计算机科学
COMPUTER SCIENCE

图像美学信息增强的视觉感知推荐系统

张凯煊, 蔡国永, 朱琨日

引用本文

张凯煊, 蔡国永, 朱琨日. 图像美学信息增强的视觉感知推荐系统[J]. 计算机科学, 2023, 50(11A): 221100083-8.

ZHANG Kaixuan, CAI Guoyong, ZHU Kunri. Image Aesthetics-enhanced Visual Perception Recommendation System [J]. Computer Science, 2023, 50(11A): 221100083-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100083

图像美学信息增强的视觉感知推荐系统

张凯煊[1]　蔡国永[2]　朱琨日[2]
1 桂林电子科技大学计算机与信息安全学院　广西 桂林 541000
2 广西可信软件重点实验室(桂林电子科技大学)　广西 桂林 541000
(devinzkx@163.com)

摘　要　视觉感知推荐系统旨在从视觉认知角度出发,通过提取物品图像的视觉特征来增强用户和物品交互的行为特征,建模用户视觉与行为相关的偏好,从而更好地进行推荐。已有的视觉感知推荐研究中,通常使用预训练的卷积神经网络(CNN)来提取视觉对象语义特征,很少考虑物品外观图像内部隐藏的美学风格特征;其次,在视觉感知推荐中用户和物品的交互行为结构嵌入信息通常被忽视。为了解决这些问题,提出了一个融合图像美学和行为交互结构嵌入的美学特征感知视觉推荐系统(ABVR)。ABVR使用预训练ViT模型提取图像的高层视觉特征——语义类别特征,利用美学提取网络挖掘出图像中的中层美学视觉特征——物品的颜色、形状等特征,利用图卷积神经网络(GCN)模块学习用户物品交互图结点的多层图结构嵌入特征,最后将3类特征关联融合,以实现美学增强的视觉推荐。在两个真实数据集上进行了大量实验,验证了ABVR模型在视觉推荐性能提升上的有效性。

关键词:视觉感知;美学特征;视觉推荐;图卷积神经网络
中图法分类号　TP391

Image Aesthetics-enhanced Visual Perception Recommendation System

ZHANG Kaixuan[1],CAI Guoyong[2] and ZHU Kunri[2]
1 College of Computer and Information Security,Guilin University of Electronic Technology,Guilin,Guangxi 541000,China
2 Key Laboratory of Guangxi Trusted Software(Guilin University of Electronic Technology),Guilin,Guangxi 541000,China

Abstract　The visual perception recommendation system aims to enhance the behavioral features of user-item interaction by extracting the visual features of item images from the perspective of visual cognition,and model the user's visual and behavior-related preferences,so as to make better recommendations. In the existing visual perception recommendation research,pre-trained convolutional neural network(CNN) is usually used to extract the semantic features of visual objects,and the hidden aesthetic style features inside the appearance image of the item are rarely considered. In addition,the embedded information of user-item interaction behavior structure is usually ignored in visual perception recommendation. To address these issues,an aesthetic feature-aware visual recommendation system is proposed that fuses image aesthetics and behavioral interaction structure embeddings(ABVR). ABVR uses the pre-trained ViT model to extract the high-level visual features of the image—semantic category features,uses the aesthetic extraction network to mine the middle-level aesthetic visual features in the image —the color,shapes and other features of the items,and uses the graph convolution neural network(GCN) module to learn the multi-layer graph structure embedding features of user item interaction graph nodes,and finally associates and fuses the three types of features to achieve aesthetically enhanced visual recommendations. Extensive experiments are conducted on two real datasets to verify the effectiveness of the ABVR model in improving visual recommendation performance.

Keywords　Visual perception,Aesthetic features,Visual recommendation,Graph convolutional neural networks

1　引言

推荐系统作为一种智能化筛选算法,旨在减少用户决策焦虑,提供个性化智能推荐服务。典型的协同过滤推荐算法利用用户和物品交互的历史行为,推断用户对物品的偏好,从而做出推荐决策。然而在实际场景下,影响用户决策的因素是多方面的。如视觉特征对人们的购物决策有着重要影响。特别是在食品、旅游、时尚等领域的推荐,图像轮廓、纹理、颜色等具有美学风格的视觉特征均会影响用户判断,吸引用户注意力,激发情感,进而影响对推荐物品的点击和购买,因此视觉推荐应运而生[1-2]。随着深度图神经网络学习的成功应用,它也被越来越多地应用到推荐领域中建模交互行为结构[3-5]。

尽管视觉感知推荐研究取得了很多进展,但其研究仍

基金项目:国家自然科学基金(61763007);广西可信软件重点实验室项目(kx202060)
This work was supported by the National Natural Science Foundation of China(61763007) and Guangxi Key Lab of Trusted Software(kx202060).
通信作者:蔡国永(ccgycai@guet.edu.cn)

存在不足之处[6-7]。(1)视觉感知推荐中常用 CNN 提取视觉语义特征,对视觉美学特征的利用不足,且它无法有效保留输入图像的空间关联信息;(2)视觉感知推荐缺少对交互行为结构信息建模,另外图神经网络推荐中对图结点表示单一、不能很好地处理多源异构数据。针对上述问题,本文将视觉的高层语义特征(如图片的类别特征)、中层美学特征(如颜色、形状特征)和用户-物品历史交互图结构嵌入特征(用户物品交互图结点的 id 标识及关联信息嵌入)共同纳入用户对物品的偏好得分进行建模。

本文第 2 章简要总结了推荐系统相关研究工作;第 3 章介绍了本文提出的模型架构;第 4 章为实验及评估分析;最后总结全文。

2 相关工作

推荐系统早期工作中,主要从非视觉角度来考虑,从用户或物品的隐藏信息中挖掘出内在关联性,如矩阵分解 MF。非视觉推荐系统中主要研究隐式反馈中用户和物品的隐因子对决策的影响。例如 Simon Funk 提出的奇异值分解 SVD 方法,将评分矩阵分解为两个低维矩阵相乘,用较小的数据来表示原始数据,去除相关冗余信息,计算隐空间下的相似度,以此提高推荐效果。但传统的 MF 模型已被证明容易过拟合。Rendle 等在 MF 基础上提出了一种更加有效的隐式反馈贝叶斯个性化排序 BPR 算法[8],利用成对排序方法(Pairwise)对三元组 (u,i,j) 数据进行处理,使用户 u 对积极物品 i 的偏好得分远远大于对消极物品 j 的偏好得分。最后进行贝叶斯分析得到最大后验概率来求解目标函数,完成对物品的排序推荐。

