三元组(h, r, t),其似然性得分(或能量得分)表述如下式:

$$s(h,r,t) = \left\| W_r e_h + e_r - W_r e_t \right\|_2^2 \tag{1}$$

其中,$W_r \in \mathbb{R}^{k \times d}$为关系$r$的变换矩阵,将$d$维实体空间中的实体投影到$k$维关系空间中。$s(h,r,t)$的分数越接近0,则表明三元组更有可能是真实的,反之亦然。

知识图谱嵌入的TransR训练考虑了有效三元组和无效三元组之间的相对顺序,并通过计算排名损失来考量它们的区分:

$$L_{KG} = \sum_{(h,r,t,t') \in \mathcal{T}} - \ln \sigma\left(s(h,r,t') - s(h,r,t)\right) \tag{2}$$

其中,$\mathcal{T} = \{(h,r,t,t') | (h,r,t) \in \mathcal{G}, (h,r,t') \notin \mathcal{G}\}$,$(h, r, t')$是通过随机替换有效三元组中的一个实体来构造得到的无效三元组。$\sigma(.)$是sigmoid函数。该层以三元组的粒度对实体和关系进行建模,作为正则化器并将直接连接注入到表示中,提高模型的表示能力。

3.4 注意力嵌入传播层

一个实体可以包含在多个三元组中,连接多个三元组和传播信息。与图卷积网络(Graph Convolutional Network, GCN)[25]或者图神经网络(Graph Sample and Aggregate, GraphSage)[30]中的信息传播不同,本文模型不仅利用了图的邻近结构,而且指定了不同重要性的邻居,建立了图卷积网络的结构,实现了沿高阶连通性的递归嵌入传播[28]。此外,利用图注意力网络的思想,得到了级联传播的注意权值,揭示了这种连通性的重要性。给定一个候选实体h,对于实体h的多模态邻接实体,将这些信息聚合到实体h,以丰富实体h的表示。$\mathcal{N}_h = \{(h,r,t) | (h,r,t) \in \mathcal{G}\}$表示与h直接相连的三元组集合。$e_A$表示一个聚合邻居实体信息的表示向量,它是每个三元组表示的线性组合,可在公式(3)中计算:

$$e_A = \sum_{(h,r,t) \in \mathcal{N}_h} \pi(h,r,t) e(h,r,t) \tag{3}$$

其中$e(h, r, t)$是每个三元组的嵌入,而$\pi(h, r, t)$是每个三元组$e(h, r, t)$的注意力分数,$\pi(h, r, t)$控制传播的三元组(h, r, t)上每次传播的衰减因子,表示在关系r的条件下从t传播到h的信息量。

保留$e(h, r, t)$和$\pi(h, r, t)$中的关系嵌入,其中的参数是可学习的。对于三元组$e(h, r, t)$,通过对head entity, tail entity 和 relation的嵌入串联进行线性变换来学习这种嵌入,表达式为:

$$e(h,r,t) = W_a (W_r e_h \| e_r \| W_r e_t) \tag{4}$$

其中W_a是权重矩阵,是可训练的模型参数,||表示串联操作。通过关系注意力机制实现$\pi(h,r,t)$,计算公式如下:

$$\pi(h,r,t) = LeakyReLU(W_b e(h,r,t)) \tag{5}$$

其中，W_b是一个可训练的权重矩阵。

按照GATs[26]中的方法选择LeakyReLU[29]作为非线性激活函数。采用softmax函数对与h相连的所有三元组的系数进行归一化：

$$\pi(h,r,t) = \frac{exp(\pi(h,r,t))}{\sum_{(h,r',t')\in\mathcal{N}_h} exp(\pi(h,r',t'))} \tag{6}$$

