计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(3):62−71 [doi: 10.15888/j.cnki.csa.009810] [CSTR: 32024.14.csa.009810]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合多模态特征的可解释推荐算法①

王紫萱, 张凯涵, 蔡江辉, 郭青松, 徐鑫芳

(中北大学 计算机科学与技术学院, 太原 030051)
通信作者: 张凯涵, E-mail: zhangkh@nuc.edu.cn

摘　要: 可解释推荐算法利用行为信息和其他相关信息不仅生成推荐结果而且提供推荐理由, 从而增加推荐的透明度和可信度. 传统的可解释推荐算法往往局限于分析评分数据和文本数据, 对图像这类数据利用并不充分, 且并没有很好地考虑模态间的有效融合方式, 难以充分挖掘不同模态之间的内在关联. 针对上述问题, 提出一种融合多模态特征的可解释推荐模型, 该模型采用特征融合技术, 从多模态角度提高推荐解释的质量与个性化. 首先, 设计多模态特征提取方法, 基于 CLIP 图像编码器和文本编码器分别提取用户和物品的文本特征和图像特征. 其次, 采用交叉注意力技术实现文本和图像的跨模态融合, 增强模态间的语义相关性. 最后, 将多模态信息与交互信息结合, 联合优化模态对齐、评分预测与解释生成任务. 实验结果表明, 所提出的方法在 3 个多模态推荐数据集上都表现出了明显优势, 尤其在提升解释质量方面.

关键词: 可解释推荐; 多模态; 特征融合; 交叉注意力; 模态对齐

引用格式: 王紫萱,张凯涵,蔡江辉,郭青松,徐鑫芳.融合多模态特征的可解释推荐算法.计算机系统应用,2025,34(3):62–71. http://www.c-s-a.org.cn/1003-3254/9810.html

Explainable Recommendation Algorithm Fusing Multimodal Features

WANG Zi-Xuan, ZHANG Kai-Han, CAI Jiang-Hui, GUO Qing-Song, XU Xin-Fang

(School of Computer Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: Explainable recommendation algorithms utilize behavioral and other relevant information to not only generate recommendation results but also provide recommendation explanations, thereby increasing the transparency and credibility of recommendations. Traditional explainable recommendation algorithms are often limited to analyzing rating data and text data and fail to fully utilize data such as images. They also do not consider effective fusion methods between modalities, making it difficult to fully unearth the intrinsic relationships between different modalities. An explainable recommendation model that fuses multimodal features is proposed to address the above-mentioned issues. This model improves the quality and personalization of recommendation explanations from a multimodal perspective through feature fusion technology. Firstly, a multimodal feature extraction method is designed based on CLIP image encoder and text encoder to extract text and image features of users and items, respectively. Secondly, cross attention technology is used to achieve cross modal fusion of text and images, enhancing semantic correlation between modalities. Finally, multimodal information is combined with interactive information to jointly optimize modal alignment, rating prediction, and explanation generation. Experimental results show that the proposed method exhibits significant advantages in the three multimodal recommendation datasets, especially in improving explanation quality.

Key words: explainable recommendation; multimodal; feature fusion; cross attention; modal alignment

① 基金项目: 国家自然科学基金 (72171137, 62401525); 山西省基础研究计划 (202203021222075, 202203021211331)
收稿时间: 2024-09-06; 修改时间: 2024-10-21, 2024-11-01; 采用时间: 2024-11-07; csa 在线出版时间: 2025-01-21
CNKI 网络首发时间: 2025-01-22

推荐系统的核心在于深入分析用户行为偏好, 以提供个性化和高效的信息服务[1]. 在传统的推荐系统[2–4]中, 推荐结果通常是基于复杂的模型和数据处理过程得出的, 对于用户而言, 推荐结果的背后机制通常是“黑箱”, 难以理解. 这种不透明性降低了用户对推荐结果的信任度, 缺乏对推荐结果的解释, 也使得开发者在优化系统时缺乏有效的指导. 在电子商务、社交媒体、音视频推荐等多个领域, 用户对推荐服务的需求不断增长[5], 不仅希望获得准确的推荐结果, 还期望了解推荐背后的逻辑和原因, 而可解释性推荐系统可以向用户提供解释说明. 例如, 当系统推荐某个好友时, 可以明确指出推荐的依据, 比如你们之间有共同的好友或共享的兴趣话题, 这种清晰的关联性让用户更容易接受推荐, 并愿意与平台互动; 当推荐一部电影时, 可以说明: “你可能会喜欢这部电影, 因为它与您之前喜欢的 X 类似, 且主演是你喜欢的演员. ”这样的解释不仅帮助用户理解推荐的逻辑, 还增加了推荐的可信度. 因此, 提高推荐算法的可解释性[6]成为当下的重要需求.

可解释推荐算法生成的解释形式有多种, 包括统计图表[7]、知识图谱[8,9]以及自然语言等. 其中, 自然语言类型的解释更加直观且易于用户理解, 已成为目前主流的解释形式. 早期自然语言类的解释依赖于预先定义好的通用文本模版[10], 将预测的关键特征词拼接进模板中形成完整的解释文本, 此类方法生成的解释较为同质化, 所有用户共享相同的解释模版, 解释文本缺乏多样性.