非视觉推荐领域中,图神经网络推荐也受到研究者们的极大关注。图神经网络模型既适用于侧重图的任务,也适用于侧重结点的任务。Wang 等提出了侧重于结点任务的图神经协同过滤 NGCF 算法[9],将用户和物品的历史交互信息构建成二部图,从消息构建(邻居消息构建+自连接消息构建+用户与物品交互行为消息构建)到消息传播(前一层的所有嵌入消息汇聚到新一层),从而捕获图结点上的协同信号。最后通过嵌入传播公式进行多层嵌入拼接融合得到最终偏好得分,一并应用到 BPR 模型中进行训练和推荐。He 等认为,NGCF 的训练时间长,主要因为是它的非线性激活、特征转换等神经网络操作,单纯针对 id 嵌入推荐任务来说,训练成本较大。因此提出了一种简化的轻量级图卷积网络 LightGCN[10],去除掉非线性激活和特征转换,聚合前一层的邻居信息进行消息传播,并将所有层的嵌入表示进行层组合加权和,缓解 GCN 的过度平滑问题。总体而言,Light GCN 更适合单纯的 id 嵌入,而 NGCF 更适合处理复杂的非结构化多模态信息数据。

在视觉推荐领域,参考仿生学,利用图片中视觉信息来建模。相较于非视觉推荐系统,视觉感知推荐能够更加有效地捕获用户对物品图像的情绪反应,并且能够在冷启动的情况下,作为辅助信息补充用户交互信息进行推荐。He 等在 BPR 算法的基础上进行了重要改进,将视觉信号引入用户的得分预测器中,提出了视觉贝叶斯个性化排序 VBPR 算法[6]。基于大型 imagenet 数据集预训练的深度卷积神经网络(CNN)提取物品图像视觉特征,起初主要是针对图像分类任务,但迁移至视觉推荐任务中仍表现较好。随后 Tang 等发现 VBPR 等视觉推荐系统不健壮的问题,对输入图像进行细小扰动将严重影响推荐精度,因而提出一种对抗性多媒体视觉推荐 AMR[11],进行对抗性训练,以增强鲁棒性和推荐精度,达到了很好的效果。Niu 等认为,传统的 BPR 中用户偏好得分计算使用用户和物品潜在向量的内积,属于赋予隐空间内每个维度相同的权重,无法捕获可变的用户偏好。其次 BPR 的本质是线性的,相较于非线性方法,表达能力不足。因此在神经协同过滤 NCF[12]的启发下,结合广义矩阵分解的思想,提出一种基于神经网络的隐式反馈个性化成对排序模型 NPR。后续又引入预训练的 CNN 提取图像特征,并使用 PCA 降维最后输入 NPR 中,得到最终视觉神经个性化排序 VNPR[13]。由于上述方法均是在物品部分建模,Chen 等提出一种新颖的注意力协同过滤 ACF 方法[7]则是对用户部分建模。针对用户历史交互物品图像的区域特征和对整体物品特征施加不同的双层注意力来加权计算最终用户的偏好得分。该方法和之前工作最大的不同之处在于,它采用卷积层的特征图(Feature Map)来提取图像特征,而不是利用以往视觉推荐中使用的全连接层来提取图像特征。

在视觉推荐领域,图像本身还蕴含着丰富的内容,如场景信息、对象信息、情感信息,以及美学信息等。从美学角度考虑,每个人的审美角度不同,个性化特点非常鲜明,因此可以从中挖掘出人类对视觉情感感知的可解释性。由于美学时尚风格高度的主观性和特征的复杂性[14],构建时尚领域的视觉推荐系统极具挑战性。Kang 等提出的深度视觉感知贝叶斯个性化排序 DVBPR[15]算法运用了对比学习中的 Siamese CNN 建模,并且使用自训练 CNN 来获取图像表示,以达到时尚推荐的目的。此外 DVBPR 模型还拥有 GAN 图像生成对抗网络模块,通过给定用户物品类别,根据用户个人历史记录生成最符合用户品味的时尚服装。因此,DVBPR 算法不仅可以从产品数据库中推荐现有物品,还可以设计具有美学风格的新产品。

传统的视觉推荐方法是在一个共同的视觉特征空间中对物品进行建模。这些视觉表示通常只能捕获图像的类别信息,而无法捕获物品的风格样式信息,因此,Liu 等提出了关于风格样式的视觉推荐 Deepstyle[16]。它的独特之处在于认定物品由风格样式和类别组成。使用预训练的 CNN 提取到的视觉特征减去图像类别特征得到物品的风格样式特征。单独对这种风格样式特征建模,得到用户对美学风格样式特征的偏好得分,最后借助 BPR 框架求目标函数最优解进行视觉推荐。受人类视觉感知和神经美学的影响,Wang 等研究了一种脑启发式深度 BDN 网络[17],主要思路是先无监督训练 4 层 SCAE 的前两层,并将前两层卷积初始化应用到所有并行通道的 FCNN 的 conv1 层和 conv2 层。初始化前两层后,对每个路径的 conv3 和 conv4 连接起来平均池化进行细分美学标签监督训练。之后将 conv4 和分类器丢弃,输入图片后将 conv3 层提取不同并行通道上的各种美学特征,如饱和度、颜色、双色调、三分法等。最后使用深度融合网络对所有美学特征融合汇总。Paul 等在 BDN 的基础上提出了一种融合美学的动态协同过滤 ADCFA 模型[18]。利用张量因子分解作为基础模型,以用户、物品和时间 3 个维度描述购买事件,然后将 BDN 生成的美学特征纳入其中进行训练,从而捕获到不同

消费者在不同时期的审美偏好差异。

虽然推荐领域之前的研究取得的成果令人印象深刻，但在视觉感知推荐中缺乏对交互图结构嵌入的运用。另外推荐中挖掘图像隐藏美学特征的也非常少。同时，利用 CNN 获取图像的视觉特征存在感受野较局限等问题，难以捕捉图像上的全局信息。因此本文设计了 ABVR 模型，将高层视觉特征（ViT 提取的类别特征）与中层美学视觉特征（颜色、形状特征）、用户物品交互结点的图结构嵌入特征相结合，获得更为细粒度的偏好，从而提升整体的推荐效果。

3 ABVR 方法

本章将详细阐述所提方法 ABVR（Aesthetics and Behavioral Interaction Structure Embedding Visual Perception Recommendation），其整体模型如图 1 所示，主要分为 3 部分：特征提取模块、特征融合模块和推荐模块。