注意力评分能够建议给予哪个邻居节点更多的注意力来捕获协同信号。当执行正向传播时，注意力流建议关注数据的某些部分，这些部分可以作为推荐背后的解释。

为了不丢失初始e_h信息，这一阶段将实体表示e_h和对应的e_A聚合为实体h的新表示。在本文中通过以下两种方法实现聚合函数$f(e_h,e_A)$。

(1)使用线性转换将e_h和e_A连接在一起，由公式(6)可以得到：

$$f_C = W_c(e_h \| e_A) \tag{7}$$

其中||表示串联操作，W_c是可训练的模型参数。

(2)考虑了e_h和e_A之间的基于元素的add特征交互，由公式(7)可得：

$$f_A = W_d e_h + e_A \tag{8}$$

其中W_d是一个权重矩阵，用来将当前表示转移到一个常见空间，是可训练的模型参数。参考残差网络[31]的流程，也选择在初始的e_h上进行线性转换。

为了实现高阶传播，通过叠加更多的传播和聚合层，进一步探索协同知识图谱中固有的高阶连通性。通常，对于n-layer模型，传入的信息是在n-hop附近积累的。

3.5 模型预测

每个实体通过知识图谱嵌入模块得到相应的嵌入，将其输入到推荐模块。与知识图谱嵌入模块类似，推荐模块也使用MKGs注意层聚合邻居实体信息。

为了保留1-n跳信息，遵循Sun等的多模态知识图谱推荐系统[28]的设置，保留了l层的候选user和item输出。不同层的输出代表不同跳的信息。采用层聚合机制[32]，将每一步的表示连接成单个向量，可以得到：

$$u_e^* = u_e^{(0)} \| \cdots \| u_e^{(L)}, i_e^* = i_e^{(0)} \| \cdots \| i_e^{(L)} \tag{9}$$

其中||为串联操作，L为MKGs注意层数。这样不仅可以通过执行嵌入传播操作来丰富初始嵌入，而且可以通过调整L来控制传播强度。

最后，通过式(10)计算user和item内积，预测匹配分数：

$$\hat{y}(u,i) = u_e^{*\mathrm{T}} i_e^* \tag{10}$$

3.6 最优化

为了进一步优化推荐模型，提高推荐的准确度，使用BPR损失函数(Bayesian Personalized Ranking Loss，BPR)[33]优化推荐预测损失。假设指示更多用户偏好观察到的记录比没观察到的记录被赋值更高的预测分数。目标函数L_{all}的计算如式(11)、(12)所示：

$$L_{CF} = \sum_{(u,i,j)\in\mathcal{O}} - \ln\sigma\left(\hat{y}(u,i) - \hat{y}(u,j)\right) \tag{11}$$

$$L_{all} = L_{KG} + L_{CF} + \lambda\|\Theta\|_2^2 \tag{12}$$

其中$\mathcal{O}=\{(u,i,j)|(u,i)\in\mathcal{R}^+,(u,j)\in\mathcal{R}^-\}$表示训练集，$\mathcal{R}^+$表示用户$u$和项目$i$之间被观察到的交互，$\mathcal{R}^-$是采样的没观察到的交互集，$\sigma(.)$是sigmoid函数。$\Theta$是模型的参数集，$\lambda$是L2归一化的参数。

交替更新MKGs嵌入模块和推荐模块中的参数，采用小批量Adam优化器[34]对嵌入损耗和预测损耗进行优化。Adam优化器是一种应用广泛的优化算法，能够自适应地控制学习速率和梯度的绝对值。特别地，对于随机抽样的一个$batch(h,r,t,t')$，更新所有实体的知识图谱嵌入，然后随机抽取一个$batch(u,i,j)$，从知识图谱嵌入中检索，对两个模块的损失函数进行交替优化。

4 实验与分析

4.1 实验设置

4.1.1 数据集

用于新闻推荐的数据集目前已有很多，但大多只有新闻文本而不包含图片(如MIND[35])。在数据集相对较多的英文新闻推荐领域，虽然也有一些开源的公共数据集包含新闻图片，如addressa[36]，但这些数据集中新闻与图片的比例很小，且许多下载新闻图片的url目前无法使用。此外，目前还没有公开的多模态中文新闻数据集。