为了生成个性化的文本解释, 研究人员利用先进的自然语言生成技术, 结合用户的评论信息、物品的属性信息等文本数据开展了广泛研究. 早期工作主要采用循环神经网络及其变体模型[11,12], 但它们难以建模文本中的长期依赖关系且计算效率较低. 近年来, Transformer 模型[13]及其变体在特征表示及建模多源信息方面均展现出优异性能, 例如, Liu 等人[14]设计了一个基于 Transformer 的模型结构应用在多模态推荐场景中, 自适应地为每个用户整合这些多模态信息, 以学习高质量的特征表示. 因此, 在可解释推荐任务中其受到大量关注.

虽然现有方法在生成个性化的文本解释方面做了诸多探索[15], 但是绝大多数工作都集中于对文本数据的建模, 多模态信息的潜力常常被忽视. 在推荐系统领域, 多模态信息主要包括物品的图片信息、用户评分数据与评论文本等, 这些信息能够从多角度反映物品特征与用户偏好. 例如, 物品图片能够直观展示商品的外观和使用场景, 用户的评分数据则提供了对物品性能的量化反馈, 而评论文本可以揭示用户的具体使用体验和情感倾向. 对此, 本文提出一种融合多模态特征的可解释推荐模型 (explainable recommendation model fusing multimodal features, ERFMF). 该模型从用户和物品各自的多模态信息入手, 设计文本编码器和图像编码器分别提取用户和物品的文本特征和图像特征, 并采用交叉注意力技术[16]实现文本特征和图像特征的跨模态融合, 进一步将协同信号 (即用户和物品 ID) 与跨模态特征相结合得到用户和物品的多模态特征; 此外, 为了缩小多种模态信息的异质性差距, 利用跨模态对齐机制引入对比损失, 促使不同模态特征在统一的语义空间中进行相互作用; 最后, 采用多任务学习框架, 联合建模模态对齐、评分预测和解释生成任务.

本文的主要贡献如下.

1) 本文提出融合多模态特征的可解释推荐模型 ERFMF, 充分建模用户和物品的评分信息、评论文本及视觉特征等多模态信息, 多模态融合主要解决信息互补和数据稀疏性的问题, 通过图像与文本信息的相互补充, 可以有效弥补数据缺失.

2) 本文提出基于 Transformer 的跨模态特征融合策略, 采用自注意力与交叉注意力建模用户与物品特征, 其主要解决信息的关联性, 物品图片、用户评分以及评论文本这 3 种信息之间存在着潜在的关联, 挖掘它们之间的联系和依赖关系, 可以有效增强模型的理解能力与适应性.

3) 在 3 个多模态推荐数据集上进行对比实验, 结果表明 ERFMF 不仅能利用多模态信息缓解数据稀疏性的影响, 而且生成的文本解释相比于其他几个对比基线质量更高.

1 相关工作

1.1 多模态推荐算法

多模态推荐方法是指利用多种类型的数据来进行推荐任务. 初期多模态推荐研究主要利用用户与物品的交互信息 (如评分) 以及文本信息 (如评论), Dong 等人[17]结合马尔可夫链和协同过滤的思想, 提出了一种基于社交标签的个性化推荐算法, 通过马尔可夫链模型计算用户对标签的兴趣程度, 然后通过推荐标签集匹配与之对应的物品. 随着多模态信息的不断丰富, 研

究人员发现图像信息反映了人们的视觉感知, 与其他模态信息可以互为补充. Truong 等人[18]采用长短期记忆 (long short-term memory, LSTM) 网络同时建模评分信息和评论信息, 当评论图片可用时, 引入视觉特征作为辅助信息. Wu 等人[19]根据新闻图片对于用户点击新闻的吸引力提出一种多模态新闻推荐方法, 该方法利用预训练的视觉语言模型 ViLBERT 对新闻文本和图像信息的内在相关性进行建模. Geng 等人[20]提出开发一个结合图像、文本和交互信息的多模态基础模型, 在一个共享的架构中处理多个模态以改进推荐. Wei 等人[21]整合多种模态信息, 开发了一个特定模态嵌入层用于研究不同模态特征提取对用户偏好建模和预测用户-物品交互的影响.

推荐领域多模态信息的研究是非常重要并且具有实际应用价值的, 而对于多模态信息的融合通常采用简单拼接或注意力机制融合不同类型数据, 这未能充分挖掘不同模态间的关联性, 容易丢失重要信息.

1.2 可解释推荐算法

自然语言生成方法目前是可解释推荐的研究重点, Li 等人[10]设计了一种改进门控循环单元 (gated recurrent unit, GRU) 模型从数据中学习相关信息, 不仅可以生成模板式的解释句, 同时也可以进行评分预测. 句子模板会限制推荐解释的表达能力, Li 等人[15]发现协同信号对于个性化的解释生成十分重要, 通过对用户和物品的协同信号、评分和评论联合建模来预测目标解释句中的单词, 生成个性化的文本解释. Jin 等人[12]利用 Transformer 和可变门控循环单元对物品信息和用户评论进行编码, 构建了一个编码器-解码器模型, 以产生真实和多样化的解释句. Chen 等人[22]引入隐式因素将用户和物品的协同信号转换为个性化表示, 并与模型识别出的重要评论和概念聚合辅助生成解释句. Xie 等人[23]通过分析评论中包含的情感极性, 将用户和物品的协同信号、评论特征和情感信息融合到多任务学习框架中同时生成评分和一个文本解释.

上述方法主要利用评分信息和文本信息, 将多模态信息融入可解释推荐的探索相对较少, 而多模态信息在提升推荐系统的准确性和个性化程度方面具有重要研究意义. 通过综合这些多样化的信息源, 推荐系统能够更全面、精准地理解用户的需求和偏好. 本文结合了多种类型数据, 包括协同信号、评分数据、文本评论以及视觉图像多种信息, 深入挖掘用户偏好与物品特征.