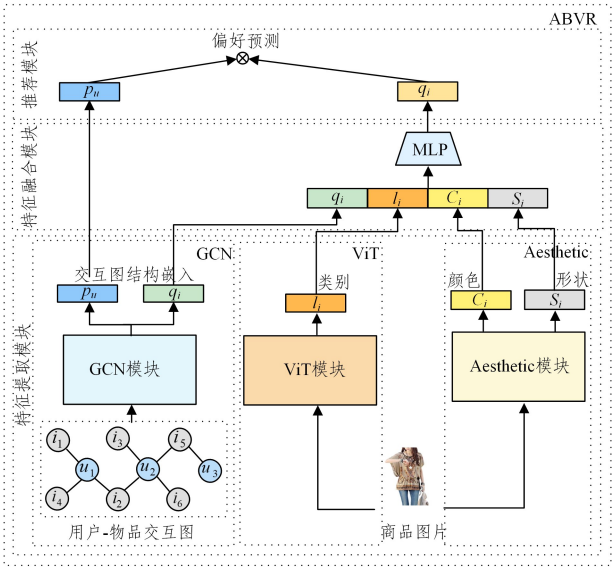


图 1 ABVR 模型整体框架结构图

Fig. 1 Overall framework structure of ABVR model

为了研究物品属性在推荐中起到的作用，设定一个物品属性由 3 部分构成：id 标识信息、类别信息 category、美学信息 aesthetic。基于此设定，通过 ABVR 模型的特征提取模块，分别获取图结构嵌入特征（id 标识及关联信息）、高层视觉（category 类别信息）特征和中层美学视觉（aesthetic 美学信息）特征。

3.1 交互图结构嵌入特征提取 GCN 模块

首先初始化嵌入层的用户/物品 id 标识嵌入，其次沿着用户和物品交互图执行多层嵌入传播，通过注入高阶连接性来完善嵌入信息。这种交互项反映了关联性质用户的历史偏好行为，即用户购买了某个物品就会被视作该物品的特征。最后将不同传播层的嵌入进行拼接，得到经过多层图卷积学习后的用户/物品嵌入特征。GCN（Graph Convolution Neural Network）模块结构如图 2 所示。

初始嵌入层：根据用户-物品交互图对用户/物品的 id 标识进行嵌入，将 id 映射至低维稠密向量中，分别作为用户 u 的初始嵌入向量 $\boldsymbol{e}_u^{(0)} \in \mathbb{R}^d$ 及物品 i 的初始嵌入向量 $\boldsymbol{e}_i^{(0)} \in \mathbb{R}^d$，其中 d 为嵌入维度。交互图中沿着图路径遍历，捕获图结点之间的协同信号，并作为图结构嵌入信息，这是传统视觉感知推荐所不具备的能力。

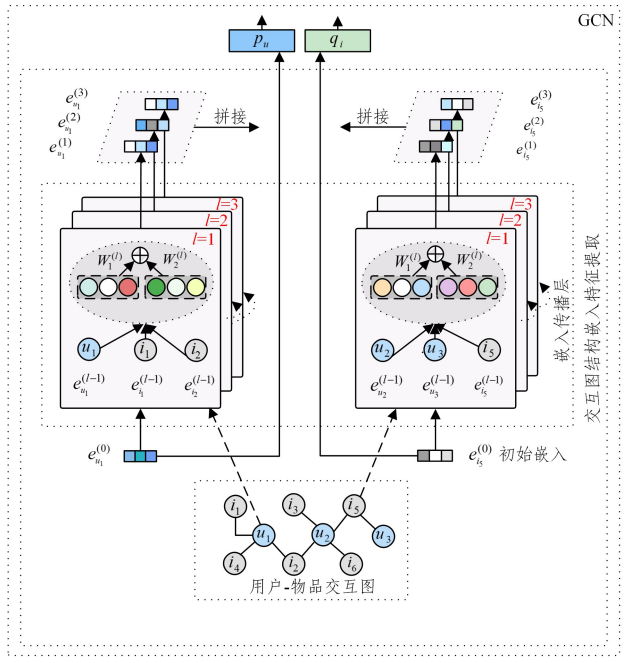


图 2 交互图结构嵌入特征提取的 GCN 模块

Fig. 2 GCN module for interaction graph structure embeddings feature extraction

一阶嵌入传播（$l=1$ 层）：针对某个用户物品对 (u,i)，定义从物品 i 到用户 u 的消息。

$$m_{u\leftarrow i} = f(e_i, e_u, z_{ui}) \tag{1}$$

其中，$f(\cdot)$ 表示消息编码函数，里面使用嵌入 e_i，e_u 作为输入，并且使用图拉普拉斯正则化项 z_{ui} 作为传播过程中的衰减因子。将消息编码函数用式(2)具体化：

$$m_{u\leftarrow i} = \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}(\boldsymbol{W}_1 e_i + \boldsymbol{W}_2(e_i \odot e_u) \tag{2}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 均为可训练的权重矩阵，$\odot$ 为哈德玛积，$\frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}$ 为图拉普拉斯正则化项，即 z_{ui}。$\mathcal{N}_u$ 与 $\mathcal{N}_i$ 分别表示用户 u 交互的物品集和物品 i 交互的用户集。

将之前每个目标结点构建的邻居消息加权和，并计算目标结点的自连接消息，经过非线性激活函数得到首层的嵌入表达：

$$e_u^{(1)} = LeakyRelu(m_{u\leftarrow u} + \sum_{i\in N_u} m_{u\leftarrow i}) \tag{3}$$

高阶嵌入传播（第 $l=2, l=3$ 层）：第一部分是新的邻居消息构建，即前一层的邻居嵌入表达与用户物品交互信息分别使用本层不同的权重加权和。最后运用 z_{ui} 进行约束得到 $m_{u\leftarrow i}^{(l)}$。第二部分是对目标结点自身的消息构建，为前一层嵌入表达与本层权重之积得到 $m_{u\leftarrow u}^{(l)}$：

$$m_{u\leftarrow i}^{(l)} = \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}}(\boldsymbol{W}_1^{(l)} e_i^{(l-1)} + \boldsymbol{W}_2^{(l)}(e_i^{(l-1)} \odot e_u^{(l-1)})) \tag{4}$$

$$m_{u\leftarrow u}^{(l)} = \boldsymbol{W}_1^{(l)} e_u^{(l-1)} \tag{5}$$

高阶传播类似于一阶传播，将目标结点之前构建的邻居消息加权和，并计算目标结点自连接消息，最后经过非线性激活函数得到第 l 层的嵌入表达。

$$e_u^{(l)} = LeakyRelu(m_{u\leftarrow u}^{(l)} + \sum_{i\in N_u} m_{u\leftarrow i}^{(l)}) \tag{6}$$