由于目前还没有包含多模式新闻信息的高质量数据集，本文自主构建了两个可用的数据集。在英文新闻数据集方面，对MIND数据集进行了处理，通过爬取数据集中每条新闻对应的url，删除无法访问的新闻后，为可用的新闻添加对应的图片url，构建了英文数据集。此外，基于新浪新闻网(2021年9月1日至

2021年9月20日)的数据,收集了部分日志,构建中文新闻数据集,其中第一周的日志用于构建用户历史,其余会话用于形成交互示例。

参考MovieLens数据集的知识图谱的构建方法[37]为数据集构建知识图谱,两个数据集的具体数据如表1中所示:

表1 数据集统计表

数据集	英文新闻	中文新闻
# 用户	805, 298	513, 627
# 新闻	121, 029	157, 464
# 交互	19, 565, 446	13, 878, 984
# 实体	1, 175, 852	1, 229, 256
# 关系	20	17
# triplets	8, 401, 955	7, 619, 827

随机选择每个用户80%的交互历史构成训练集,剩余的作为测试集。从训练集中,随机选择10%的交互作为验证集优化超参数。对于每个观察到的用户项交互,将其视为一个正实例,然后执行负抽样策略,将其与用户以前未选择的负项配对。

4.1.2 评价标准

对于测试集中的每个用户,将未与用户交互的项视为负项,每个方法输出用户对所有项目(除了训练集中的正项)的偏好得分。随机选择20%的交互作为测试的有效值,其余的交互作为训练。为了评价top-k推荐和偏好排名的有效性,采用两个常用的评价指标[38-39]:$recall@k$和归一化折损累计增益(Normalized Discounted Cumulative Gain, NDCG)$ndcg@k$。公式表示为:

$$Recall@k = \frac{TP@k}{TP@k + FN@k} \tag{13}$$

其中$TP@k$表示真正例,$FN@k$表示假负例(False Negative)。

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{14}$$

其中

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \tag{15}$$

rel_i指第i个结果的真实相关性分数。

$$IDCG@k = \sum_{i=1}^{|REL|} \frac{rel_i}{\log_2(i+1)} \tag{16}$$

$IDCG$(Ideal DCG)为理想的DCG。$|REL|$表示结果按照真实相关性从大到小排序,取前k个结果,k默认值为20。

4.1.3 基线

将MKGAT模型与一些较高效的基线进行比较,包括基于FM(Factorization Machines)的方法(Neural Factorization Machines, NFM)、基于KG的方法(CKE、KGAT)、多模态方法(MMGCN、MKGAT)。

神经分解机(NFM)[40]是最先进的因子分解机器(FM),它将FM纳入神经网络。特别是,本文按照基于递归神经网络的时间异构反馈推荐[4]中的建议,在输入特性上使用一个隐藏层。

基于嵌入的协同知识(Collaborative Knowledge Base Embedding, CKE)[19]将CF与结构化知识、文本知识和可视化知识结合在统一的推荐框架中,本文将CKE实现为CF+结构化知识模块。

知识图谱注意力网络(Knowledge Graph Attention Network, KGAT)[9]首先应用TransR模型[13]获得实体的初始表示,然后从实体本身向外运行实体传播。因此用户表示和项表示可以利用相应的邻居信息。

多模态图卷积网络(Multi-modal Graph Convolution Network, MMGCN)[41]是一种多模式模型,考虑每个模式的用户-项目交互。MMGCN为每个模态构建用户-项目二分图,然后使用GCN对每个二分图进行训练,合并不同模态的节点信息。

多模态知识图谱注意力网络(Multi-modal Knowledge Graphs Attention Network, MKGAT)[28]利用Resnet50模型和SIF模型分别将图像和文本转化为相应的向量,然后使用transE模型学习知识图谱的结构化表示。其与KGAT类似,都是从实体本身向外运行实体传播,将相邻实体的信息进行聚合。