2 模型介绍

ERFMF 模型结构如图 1 所示, 该模型主要包括 3 部分: 1) 特征提取, 其中包括文本编码器和图像编码器, 分别从历史评论与视觉图像中捕获用户和物品的文本特征和图像特征. 2) 多模态融合, 该模块用于融合协同信号、文本和图像这 3 种模态特征, 分别获得用户多模态特征和物品多模态特征. 3) 多任务联合优化, 将图像特征和用户-物品对的真实评论文本特征进行模态对齐, 并且实现高效的评分预测和解释生成.

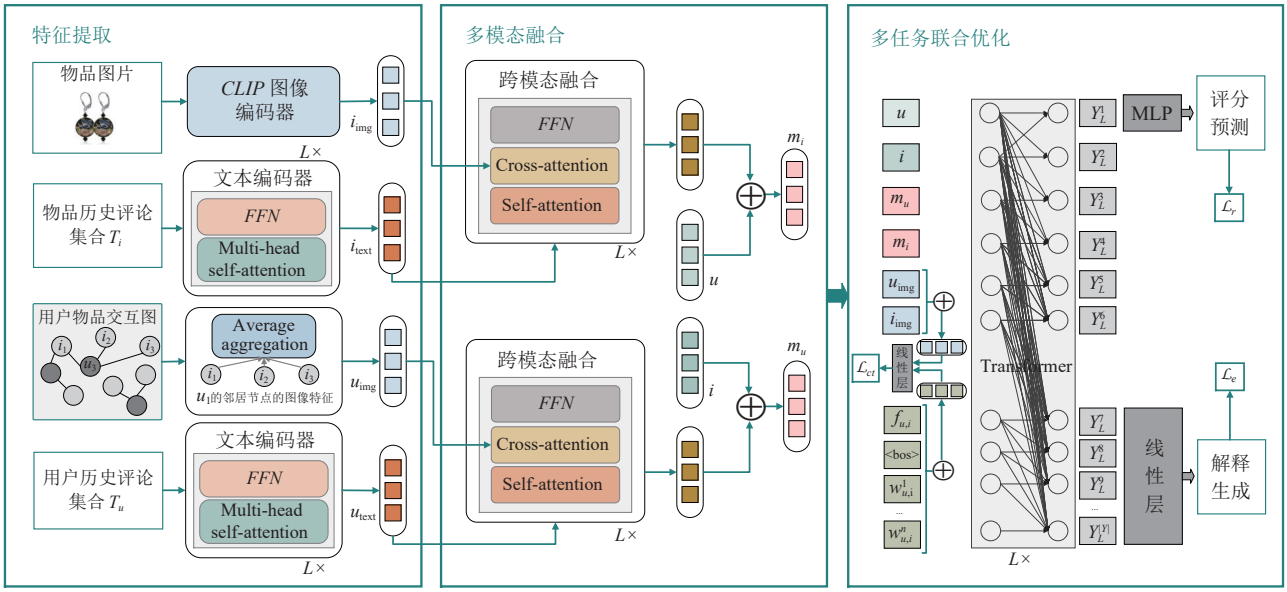


图 1 ERFMF 模型框架

与传统模型相比, ERFMF 模型在处理复杂数据时能够更深入捕捉用户偏好, 综合运用先进技术, 在可解释推荐系统中表现出色, 推动推荐精度与解释质量的双重提升. 其平行网络结构能够充分挖掘多模态信息, 将文本和视觉特征有效融入可解释推荐算法中, 最大程度发挥不同数据类型的优势. 技术上, 该模型基于 Transformer 结构, 联合优化评分预测和解释生成任务, 确保推荐与解释之间的协同作用.

2.1 问题描述与符号定义

给定数据集D, 数据集中的样本格式为$(u,i,r_{u,i},p_i,T_u,T_i,F_{u,i},W_{u,i})$, 其中$r_{u,i}$表示用户$u$对物品$i$在 1–5 整数范围内的评分, p_i表示物品i的一幅图像. T_u表示用户u对物品的历史评论词集合, T_i表示物品i收到的历史评论词集合. $F_{u,i}$表示物品i的特征词集合, $W_{u,i}$表示用户u对物品i的真实评论. 特别地, 定义了一个词汇表$V=\{v_1,\cdots,v_{|V|}\}$, 它包括数据集中所有评论词. 本文任务是根据给定的数据集D, 预测出评分$\hat{r}_{u,i}$, 同时生成一个自然语言文本$\hat{W}_{u,i}$, 进一步阐明为什么模型将物品i推荐给用户u. 本文所用数学符号及定义见表 1.

表 1 符号说明

符号	定义
D	数据集
$W_{u,i}$	用户u对物品i的真实评论
$F_{u,i}$	物品i的特征词集合
$r_{u,i}$	用户u对物品i的真实评分
T_u	用户u对物品的历史评论词集合
T_i	物品i收到的历史评论词集合
m_u	用户u的多模态特征
m_i	物品i的多模态特征
u_{img}	用户u的图像特征
i_{img}	物品i的图像特征
u_{text}	用户u的文本特征
i_{text}	物品i的文本特征

2.2 特征提取

为充分挖掘用户与物品的多模态特征, 本文分别设计文本编码器与图像编码器, 通过注意力机制从历史评论中提取用户和物品的文本特征, 并采用预训练模型 CLIP 从图像信息中提取物品的视觉特征, 进一步地, 利用消息传播机制获取用户的视觉特征.