类似地，对物品进行一阶嵌入传播再到高阶嵌入传播，得到物品 i 在第 l 层的嵌入表达 $e_i^{(l)}$。

最后将所有层的用户/物品的嵌入表示级联拼接作为最终用户/物品结点的交互图结构嵌入特征向量。

$$p_u = e_u^{(0)} \| \cdots \| e_u^{(L)} \tag{7}$$

$$q_i = e_i^{(0)} \| \cdots \| e_i^{(L)} \tag{8}$$

3.2 高层视觉特征提取 ViT 模块

物品图像的高层视觉特征提取采用 timm 库中的预训练 ViT(Vision Transformer)模型，主要是因为 ViT 模型相比 CNN 更能捕捉图像空间上的全局信息。自注意力使得全局信息得以在早期聚合，残差连接能将特征从底层传播到较高层，因而更适合用于视觉特征提取和迁移下游任务。另外传统 Transformer Encoder 主要针对序列，因此处理图像数据时，需要将图像打成序列状，进而提取图像的高层视觉(类别)特征，ViT 模块如图 3 所示，共分为以下 4 个阶段。

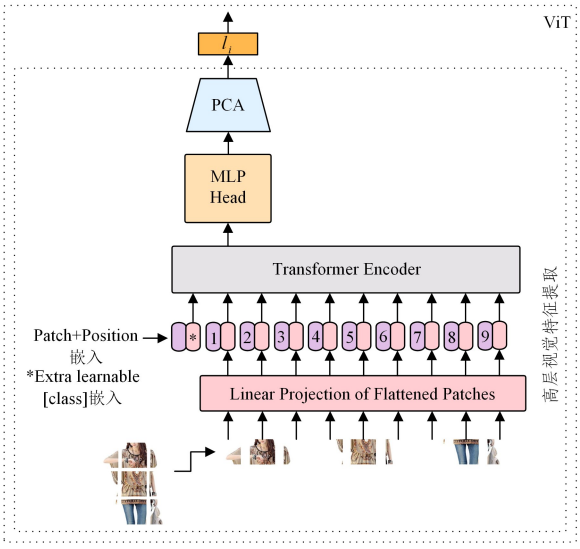


图 3 高层视觉特征提取的 ViT 模块

Fig. 3 ViT module for high level visual feature extraction

Patch 嵌入阶段。首先对输入图片进行微调缩放，统一设置大小为 $224\times224\times3$ 的图片。将图像 $x\in\mathbb{R}^{H\times W\times C}$ 打成固定大小的 N 个 patch($N=196$)，即 x_p^N，每个 Patch 大小为 16×16。将 Patch 输入到线性映层得到最终的 Patch 嵌入表示 P_e：

$$P_e = [x_{\text{class}}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] \tag{9}$$

其中，Class Embedding，即 x_{class}，在图 3 中用 * 表示，用于学习图像的类别信息；$2D$ Patches $x_p\in\mathbb{R}^{N\times(P^2\cdot C)}$，$C$ 是通道数，P 是 Patch 大小，一共有 N 个 Patches，$N=HW/P^2$，$E\in\mathbb{R}^{(P^2\cdot C)\times D}$ 为线性映射层嵌入。

Positional 编码阶段。对每一个 Patch 进行位置索引编码 E_{pos}，防止输入 Encoder 后丢失位置信息。最终得到加入位置编码的初始嵌入 z_0：

$$z_0 = P_e + E_{\text{pos}} \tag{10}$$

Transformer 编码阶段。内部采用 12 层堆叠的 Encoder Block，将初始 z_0 扩展到第 L 层 $z_{\ell-1}$，在内部首先经过 Layer Norm 层，并且使用多头注意力机制 $MSA(\cdot)$ 处理，同时连接一个用于保留原始输入特征的残差结构 $z_{\ell-1}$，进而得到 z_ℓ'。其次在经过第二个 Layer Norm 层后，使用 $MLP(\cdot)$ 进行处理，同时再接一个残差结构 z_ℓ，最终得到 Encoder 的输出 z_ℓ：

$$z_\ell' = MSA(LN(z_{\ell-1})) + z_{\ell-1}, l=1,\cdots,L \tag{11}$$

$$z_\ell = MLP(LN(z_l')) + z_\ell', l=1,\cdots,L \tag{12}$$

MLP 阶段。从前面多层 Encoder Block 的输出结果 z_ℓ 中抽取类别部分的一维特征的 z_L^0，然后送入 Layer Norm 层得到 y，最后使用 $MLP(\cdot)$ 进行进一步处理，得到初始视觉特征 y'：

$$y = LN(z_L^0) \tag{13}$$

$$y' = MLP(y) \tag{14}$$

最后将输出的初始视觉特征 y' 输入到设定的另一个 Encoder 进行降维，这里采用 PCA 对图像的语义类别特征进行编码，得到图像的高层视觉特征向量。

$$l_i = PCA(y') \tag{15}$$

3.3 中层美学视觉特征提取 Aesthetic 模块

物品图像的中层美学视觉特征具备一定可解释性，关于审美有很多特征选项，如颜色、形状、纹理等。但即使是时尚专家也很难定义决定一件物品优劣的具体美学特征因素。本文中采用了一种美学视觉特征提取网络，如图 4 所示，分别从图像中提取颜色 color、形状/纹理 shapes 特征。

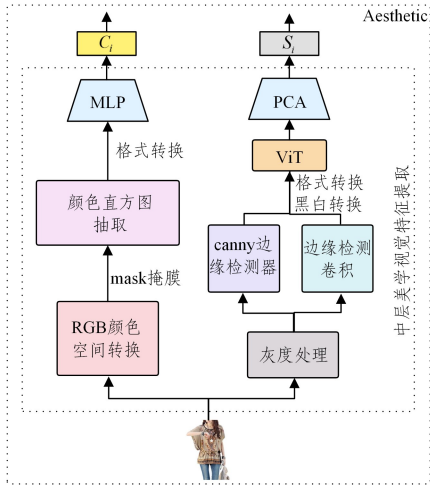


图 4 中层美学视觉特征提取的 Aesthetic 模块

Fig. 4 Aesthetic module for mid-level aesthetic visual feature extraction

颜色 color 部分。首先将输入图片 x 送入美学视觉特征提取器中，通过 8 位的 RGB 颜色直方图获取。具体而言，使用 opencv 库中颜色空间转换函数 $cvtColor(\cdot)$ 将原始图像转换至 RGB 颜色空间内，并找到像素簇如式(16)所示，然后利用掩膜技术 mask 屏蔽掉黑色轮廓线条区域，并利用 $calcHist(\cdot)$ 计算图像经过掩膜后的主要颜色 Hist，如式(17)所示。再将生成出的 Hist 转换为 numpy 格式数据，并使用 Encoder 编码进行降维处理，这里 Encoder 采用多层感知机 MLP-based 进行编码，颜色特征如式(18)所示。最后用降维后的颜色向量 $\boldsymbol{c}_i$ 来表征单个物品的颜色部分的美学视觉特征。其中 $\boldsymbol{c}_i$ 与图卷积 GCN 模块生成的 p_u，q_i 的维度保持一致。