4.1.4 参数设置

本文使用Xavier初始化器[42]初始化模型参数,并使用Adam优化器[34]优化模型。mini-batch大小和学习率分别在[1024;5120;10240]和[0:0001;0:0005;0:001]中选取。对于NFM和KGAT,丢失率在{0.0, 0.1, ⋯, 0.8}中调整。对于MKGAT,其视觉实体使用Resnet最后一层隐藏层的2048维特征,同时文本实体使用word2vec训练300维的词嵌入,并使用SIF算法生成相应的句子向量。本文的VMKGAT微调了ViLBERT的最后三个层,并将负抽样比设为4。MKGAT和VMKGAT的所有实体维度设置为64。

4.2 实验结果

4.2.1 模型表现

所有模型的实验结果如表2所示。VMKGAT模

型(其中的模式包括结构化知识、文本和视觉;模型深度设置为2;组合层设置为Add聚合层)的*recall*和*ndcg*在两个数据集上都优于所有基线。

VMKGAT始终在两个数据集上最佳。特别是VMKGAT在英文新闻数据集和中文新闻数据集中分别比最强的基于KG的单模态基线KGAT在ndcg@20上提高了13.11%和13.23%,在recall@20上提高了12.98%和11.73%,由此验证了多模式知识图谱的有效性。结合表2和表3,在引入多模态实体的情况下,本方法比其他基于KG的方法有更大的改进。这验证了本文方法比其他方法对多模态信息更友好。

在所有的比较方法中,基于KG的方法(即CKE和KGAT)在两个数据集上优于基于CF的方法(即NFM),这表明使用KG确实大大提高了推荐性能。

比较两种基于KG的方法CKE和KGAT的性能,发现KGAT在两种指标上都比CKE有更好的性能,由此可见图卷积网络在推荐系统中的强大作用。

值得一提的是,VMKGAT在两个数据集上的指标都比MKGAT模型提高了2%到3%。这表明本文方法可以更加合理地利用多模态信息。

表2 不同推荐模型的总体表现

模型	英文新闻		中文新闻	
	recall	ndcg	recall	ndcg
NFM	0.3213	0.2951	0.3165	0.2923
CKE	0.3454	0.3096	0.3323	0.3025
KGAT	0.3652	0.3203	0.3459	0.3166
MMGCN	0.3695	0.3229	0.3531	0.3202
MKGAT	0.4018	0.3520	0.3778	0.3480
VMKGAT	**0.4126**	**0.3623**	**0.3865**	**0.3585**
%Improv.	2.6%	2.9%	2.3%	3.0%

4.2.2 多模态特征的影响

为了探讨不同模式的影响,比较了KGAT、MKGAT和VMKGAT模型在英文新闻数据集上不同模式的结果。性能对比结果见表3。

KGAT、MKGAT和VMKGAT多模态特征的方法普遍优于单模态特征的方法,如表3所示。

VMKGAT模型本质上也是一种基于KG的方法,与MKGAT相比,VMKGAT模型可以更好地利用图像信息以提高推荐性能。与其他基于KG的方法相比,当引入多模态信息时,方法会有更大的改进。这是因为在训练知识图谱嵌入时,VMKGAT能更好地将图像实体的信息聚合成item实体,如表3所示。

表3 推荐的性能:不同模式对英语新闻数据集的影响

模型	KGAT		MKGAT		VMKGAT	
	recall	ndcg	recall	ndcg	recall	ndcg
基本	0.3652	0.3203	0.3675	0.3231	0.3719	0.3254
基本+文本	0.3709	0.3271	0.3882	0.3424	0.3958	0.3490
%提升	1.6%	2.1%	5.6%	6.0%	6.4%	7.3%
基本+图像	0.3771	0.3329	0.3999	0.3501	0.4055	0.3565
%提升	3.5%	3.9%	8.8%	8.4%	9.0%	9.6%
基本+文本+图像	0.3817	0.3359	0.4018	0.3520	0.4126	0.3623
%提升	4.5%	4.9 %	9.3%	9.0%	10.9%	11.3%

4.2.3 模型深度的影响

改变VMKGAT的深度(L)研究多个嵌入传播层的使用效率。具体来说,在{1,2,3,4}范围内搜索层数,VMKGAT1表示使用一层模型,将结果汇总在表4中,有以下观察结果:

增加VMKGAT深度能够大幅提升性能。显然,VMKGAT2和VMKGAT3在所有方面都比VMKGAT1实现了持续的改进。这种改进归功于对用户、物品和实体之间的高阶关系的有效建模,这些关系分别由二阶连接和三阶连接承载。

在VMKGAT3上再叠加一层,观察到VMKGAT4只实现了边际改进。这表明考虑实体之间的三阶关系就足以捕获协作信号。同时可以发现,当两个数据集中的层数大于3时,评价指标会下降。即当层的数量增加到某个级别时,评估指标就会下降。这可能是数据的稀疏性导致的过拟合。

联合分析表2和表4,多数情况下,VMKGAT1始终优于其他基线。再次验证了该方法的有效性,经验表明该方法能较好地模拟一阶关系。

表4 嵌入传播层数的影响

模型	英文新闻		中文新闻	
	recall	ndcg	recall	ndcg
VMKGAT1	0.4015	0.3603	0.3810	0.3565
VMKGAT2	0.4126	0.3623	0.3865	0.3581
VMKGAT3	0.4133	0.3645	0.3877	0.3589
VMKGAT4	0.4130	0.3643	0.3871	0.3580

4.2.4 聚合层的影响

本节研究聚合层在模型中的作用,使用两种类型的聚合层,即Add层和Concatenate层来学习知识图谱的嵌入。模型深度固定为2。实验结果如表5所示,CONCAT标记的层连接方法优于ADD标记的层连接方法。一个可能的原因是,每个实体的相邻实体包含文本和可视化信息,这些信息与知识图中的一般实体

是异构的,它们不在同一个语义空间中。ADD实际上是一种元素与元素的特征交互方法,适用于相同语义空间中的特征。因为在相同的语义空间中,每个特征的每个维度的含义相同,所以把每个特征的每个维度相加是有意义的。而CONCAT是特征之间维度的扩展,更适合不同语义空间中特征的交互。

表5 聚合层的影响

聚合方式	英文新闻		中文新闻	
	recall	ndcg	recall	ndcg
ADD	0.4126	0.3623	0.3865	0.3585
CONCAT	0.4139	0.3688	0.3891	0.3602

5 结论

本文提出了一种基于知识图谱的推荐模型——结合ViLBET的多模态知识图谱注意网络(VMKGAT),在新闻推荐系统中引入了ViLBERT多模态预训练模型和多模态知识图谱组合的模型。VMKGAT模型通过学习实体之间的推理关系,将每个实体的相邻实体信息聚合到自身,可以更好地利用多模态实体信息。在两个真实数据集上的大量实验证明了VMKGAT模型的合理性和有效性。

本文对多模态知识图谱在推荐系统中的应用进行了初步探索,并在此基础上进行了进一步的研究。未来在多模态知识图谱的框架下,探索更多的多模态融合方式。

参考文献(References):

[1] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70.

[2] Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008: 426-434.

[3] Salakhutdinov R, Mnih A, Hinton G. Restricted Boltzmann machines for collaborative filtering[C]//24th International Conference on Machine Learning, 2007: 791-798.

[4] Wu C, Wang J, Liu J, et al. Recurrent neural network based recommendation for time heterogeneous feedback[J]. Knowledge-Based Systems, 2016, 109: 90-103.

[5] Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations[C]//Proceedings of the 10th ACM Conference on Recommender Systems, 2016: 191-198.

[6] Sun Z, Yang J, Zhang J, et al. Recurrent knowledge graph embedding for effective recommendation[C]//Proceedings of the 12th ACM Conference on Recommender Systems, 2018: 297-305.

[7] Wang H, Zhang F, Wang J, et al. Ripplenet: propagating user preferences on the knowledge graph for recommender systems[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018: 417-426.

[8] Wang H, Zhang F, Xie X, et al. DKN: deep knowledge-aware network for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference, 2018: 1835-1844.

[9] 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.