2.2.1 文本编码器

用户对物品的历史评论能够反映用户的兴趣偏好与物品的细粒度特征等, 因此, 本文设计面向用户和物品具有相同网络结构的文本编码器, 从评论词集合T_u和T_i中提取文本特征. 接下来, 以用户为例介绍文本编码器结构.

首先, 通过嵌入层获取初始词表示$T_u^0=[t_1,\cdots,t_q]$, 其中$t_k\in\mathbb{R}^{1\times d}$表示评论中第$k$个词的嵌入表征, q表示历史评论集中单词个数. 然后, 将其输入文本编码器中, 它采用L层 Transformer 编码器组成, 每个层包括两个子层: 多头自注意力层 (multi-head self-attention) 和前馈神经网络层 (feed forward network, *FFN*). 在用户偏好建模过程中, 考虑到评论句中不同词汇的贡献差异, 采用多头自注意力计算每个单词的重要性, 计算公式如下:

$$\begin{cases} M_u^l = Concat(A_{l,1},\cdots,A_{l,H})W_l^O \\ A_{l,h} = Softmax\left(\dfrac{Q_{l,h}K_{l,h}^{\mathrm{T}}}{\sqrt{d}}\right)V_{l,h} \\ Q_{l,h}=T_u^{l-1}W_{l,h}^Q,\ K_{l,h}=T_u^{l-1}W_{l,h}^K,\ V_{l,h}=T_u^{l-1}W_{l,h}^V \\ T_u^{l-1}=FFN_{l-1}(M_u^{l-1}) \end{cases} \tag{1}$$

其中, $l\in[1,L]$, $h\in[1,H]$表示对应层的第h个注意力头. M_u^l表示多头自注意力第l层的输出, T_u^{l-1}是第$l-1$层编码器的输出, $W_{l,h}^Q$、$W_{l,h}^K$、$W_{l,h}^V\in\mathbb{R}^{d\times d_z}$, $W_l^O\in\mathbb{R}^{d_z|H|\times d_z}$, $d_z=d/H$表示每个注意力头的维数.

最后, 将第L层多头自注意力的输出M_u^L输入到前馈神经网络中计算最终用户u的文本特征$u_{\text{text}}\in\mathbb{R}^{1\times d}$, 计算公式如下:

$$u_{\text{text}}=FFN_L(M_u^L) \tag{2}$$

同理, 可以得到最终物品i的文本特征$i_{\text{text}}\in\mathbb{R}^{1\times d}$.

2.2.2 图像编码器

在图像上采用预训练模型 CLIP[24]的图像分支作为图像编码器. 具体来说, 给定一张物品图像p_i, 将其输入到 CLIP 图像编码器中,来提取物品i的图像特征$i_{\text{img}}\in\mathbb{R}^{1\times d}$, 计算公式如下:

$$i_{\text{img}}=CLIP(p_i) \tag{3}$$

本文构建一张用户物品交互图, 用以获取每个用户所交互的物品集合, 将该集合中物品的图像特征聚合计算用户u的图像特征$u_{\text{img}}\in\mathbb{R}^{1\times d}$. 计算公式如下:

$$u_{\text{img}}=\frac{1}{|N_u|}\sum_{i\in N_u}i_{\text{img}} \tag{4}$$

其中, N_u表示用户u的交互物品集合.

2.3 多模态特征融合

多模态信息可以反映物品不同角度的特征以及用

户对不同方面特征的偏好程度, 为了更加深入地挖掘不同模态信息之间的关联性, 本文提出一个跨模态融合模块, 充分融合两种模态的特征信息. 由于用户和物品的跨模态特征融合是结构相同的两条平行网络, 因此本节以用户部分为例进行说明.

具体来说, 该模块由L层带有交叉注意力的 Transformer 块组成, 每一层由自注意力层 (SA)、交叉注意力层 (CA) 和前馈神经网络 (FFN) 组成. 首先将文本特征输入到自注意力层, 然后通过交叉注意力层计算文本和图像特征之间的跨模态交互, 之后通过前馈神经网络层得到跨模态特征. 计算公式如下:

$$u^l_{\text{text},SA} = LN(SA(u^{l-1}_{\text{text}}) + u^{l-1}_{\text{text}}) \tag{5}$$

$$u^l_{\text{text},CA} = LN(CA(u^l_{\text{text},SA}, u^{l-1}_{\text{img}}) + u^l_{\text{text},SA}) \tag{6}$$

$$u^l_{\text{text},FFN} = LN(FFN(u^l_{\text{text},CA}) + u^l_{\text{text},CA}) \tag{7}$$

其中, u^{l-1}_{text}和u^{l-1}_{img}表示$l-1$层的输出, LN 表示层归一化 (layer normalization, LN). 最后将第L层跨模态特征与协同信号拼接融合计算最终用户u的多模态特征$m_u \in \mathbb{R}^{1\times d}$.

$$m_u = [u^L_{\text{text},FFN}; u_{id}] \tag{8}$$

其中, ;表示拼接符号, $u_{id} \in \mathbb{R}^{1\times d}$表示用户$u$的 ID 表示. 同理, 可以得到物品$i$的多模态特征$m_i \in \mathbb{R}^{1\times d}$.