$$h = cvtColor(x, COLOR\_BGR2RGB) \tag{16}$$

$$Hist = calcHist(h, mask) \tag{17}$$

$$\boldsymbol{c}_i = MLP(Hist) \tag{18}$$

形状 shapes 部分。首先将输入图片 x 送入美学视觉特征提取器中，然后对原始图像进行灰度处理。之后再对生成的灰度图使用预训练的 ViT 提取物品的外观形状信息。具体而言，设定输入物品图像为 x，针对图像 x 应用 Canny 边缘

检测器[19],获得置灰的形状图像x_{e_1}。然后对x使用如下的3×3卷积核f如式(19)所示,得到$x_{e_2}=x*f$。

$$f=\begin{bmatrix}-1 & -1 & -1\\ -1 & 8 & -1\\ -1 & -1 & -1\end{bmatrix} \tag{19}$$

Canny检测器拥有出色且全面的边缘特征提取能力及噪声抑制能力。而边缘检测卷积由于其特殊的卷积核则更注重判断图像中处于水平和倾斜的边缘区域。将两者处理后的数据综合起来,进行相加获得更为完整、更细粒度的图片形状细节表示,即式(20):

$$x_e=x_{e_1}+x_{e_2} \tag{20}$$

对x_e运用Black-White Inverted[20]方法,将图片中的黑色区域和白色区域相互转换,最终得到转换后的数据表示。

$$x_e end=Clicp(255-x_e,0,255) \tag{21}$$

将$x_e end$转换为tiff图像格式,使用小型的预训练ViT模型提取该图片得到初始形状向量。之后再添加一层Encoder编码层对初始形状向量进行进一步降维。这里的Encoder编码层采用类似ViT部分的$PCA(\cdot)$降维如式(22)所示,最后生成与l_i维度一致的形状部分的美学视觉特征向量$\boldsymbol{s}_i$。

$$\boldsymbol{s}_i=PCA(ViT(x_e end)) \tag{22}$$

3.4 ABVR特征融合与推荐模块

此前的工作中,特征提取模块分别使用GCN模块获取物品的图结构嵌入特征向量$\boldsymbol{q}_i$、ViT模块获取图像的高层视觉特征向量$\boldsymbol{l}_i$、Aesthetic模块获取图像的中层美学视觉特征向量(color向量$\boldsymbol{c}_i$、shapes向量$\boldsymbol{s}_i$)。其中提取的特征向量均统一设置为K维度,后续将进一步探索隐因子空间下K数值与推荐性能的内在联系。

特征融合模块。采用早期融合方式,使用$concat(\cdot)$作为拼接函数,将上述4种特征向量作为输入得到拼接融合后的物品部分的特征q_i^*:

$$\boldsymbol{q}_i^*=concat(\boldsymbol{q}_i,\boldsymbol{l}_i,\boldsymbol{c}_i,\boldsymbol{s}_i) \tag{23}$$

将拼接后的物品部分向量$\boldsymbol{q}_i^*$,使用多层感知机$MLP(\cdot)$进行降维操作,同时保持与GCN模块生成的用户的图结构嵌入向量p_u维度一致,得到最终的物品表示向量$\boldsymbol{q}_i'$:

$$\boldsymbol{q}_i'=MLP(\boldsymbol{q}_i^*) \tag{24}$$

推荐模块。结合式(7)和式(24),ABVR模型最终的偏好得分计算式如式(25)所示:

$$\hat{y}_{\mathrm{ABVR}}ui=p_u^{\mathrm{T}}\boldsymbol{q}_i' \tag{25}$$

目标函数。采用视觉感知推荐系统中广泛应用的BPR Loss。通过给定用户、积极物品、消极物品(u,i,j)三元组,最大化用户对积极物品和消极物品之间偏好得分之差,使得没有观察到的物品的偏好得分远低于观察过的物品的偏好得分。ABVR目标函数具体如下:

$$Loss=\sum_{(u,i,j)\in M}-\ln\sigma(\hat{y}_{ui}-\hat{y}_{uj})+\lambda\|\theta\|_2^2 \tag{26}$$

其中,三元组$M=\{(u,i,j)\mid(u,i)\in R^+,(u,j)\in R^-\}$表示成对排序的训练数据,$R^+$表示积极物品(观察过的物品),$R^-$表示消极物品(没有观察到的物品),$\sigma(\cdot)$表示sigmod激活函数,$\theta$表示可训练的模型参数,并且使用$\lambda$控制$L_2$正则化约束,防止过拟合。

4 实验

4.1 数据集

对本文所提出的方法在两个开源的数据集——亚马逊男孩和女孩数据集(Amazon_boys_girls)、亚马逊男人数据集(Amazon_men)上进行评估。这两个数据集是亚马逊产品类别Clothing,Shoes和Jewelry的子类别,并且是专门为视觉感知推荐系统构建的。在过滤阶段,只考虑2010年以后记录的交互,并去掉少于5次交互的物品和用户(对物品和用户应用5核技术)。最后,应用时间留一法将数据集拆分为训练集、验证集和测试集。对于每个数据集而言,每个物品都至少包含一张物品图片,具体图片样本如图5所示。数据集中用户和物品交互矩阵高度稀疏,如表1所列,主要记录了实验的统计数据。

(a) Amazon_boys_girls数据集　　(b) Amazon_men数据集

图5　不同数据集的示例

Fig. 5　Examples of different datasets

表1　实验数据集统计

Table 1　Statistics of datasets used in experiment

Dataset	Users	Items	Interactions	Sparsity/%
Amazon_boys_girls	1425	5019	9213	99.87
Amazon_men	6001	7370	16522	99.96