[10] Mousselly-Sergieh H, Botschen T, Gurevych I, et al. A multimodal translation-based approach for knowledge graph representation learning[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, 2018: 225-234.

[11] Xie R, Liu Z, Luan H, et al. Image-embodied knowledge representation learning[DB/OL]. arXiv: 1609.07028, 2016.

[12] Pezeshkpour P, Chen L, Singh S. Embedding multimodal relational data for knowledge base completion[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3208-3218.

[13] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//29th AAAI Conference on Artificial Intelligence, 2015: 2181-2187.

[14] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//IEEE International Conference on Computer Vision(ICCV), 2017.

[15] Lu J, Batra D, Parikh D, et al. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[DB/OL]. arXiv: 1908.02265, 2019.

[16] Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019: 4171-4186.

[17] Wang H, Zhang F, Hou M, et al. Shine: Signed heterogeneous information network embedding for sentiment link prediction[C]//ACM International Conference on Web Search and Data Mining, 2018: 592-600.

[18] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[19] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge

base embedding for recommender systems[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2016：353-362.

[20] Wang H，Zhang F，Zhao M，et al. Multi-task feature learning for knowledge graph enhanced recommendation[C]// The World Wide Web Conference，2019：2000-2010.

[21] Yu X，Ren X，Sun Y，et al. Personalized entity recommendation：a heterogeneous information network approach[C]// Proceedings of the 7th ACM International Conference on Web Search and Data Mining，2014：283-292.

[22] Zhao H，Yao Q，Li J，et al. Meta-graph based recommendation fusion over heterogeneous information networks[C]//23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，2017：635-644.

[23] Li Q，Tang X，Wang T，et al. Unifying task-oriented knowledge graph learning and recommendation[J]. IEEE Access，2019，7：115816-115828.

[24] Liu Y，Li H，Garcia-Duran A，et al. MMKG：multi-modal knowledge graphs[C]//European Semantic Web Conference，2019：459-474.

[25] Kipf T N，Welling M. Semi-supervised classification with graph convolutional networks[C]//International Conference on Learning Representations（ICLR），2017.

[26] Veličković P，Cucurull G，Casanova A，et al. Graph attention networks[DB/OL]. arXiv：1710.10903，2017.

[27] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[C]//31st International Conference on Neural Information Processing Systems，2017：6000-6010.

[28] 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.

[29] Maas A L，Hannun A Y，Ng A Y. Rectifier nonlinearities improve neural network acoustic models[C]//ICML Workshop on Deep Learning for Audio，Speech and Language Processing，2013.

[30] Hamilton W L，Ying R，Leskovec J. Inductive representation learning on large graphs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems，2017：1025-1035.

[31] 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.

[32] Xu K，Li C，Tian Y，et al. Representation learning on graphs with jumping knowledge networks[C]//35th International Conference on Machine Learning，2018.

[33] Rendle S，Freudenthaler C，Gantner Z，et al. BPR：bayesian personalized ranking from implicit feedback[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial IntelligenceJune，2009：452-461.

[34] Kingma D P，Ba J. Adam：a Method for Stochastic Optimization[DB/OL]. arXiv：1412.6980，2014.

[35] Wu F，Qiao Y，Chen J-H，et al. Mind：a large-scale dataset for news recommendation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：3597-3606.

[36] Gulla，J A，Zhang L，Liu P，et al. The adressa dataset for news recommendation[C]//Proceedings of the International Conference on Web Intelligence，2017：1042-1048.

[37] Wang H，Zhang F，Zhang M，et al. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，2019：968-977.

[38] He X，Liao L，Zhang H，et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web，2017：173-182.

[39] Yang J-H，Chen C-M，Wang C-J，et al. HOP-rec：high-order proximity for implicit recommendation[C]//12th ACM Conference on Recommender Systems，2018：140-144.

[40] He X，Chua T-S. Neural factorization machines for sparse predictive analytics[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval，2017：355-364.

[41] 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.

[42] Glorot X，Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//13th International Conference on Artificial Intelligence and Statistics，2010.

编辑：王谦