2.4 多任务联合优化

本节包括 3 个任务, 分别是模态对齐、评分预测与解释生成. 首先, 将用户u与物品i的 ID 特征、多模态特征、u对i的评论词特征拼接为输入序列$Y = [u, i, m_u, m_i, u_{\text{img}}, i_{\text{img}}, f_{u,i}, bos, w^1_{u,i}, \cdots, w^n_{u,i}]$, 其中$m_u$和$m_i$虽然已经融入视觉特征, 但是其更加偏向语言特征, 为了增强模型学习时的视觉特征, 防止模型过于依赖语言信号而疏忽视觉信号, 本文进一步将图像特征u_{img}和i_{img}引入输入序列Y中. 其中, $f_{u,i} \in \mathbb{R}^{1\times d}$为$F_{u,i}$中特征词经过嵌入操作后的词表示, $w^k_{u,i} \in \mathbb{R}^{1\times d}$为$W_{u,i}$中真实评论词经过嵌入操作后得到的第$k$个词表示, $k \in [1, n]$, n表示评论长度, bos为设定的开始标记.

2.4.1 模态对齐

为了将图像特征和用户-物品对的文本特征映射到公共语义空间, 使其能够更好地进行模态间的学习, 本文采用一个线性层作为映射网络, 运用跨模态对比思想采用均方误差 (mean square error, MSE) 作为对齐图像特征和文本特征的损失函数. 计算公式如下:

$$\mathcal{L}_{ct} = \frac{1}{|\mathcal{S}|}\sum_{(u,i)\in\mathcal{S}} (TF_{\text{mean}} - IF_{\text{mean}})^2 \tag{9}$$

其中, $TF_{\text{mean}} = f_{u,i} \oplus bos \oplus w^1_{u,i} \oplus \cdots \oplus w^n_{u,i}$表示平均融合后文本信息, $IF_{\text{mean}} = u_{\text{img}} \oplus i_{\text{img}}$表示平均融合后的图像信息, $\mathcal{S}$表示样本集合.

2.4.2 评分预测

推荐过程的目标是预测用户u对物品i的评分$\hat{r}_{u,i}$. 将输入序列Y经过L层 Transformer 得到输出序列表示$Y_L = [Y^1_L, Y^2_L, \cdots, Y^{|Y|}_L]$, 由于模型中的$u$和$i$是相互关注的, 通过这种方式捕获它们之间的交互信息, 所以模型可以采用Y^1_L或者Y^2_L来预测评分. 具体而言, 本文采用一个具有隐藏层的多层感知器 (multi-layer perceptron, MLP) 将Y^1_L映射为一个标量. 计算公式如下:

$$\hat{r}_{u,i} = w^r \sigma(W^r Y^1_L + b^r_1) + b^r_2 \tag{10}$$

其中, $w^r \in \mathbb{R}^{1\times d}$, $W^r \in \mathbb{R}^{d\times d}$, $b^r_1 \in \mathbb{R}^d$, $b^r_2 \in \mathbb{R}$是权重参数, $\sigma(\cdot)$是 Sigmoid 函数. 对于该任务我们使用均方误差作为损失函数, 计算公式如下:

$$\mathcal{L}_r = \frac{1}{|\mathcal{S}|}\sum_{(u,i)\in\mathcal{S}} (\hat{r}_{u,i} - r_{u,i})^2 \tag{11}$$

其中, $r_{u,i}$为用户u对物品i真实评分.

2.4.3 解释生成

在可解释推荐算法研究中, 不仅需要预测推荐列表, 还需要提供推荐理由, 该模块以用户和物品的多模态表征为输入, 采用 Transformer 解码器生成一句文本样式的推荐理由, 以向用户解释为什么推荐某些物品. 采用负对数似然 (negative log-likelihood, NLL) 函数作为解释任务的损失函数, 并计算训练集中用户-物品对的均值, 其公式如下:

$$\mathcal{L}_e = \frac{1}{|\mathcal{S}|}\sum_{(u,i)\in\mathcal{S}} \frac{1}{n+1}\sum_{t=1}^{n+1} -\log c^{w_t}_{8+t} \tag{12}$$

其中, $c^{w_t}_{8+t}$表示输入序列第$8+t$位置上词$w_t \in W_{u,i}$的概率分布, 因为从序列第 8 位开始为模型生成的解释文本.

在测试阶段, 模型使用$u, i, m_u, m_i, u_{\text{img}}, i_{\text{img}}$和特征词$f_{u,i}$作为初始输入序列, 并设定一个开始标记$bos$, 根据其产生的概率分布预测下一个可能出现的单词. 为了实现这一点, 本文采用了一种贪婪解码方式, 每次生成时直接选择模型认为概率最高的那个词. 随后, 将这

个预测出来的单词与序列末端相连接, 形成一个新的序列输入给模型. 这个过程可以不断重复进行, 直到模型得到一个特殊的结束标记*eos*, 或者当生成的解释语句$\hat{W}_{u,i}$达到预先设定的长度时为止.

2.5 模型优化

模型采用多任务学习框架, 通过将不同任务的损失函数进行线性组合, 通过一个加权和的形式, 将各个任务的损失整合为一个总损失, 实现联合学习. 总损失计算公式如下:

$$\mathcal{L} = \min(\lambda_e \mathcal{L}_e + \lambda_r \mathcal{L}_r + \lambda_{ct} \mathcal{L}_{ct}) \tag{13}$$

其中, λ_e, λ_r, λ_{ct}是平衡不同任务学习的正则化权重. 这种策略不仅可以提高多任务学习的效率, 还能增强模型在不同任务间的泛化能力. 通过并行训练多个相关任务, 模型能够更有效地捕捉用户行为的复杂性, 提升推荐结果的准确性, 同时生成高质量的解释.