4.2 对比方法

为了评估本文提出的ABVR算法,使用以下Baseline方法进行对比,且均在Amazon_boys_girls,Amazon_men数据集上进行评估测试,输入图像尺寸统一设置为224×224,对比模型中的图像视觉特征提取器IFE均采用Elliot框架[1]下的resnet50,具体说明如下:

BPRMF[8]采用用户的隐式反馈(如点击、收藏等),通过对问题进行贝叶斯分析得到的最大后验概率来对物品进行推荐。

VBPR[6]是BPRMF算法的扩展,在预测偏好得分中添加视觉贡献部分的同时,考虑了视觉和非视觉因素对推荐的影响。

DeepStyle[16]在VBPR的基础上引入美学风格特征,纳入最终偏好得分计算,其中美学风格样式特征由提取到的图像视觉特征减去图像类别特征得到。

DVBPR[15]使用自定义的卷积神经网络CNN-F替换传统imagenet预训练的CNN,并且简化了VBPR的偏好得分预测部分,单纯考虑视觉因素对推荐影响。

ACF[7]通过对用户历史记录中的物品建模,即对物品图片不同区域以及图片本身赋予不同的注意力偏好权重,并且在提取特征时使用的是feature map。

AMR[11]是VBPR在图像攻防领域的一种扩展,运用经典的FGSM图像对抗算法,将提取到的高层视觉特征添加对抗噪声,进行对抗训练以增强系统鲁棒性。

VNPR[12]对 BPRMF 中用户偏好得分采用用户和物品潜在向量内积的做法进行了更改，借助广义矩阵分解的思想，使用非线性的方法去计算最终偏好得分。其次在提取图像的视觉特征后，施加 PCA 对其进行降维。

4.3 评估指标

为了验证本文模型的性能，进行了大量的实验。在 top@k 推荐列表中应用准确率相关评估指标，使用 HR@k 和 NDCG@k 来衡量整体推荐系统的性能优劣。在整个实验中，默认设置 k 为 20。所有模型均采用一个双 GPU 核心的 RTX3090 进行半精度浮点运算来训练。

命中率(Hit Ratio，HR)用于评估推荐列表 top@k 的物品是否命中，命中即为 1，否则为 0。

归一化折损累计增益(Normalized Discounted Cumulative Gain，NDCG)主要应用于推荐任务，衡量推荐列表 top@k 的优秀程度，即评估排序结果的好坏。NDCG 值越大，推荐的质量越高。

$$NDCG(k)=Z_n\sum_{i=1}^{k}(2^{r(j)}-1)/\log(1+j) \tag{27}$$

其中，Z_n表示正则化，j 表示物品在推荐列表 top@ k 中的位置索引，$r(j)$属于 Gain 增益，用于表示物品的相关性得分，$\sum_{i=1}^{k}(2^{r(j)}-1)$表示 CG 累积增益，$\sum_{i=1}^{k}(2^{r(j)}-1)/\log(1+j)$ 属于 DCG 折损累积增益，用于考虑具体排序的因素，其中使用 Discounted 目的是让排名靠前的物品增益更高，排名靠后增益进行折损。

4.4 参数设置

实验中，针对所提出的模型，对高层视觉特征和中层美学视觉特征以及图结构嵌入特征部分建模。由于每部分特征设置了降维操作，因此其对应降维部分的隐因子维度 K(嵌入维度)均在{32，64，128，256}中挑选。另外在实验中将采用 Adam 优化器对模型的可学习参数进行优化，学习率在{0.0001，0.001，0.01，0.1}范围内探索。

4.5 实验结果分析

4.5.1 Baseline 性能对比实验

对比 ABVR 和 baseline 模型之间的性能差异，表 2、表 3 分别列出了模型在 Amazon_boys_girls，Amazon_men 数据集上的 top@20 的性能指标。所有性能指标都进行了 5 次实验，取 5 次实验的平均结果，并用黑体和下划线分别标注最优结果和次优结果。

首先，从基线模型角度而言，无论是在表 2 还是表 3 中，BPRMF 在数据集中表现均最差，这表明了单纯的用户和物品内积计算偏好得分不足以捕获用户和物品之间的复杂关系。相较于传统的 BPRMF 算法，VBPR 在引入视觉特征后，HR 和 NDCG 均提升较为明显，这反映了视觉特征对推荐的重要性。Deepstyle 在使用美学风格样式特征替代直接使用高层者视觉特征的方法后，效果比 VBPR 有了更进一步的提升，也证明了美学风格特征的重要性。而 ACF 整体效果不如 VBPR，可能是因为 ACF 选择了 resnet50 作为 IFE，使用特征图来提取高层视觉特征而非全连接层最后一层，因此对于 ACF 而言，选择简单的网络结构，如 alexnet，可能效果更佳。AMR 针对 VBPR 在视觉上的不健壮性，加入对抗噪声进行对抗训练，提升系统鲁棒性，性能相比 VBPR 也有些许提升。VNPR 由于引入非线性方式计算偏好得分，略优于 VBPR。

其次，与近年来最先进的基线模型相比，本文提出的 ABVR 模型在表 2 的 Amazon_boys_girls 上 HR 指标表现最佳，其次是 Deepstyle。在 NDCG 指标上排名次优，仅次于 Deepstyle。在表 3 的 Amazon_men 上 HR 和 NDCG 指标则均达到了 SOTA。这些都证明了，在推荐系统中引入高层视觉特征和中层美学视觉特征以及图结构嵌入特征后拥有极强的竞争力，整体效果优于大部分基线模型。单独从视觉角度而言，在两个数据集上对比 VBPR，VNPR，ACF 模型，HR 和 NDCG 均提升明显，说明模型能够很好地胜任视觉推荐任务，效果显著。而从美学任务的角度而言，对比 Deepstyle，DVBPR，HR 指标均优于基线模型，但在 Amazon_boys_girls 上，NDCG 略低于 Deepstyle，但优于 DVBPR，主要是因为该数据集中同类别商品风格样式较多，特别是针对青少年用户(男孩/女孩)更多考虑的是商品外观风格样式上的差异性，而不是类别的差异性。这也从侧面反映了引入美学特征后，对推荐起积极影响。因此本文提出的 ABVR 在模型设计中除引入中层美学视觉特征外，还引入了高层视觉特征来共同增强物品的嵌入部分表示，实验结果从视觉及美学角度证明了本文方案的有效性和优异性。

表 2 Amazon_boys_girls 上模型 top@20 性能指标对比

Table 2 Top@20 performance metrics comparison on Amazon_boys_girls

Dataset	Model	*HR*	*NDCG*
Amazon_boys_girls	BPRMF	0.01474	0.00508
	VBPR	0.03018	0.01287
	DeepStyle	<u>0.03719</u>	**0.01543**
	DVBPR	0.03491	0.01211
	ACF	0.01544	0.00482
	AMR	0.03213	0.01321
	VNPR	0.03053	0.01229
	Our Method(ABVR)	**0.04140**	<u>0.01480</u>

表 3 Amazon_men 上模型 top@20 性能指标对比

Table 3 Top@20 performance metrics Comparison on Amazon_men

Dataset	Model	*HR*	*NDCG*
Amazon_men	BPRMF	0.01347	0.00473
	VBPR	0.02952	0.01225
	DeepStyle	<u>0.03434</u>	<u>0.01340</u>
	DVBPR	0.03023	0.01136
	ACF	0.02548	0.01029
	AMR	0.03361	0.01336
	VNPR	0.02958	0.01303
	Our Method(ABVR)	**0.03533**	**0.01347**

4.5.2 消融实验

为了更好地了解提出的 ABVR 模型及其各个组成部分的效果，分别在两个数据集上针对模型的关键部分进行消融分析——图结构嵌入特征提取 GCN 模块、高层视觉特征 ViT 提取模块、中层美学视觉特征 color 提取模块、中层美学视觉特征 shapes 提取模块。表 4 列出了对应模块的评估结果，其中 GCN 和 GCN+ViT 分别指仅使用图结构嵌入特征模块和同时引入图结构嵌入特征和高层视觉特征两者的模块。GCN+ViT+Color 和 GCN+ViT+Shapes 则分别指同时引入图结构嵌入特征、高层视觉特征、中层美学视觉特征 color 三者的模块以及同时引入图结构嵌入特征、高层视觉特征和中层美学视觉特征 shapes 三者的模块。

表 4　不同特征配置下的消融实验

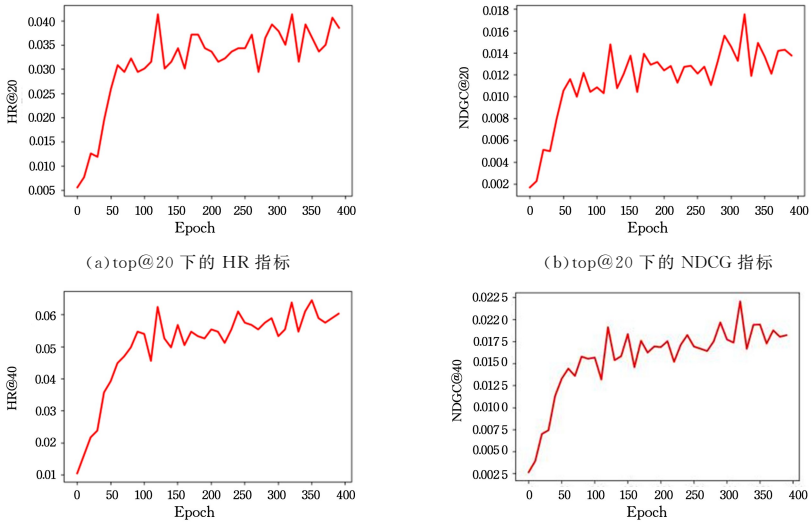
Table 4　Ablation experiments with different features

Model	Amazon_boys_girls		Amazon_men		Params/M
	HR@20	NDCG@20	HR@20	NDCG@20	
GCN	0.00912	0.00238	0.00843	0.00211	11
GCN＋ViT	0.03859	0.01376	0.03166	0.01259	97
GCN＋ViT＋Color	0.03866	0.01381	0.03333	0.01318	109
GCN＋ViT＋shapes	0.03989	0.01396	0.03449	0.01343	107
ABVR	**0.04140**	**0.01480**	**0.03533**	**0.01347**	**119**

从表 4 中可以看出，仅使用 GCN 在 HR 和 NDCG 上效果不佳，而在引入了 ViT 后，参数量增长的同时，效果也提升显著，这证实了对物品视觉信息进行建模的重要性，增强了图神经网络的性能。其次在性能上 GCN＋ViT 组合也优于表 2、表 3 中的大部分 baseline 模型，同时也反映了图结构嵌入特征对视觉推荐上的作用是相辅相成的。另外在引入了中层美学视觉特征 color 后，相较于 GCN＋ViT 性能有小幅度提升，表明颜色特征对视觉推荐具备一定增强效果。在引入中层美学视觉特征 shapes 后，也展现了优于 GCN＋ViT 的性能。其次在引入高层视觉特征和中层美学视觉特征后，Amazon_men 上的整体性能表现略低于 Amazon_boys_girls，推测和数据集自身的图像色彩和形状有关，相比之下 Amazon_boys_girls 数据集中的图像针对青少年，所以更生动、更具美感、也更有视觉冲击力，因此表现更佳。最后，从表中可以看出，以参数量增长为真实代价带来的推荐效益是显著的，即同时使用 4 种特征的 ABVR 模型优于所有消融部分，也证实了中层美学视觉特征和高层视觉特征以及图结构嵌入特征融合的有效性。

4.5.3　性能评估实验

在性能评估实验中，选取图像更为鲜艳丰富的 Amazon_boys_girls 数据集，并设定 400 个 epoch 进行训练，指标分别选择 top@20 和 top@40 的 HR 和 NDCG。从图 6(a)、图 6(b)可以观察到 top@20 的 HR 最大值在 0.041 左右，NDCG 最大值在 0.017 左右。作为对比，图 6(c)、图 6(d)在 top@40 的 HR 最大值达到 0.062 左右，NDCG 最大值在0.021 2左右。因此随着迭代次数的增加，HR 和 NDCG 指标数值整体保持上升的趋势。在训练过程中前 50～100 个 epoch 上，性能评估指标(HR 和 NDCG)增长迅速，而后 100～400 个 epoch 上整体增长速度放缓，局部小幅波动，与学习率设定有关，在不断震荡的迭代区间中寻找最佳点。总体而言，top@40 下整体性能要优于 top@20，这意味着更大的推荐序列 k，模型拥有更佳的推荐性能。