3 实验与结果分析

3.1 数据集的构建与评价指标

3.1.1 数据集的构建

本文在 MoviesAndTV (MT)、TripAdvisor (TA) 和 ClothingShoesAndJewelry (CSJ) 这 3 个真实的可解释推荐基准数据集上评估模型性能. MT 和 CSJ 数据集均收集自 Amazon (http://jmcauley.ucsd.edu/data/amazon) 电子商务平台, 包含用户对物品的评分信息、文本评论以及物品图片信息. TA 数据集收集自旅游网站 TripAdvisor (https://www.tripadvisor.com), 包含用户对酒店的评分信息与文本评论, 由于该数据集缺乏视觉特征, 本文爬取了各酒店的图片集以补充该数据集的视觉信息. 所有数据集中的每条记录都由用户 ID、物品 ID、评分、物品特征、真实评论、用户和物品的历史评论集组成, 数据集的统计信息如表 2 所示.

表 2 数据集的统计信息

数据集	MT	CSJ	TA
用户数量	7 506	38 764	9 765
物品数量	7 360	22 919	6 280
总记录数量	441 783	179 223	320 023
特征词数量	5 399	1 162	5 069
图像数量	7 332	22 385	6 061

3.1.2 评价指标

本文同时从推荐精度和解释质量两方面评估模型性. 对于推荐精度, 采用两个常用的度量: 均方根误差 (root mean square error, *RMSE*) 和平均绝对误差 (mean absolute error, *MAE*). *RMSE* 值计算的是预测数据与原始数据对应样本误差平方和的均值, 能够衡量预测值和真实值之间的偏差, 其公式如下:

$$RMSE = \sqrt{\frac{1}{\mathcal{S}} \sum_{u,i} (\hat{r}_{u,i} - r_{u,i})^2} \tag{14}$$

MAE 值则是用来计算预测值与真实值误差绝对值的平均值. 公式为:

$$MAE = \frac{1}{\mathcal{S}} \sum_{u,i} |\hat{r}_{u,i} - r_{u,i}| \tag{15}$$

RMSE 值与 *MAE* 值越低表示模型推荐精度越高.

对于解释质量, 从两个角度来度量: 文本生成质量和个性化程度. 其中, 文本生成质量采用机器翻译中的 BLEU 和文本摘要中的 ROUGE 指标, 这两者均度量了生成的解释文本与真实文本之间的相似度, 其值越大说明生成的解释越好. 与文献[9]类似, 本文分别采用了 BLEU-1、BLEU-4、ROUGE-1 和 ROUGE-2 的 Precision、Recall、*F*1 指标对生成文本进行不同粒度的度量, 在表 3 中简写为 B1、B4、R1-P、R1-R、R1-F、R2-P、R2-R 和 R2-F.

对于第 2 个角度, 即个性化程度, 本文采用文献[8]中提出的 3 个指标进行评估, 分别是 unique sentence ratio (USR)、feature matching ratio (FMR) 以及 feature diversity (DIV). DIV 越低表明特征集之间的重叠越小, 具有较高的多样性. 对于 USR 和 FMR, 分数越高性能越好.

3.2 对比模型

为了评估 ERFMF 模型性能, 本文将其与以下 4 个基准模型进行比较.

(1) Att2Seq[25]模型. 将 MLP 与 LSTM 结合, 对一组交互信息 (即用户、物品和评分) 进行建模, 只进行解释生成.

(2) NETE[10]模型. 使用改进的 GRU 为评分信息和评论信息建模, 并生成模板式的解释句.

(3) PETER[15]模型. 结合 Transformer 模型连接协同信号和评论词, 利用协同信号来预测目标解释中的单词, 实现个性化的可解释推荐.

(4) ExBERT[26]模型. 采用 BERT 模型对评论文本、评分和协同信号进行建模, 并将用户和物品的历史评论纳入语义表示中.

表 3 不同模型的结果对比

数据集	模型	推荐精度		个性化程度			文本质量							
		RMSE↓	*MAE*↓	DIV↓	USR↑	FMR↑	B1↑	B4↑	R1-P↑	R1-R↑	R1-F↑	R2-P↑	R2-R↑	R2-F↑
TA	Att2Seq	—	—	4.32	0.17	0.06	15.27	1.03	18.97	14.72	15.92	2.40	2.03	2.09
	NETE	0.96	0.73	2.22	0.57	0.78	22.39	3.66	35.68	24.86	27.71	10.20	6.98	7.66
	PETER	0.81	**0.63**	1.61	0.25	**0.89**	24.32	4.55	37.48	29.21	30.49	11.92	8.98	9.24
	ExBERT	0.85	0.66	1.61	**0.75**	0.37	25.71	4.83	34.21	28.66	29.66	10.62	9.22	9.21
	ERFMF	**0.80**	**0.63**	**1.43**	0.49	**0.89**	**28.20**	**6.09**	**42.53**	**32.83**	**35.10**	**15.32**	**11.67**	**12.25**
MT	Att2Seq	—	—	2.74	0.33	0.12	12.56	0.95	20.79	13.31	15.35	2.62	1.78	1.99
	NETE	0.96	0.73	1.93	0.57	0.71	18.76	2.46	33.87	21.43	24.81	7.58	4.77	5.46
	PETER	0.95	0.71	1.20	0.46	**0.77**	19.75	3.06	34.71	23.99	26.35	9.04	6.23	6.71
	ExBERT	0.95	0.74	1.65	**0.82**	0.34	22.72	3.89	30.28	25.17	26.34	8.34	7.33	7.38
	ERFMF	**0.94**	**0.70**	**1.19**	0.67	**0.77**	**23.77**	**4.23**	**39.07**	**27.64**	**30.72**	**11.49**	**8.43**	**9.10**
CSJ	Att2Seq	—	—	0.14	0.03	0.05	12.83	0.85	15.57	13.41	13.37	1.85	1.70	1.60
	NETE	1.05	0.84	0.08	0.42	0.35	13.56	1.56	23.78	15.71	17.66	4.76	3.13	3.44
	PETER	1.05	0.82	**0.05**	0.22	**0.95**	22.66	4.30	38.35	29.84	30.69	12.24	9.00	9.16
	ExBERT	**1.03**	**0.80**	**0.05**	**0.43**	0.30	18.77	2.74	25.90	21.94	22.16	6.64	5.58	5.54
	ERFMF	1.04	0.82	**0.05**	0.22	**0.95**	**23.13**	**4.48**	**40.01**	**30.49**	**32.08**	**13.21**	**9.57**	**10.01**

注: 加粗部分表示最佳性能, 下划线部分表示次优性能

3.3 实验设置

本文将每个数据集按照 8:1:1 的比例随机分为训练集、验证集和测试集. 在数据预处理时对评论词按照出现频率进行了排序, 删除评论文本中的低频词, 为每个数据集构建一个保留前 20 000 个高频词的词汇表V.

本文模型中参数L和H均设置为 2, 嵌入维度d设置为 512, 评论词数量n设置为 15, 批量大小设置为 128, dropout 值设置为 0.2, 采用 SGD (stochastic gradient descent) 优化器优化模型, 正则化权重λ_e, λ_r, 和λ_{ct}分别设置为 1、0.1 和 1, 初始学习率设置为 1.

3.4 结果分析

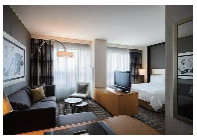
不同模型的结果对比如表 3 所示, 其中, 最佳结果值加粗显示, 次优结果带下划线显示, BLEU 和 ROUGE 为百分比值, 其他为绝对值. 在推荐精度方面, 本文模型 ERFMF 在两个较大的数据集 TripAdvisor 和 MoviesAndTV 上的推荐精度均优于其他所有对比模型, 这证明了本文推荐模块的优越性. 在 ClothingShoesAndJewelry 这个较小的数据集上, ERFMF 模型虽然没有展现出最佳的推荐性能, 但也取得了次优结果, 这可能是由于 ERFMF 模型有更多的参数可以学习, 导致它在较小数据集上存在过拟合的问题. 在个性化程度方面, ERFMF 在 DIV 指标和 FMR 指标上是最优的, 两者关注的是生成的解释句多样性, 这在实际应用中是非常重要的. 在文本质量方面, ERFMF 模型显著地优于所有对比模型, 这说明本文的模型在生成高质量句子方面的有效性, 也验证了多模态信息在指导解释生成任务上的价值. 总体而言, ERFMF 相较于其他对比模型而言, 无论是在推荐结果准确性方面, 还是在生成推荐解释质量方面, 都展现出了自身的优势.

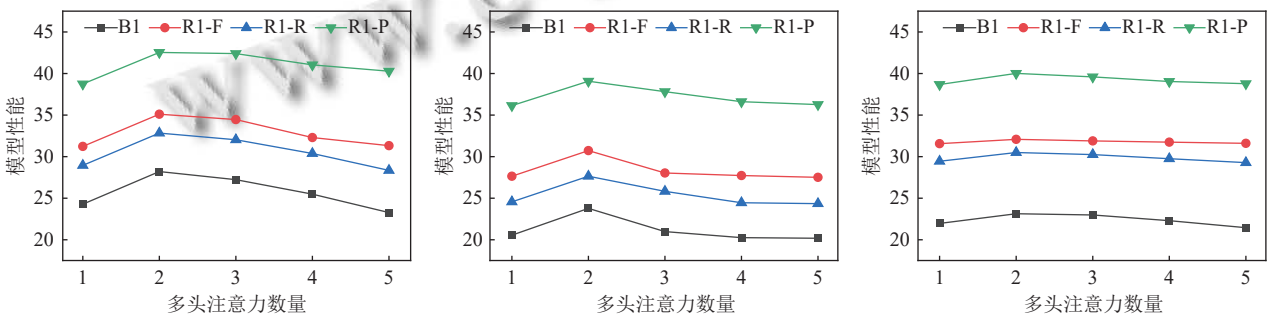
ERFMF 模型在 TripAdvisor 数据集上生成的推荐解释样例如表 4 所示, 其中, 每个样例提供了物品图像以及评分数据作为参考信息. 从中可以看出, ERFMF 模型生成的推荐解释不仅符合真实评论所表达的含义, 并且更加贴近图像所传达出的视觉信息, 能够有效地捕捉到多模态信息中的相关性, 证明了本文方法的可行性.

3.5 多头注意力数量分析

多头自注意力机制的注意力头数对提取文本的全局相关性的性能有着重要的作用, 在本节我们测试文本编码器中多头注意力数量对 ERFMF 模型性能的影响, 实验在 3 个数据集上进行了研究, 将注意力头数的取值范围设定在 1–5 之间. 从图 2 中可以发现, 当多头注意力数量设置为 2 时, ERFMF 模型在所有评估指标上均取得最佳效果, 由于所有指标趋势一致, 因此我们选取 B1、R1-F、R1-R 和 R1-P 这 4 个指标进行可视化展示. 当多头注意力数量为 1 时, 模型性能急剧下降, 不能更好地捕捉全局信息的相关性. 当多头注意力数大于 2 时, 模型性能几乎都会降低, 这是由于随着注意力头数的过多增加使建模参数同步增加, 产生了过拟合现象. 因此, 实验最后将注意力头数设置为 2.