(a)top@20 下的 HR 指标　(b)top@20 下的 NDCG 指标

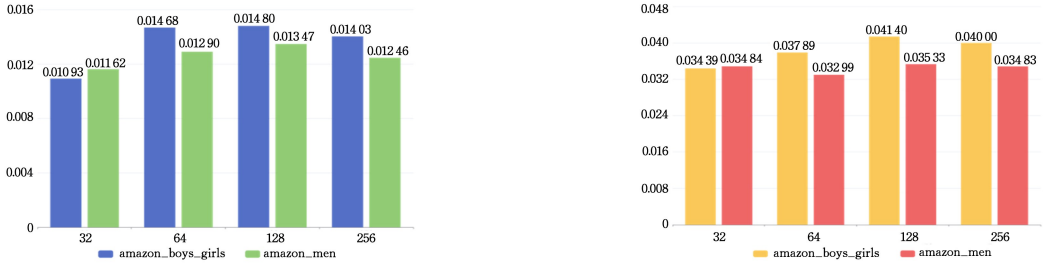
(c)top@40 下的 HR 指标　(d)top@40 下的 NDCG 指标

图 6　Amazon_boys_girls 上不同 top@k 下的性能趋势

Fig. 6　ABVR performance trends on Amazon_boys_girls with different top@k

4.5.4　参数探索实验(K 隐因子探索)

不同类型特征在经过降维操作后，统一映射在隐空间下进行信息融合后，为了探索隐空间下 K 隐因子(嵌入维度)对模型性能的影响，本文进行了 K 隐因子探索实验。图 7 给出了数据特征降维至 K 隐因子的维度大小时，不同 K 值对模型的影响。隐因子维数在{32，64，128，256}内选择。



(a)K 隐因子在不同数据集上的 NDCG@20 指标表现　(b)K 隐因子在不同数据集上 HR@20 指标表现

图 7　K 隐因子在不同数据集上的影响(电子版为彩图)

Fig. 7　Effect of K implicit factor on dfferent datasets

从图 7(a)、图 7(b)可以看出，不论是在 Amazon_boys_girls 数据集上(分别用蓝色和黄色表示)，还是在 Amazon_men 数据集上(分别用绿色和红色表示)，整体均呈上升趋势，当 K 隐因子维度为 128 维时，本文提出的模型性能指标表现最佳，这表明了随着隐因子维度增加，空间上映射到的点会更多，因此不同特征用于表示的信息也就越多，承载的内容也就越丰富，性能也随之提升。但当 K 隐因子维度为 256 维时，性能出现下降，主要是因为维度太大时容易出现过拟合。因此，本文实验中 K 隐因子维度统一设置为 128。

结束语 本文对现阶段国内外视觉感知推荐、图神经网络推荐、美学推荐的研究现状进行了梳理，同时总结了当前推荐中存在的一些问题。针对这些问题，此项研究工作表明将图像划分为高层视觉特征和中层美学视觉特征，并从美学的角度来表征物品可以取得良好的效果，且更符合实际生活中的各种时尚领域推荐应用场景。其次在高层视觉特征提取上创新性地采用 ViT 替代传统的 CNN 来处理图像数据，在消融实验中验证了 ViT 的有效性。最后，将图卷积模块提取的图结构嵌入特征与高层视觉特征以及中层美学视觉特征共同映射到潜在空间中进行融合，并整体应用到 BPR 贝叶斯个性化框架下进行推荐。实验结果表明，所提方法优于其他仅利用视觉感知推荐方法、图神经网络推荐方法和美学推荐方法。

在未来的研究工作中，针对物品部分将引入更多的辅助信息来丰富目前的视觉感知推荐系统，如文字描述信息、地理位置信息等，可以将图片视觉特征与文本语义特征的结合共同纳入后续研究中。如针对用户部分将考虑用户画像信息、用户的序列信息等。此外，尝试捕捉视觉推荐中动态美学(用户随时间变化审美会逐渐变化)对用户的影响。

参 考 文 献

[1] ANELLI V W, BELLOGÍN A, FERRARA A, et al. Velliot: Design, evaluate and tune visual recommender systems[C]// Fifteenth ACM Conference on Recommender Systems. 2021: 768-771.

[2] ANELLI V W, DELDJOO Y, DI NOIA T, et al. A study of defensive methods to protect visual recommendation against adversarial manipulation of images[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1094-1103.

[3] WANG H, ZHAO M, XIE X, et al. Knowledge graph convolutional networks for recommender systems[C]// The World Wide Web Conference. 2019: 3307-3313.

[4] 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.

[5] WANG H, ZHANG F, ZHAO M, et al. Multi-task f-eature learning for knowledge graph enhanced recommendation[C]// The World Wide Web Conference. 2019: 2000-2010.

[6] HE R, MCAULEY J. VBPR: visual bayesian personalized ranking from implicit feedback[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2016: 144-150.

[7] 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 Developmentin Information Retrieval. 2017: 335-344.

[8] RENDLE S, FREUDENTHALER C, GANTNER Z, et al. Bayesian personalized ranking from implicit feedback[C]// Proceedings of Uncertainty in Artificial Intelligence. 2014: 452-461.

[9] 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.

[10] HE X, DENG K, WANG X, et al. Lightgcn: Simp-lifying and powering graph convolution netw-ork for recommendation[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 639-648.

[11] TANG J, DU X, He X, et al. Adversarial training towards robust multimedia recommender syst-em[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(5): 855-867.

[12] HE X, LIAO L, ZHANG H, et al. Neural collabor-ative filtering[C]// Proceedings of the 26th International Conference on World Wide Web. 2017: 173-182.

[13] NIU W, CAVERLEE J, LU H. Neural personalized ranking for image recommendation[C]// Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. 2018: 423-431.

[14] GONG W, KHALID L. Aesthetics, Personalization and Recommendation: A survey on Deep Learning in Fashion[J]. arXiv: 2101.08301, 2021.

[15] KANG W C, FANG C, WANG Z, et al. Visually-aware fashion recommendation and design withgenerative image models[C]// 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017: 207-216.

[16] LIU Q, WU S, WANG L. Deepstyle: Learning user preferences for visual recommendation[C]// Proceedings of the 40th international ACM Sigir Conference on Research and Development in Information Retrieval. 2017: 841-844.

[17] WANG Z, CHANG S, DOLCOS F, et al. Brain-inspired deep networks for image aesthetics assessment[J]. Michigan Law Review, 2016, 52(1): 123-128.

[18] PAUL A, WU Z, LIU K, et al. Robust multi-objective visual bayesian personalized ranking for multimedia recommendation [J]. Applied Intelligence, 2022, 52(4): 3499-3510.

[19] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1986: 679-698.

[20] TANGSENG P, OKATANI T. Toward explainable fashion recommendation[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2153-2162.

ZHANG Kaixuan, born in 1996, postgraduate. His main research interests include recommendation system and graph deep learning.

CAI Guoyong, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include social media and data mining.