表 4 解释语句样例

图像	评分		解释
	4.0	真实评论	The **rooms** are **nicely** done and **breakfast** was **good**.
		生成解释	The **rooms** are very **nice** and the **breakfast** was **good**.
	5.0	真实评论	**Pool** was **small** but on **roof** so that was **nice**.
		生成解释	This is a **small pool** and **good enough** on **roof**.
	1.0	真实评论	**Breakfast** was only **so-so** in terms of taste and **service** was pretty **slow**.
		生成解释	The **breakfast** was **not good** and the **service** was **slow**.



(a) TripAdvisor (b) Amazon-MoviesAndTV (c) Amazon-ClothingShoesAndJewelry

图 2 注意力头数对模型性能的影响

3.6 消融实验

为了验证模型框架中不同模块对整个模型性能的影响, 设计了两组消融实验, 实验结果如表 5 所示, 分别在 3 个数据集上加以验证. 通过设置$\mathcal{L}_{ct}=0$来移除模态对齐部分, 可以看出去除该模块后生成文本质量和个性化程度均有所下降. 该结果表明, 对齐不同模态能够使得模型学习更加合理. 模型移除用户和物品的多模态特征m_u和m_i后, 推荐精度随之降低, 解释性能急剧下降, 表明不同模态的信息可以相互补充, 多模态信息的结合可以更全面、准确地反映用户的偏好. 总体而言, ERFMF 达到了一个最优的情况, 其可解释性和推荐准确性都比较好.

表 5 消融分析结果

数据集	模型	推荐精度		个性化程度			文本质量							
		RMSE	*MAE*	DIV	USR	FMR	B1	B4	R1-P	R1-R	R1-F	R2-P	R2-R	R2-F
TA	移除$\mathcal{L}_{ct}$	0.80	0.63	1.46	0.45	0.86	28.12	5.99	42.30	32.33	34.92	14.60	11.14	11.81
	移除m_u和m_i	0.81	0.64	1.60	0.22	0.89	24.16	4.50	30.75	29.03	30.45	12.04	8.96	9.24
	ERFMF	**0.80**	**0.63**	**1.43**	**0.49**	**0.89**	**28.20**	**6.09**	**42.53**	**32.83**	**35.10**	**15.32**	**11.67**	**12.25**
MT	移除$\mathcal{L}_{ct}$	0.95	0.71	1.19	0.61	0.76	23.23	3.95	37.99	26.82	29.71	10.64	7.90	8.46
	移除m_u和m_i	0.95	0.71	1.22	0.41	0.74	19.65	3.01	35.62	23.95	26.62	9.33	6.30	6.85
	ERFMF	**0.94**	**0.70**	**1.19**	**0.67**	**0.77**	**23.77**	**4.23**	**39.07**	**27.64**	**30.72**	**11.49**	**8.43**	**9.10**
CSJ	移除$\mathcal{L}_{ct}$	1.04	0.82	0.06	0.20	0.94	22.63	4.46	39.97	29.86	31.85	13.18	9.22	9.75
	移除m_u和m_i	1.05	0.83	0.08	0.16	0.95	22.11	4.20	39.60	28.67	31.09	12.75	8.93	9.31
	ERFMF	**1.04**	**0.82**	**0.05**	**0.22**	**0.95**	**23.13**	**4.48**	**40.01**	**30.49**	**32.08**	**13.21**	**9.57**	**10.01**

4 结语

本文提出了一个融合多模态信息的可解释推荐系统, 用于深入挖掘用户偏好并缓解数据稀疏性问题. 通过结合多种模态特征, 同时进行评分预测和解释生成任务, 确保模型可以从多角度捕获用户与物品之间的关系, 并且提供了更加细粒度的推荐解释. 在真实数据集上的实验结果表明, 相比于其他基线方法. 该方法能进一步提高文本解释的质量和个性化程度. 在未来的工作中, 将进一步探索引入时间因素, 在动态场景下融合多模态信息的方法, 以便更好地捕捉数据之间的关系, 从而提高推荐解释的质量.

参考文献

1 Zanjani MD, Aghdam MH. The explainable structure of deep neural network for recommendation systems. Future Generation Computer Systems, 2024, 159: 459–473. [doi: 10.1016/j.future.2024.05.036]

2 Wei W, Huang C, Xia LH, *et al*. Multi-modal self-supervised learning for recommendation. Proceedings of the 2023 ACM Web Conference. Austin: ACM, 2023. 790–800. [doi: 10.1145/3543507.3583206]

3 冯兴杰, 曾云泽. 基于评分矩阵与评论文本的深度推荐模型. 计算机学报, 2020, 43(5): 884–900. [doi: 10.11897/SP.J.1016.2020.00884]

4 Bakariya B, Singh A, Singh H, *et al*. Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. Evolving Systems, 2024, 15(2): 641–658. [doi: 10.1007/s12530-023-09506-z]

5 Feng L, Yuan H, Ye QW, *et al*. Exploring the impacts of a recommendation system on an e-platform based on consumers' online behavioral data. Information & Management, 2024, 61(2): 103905. [doi: 10.1016/J.IM.2023.103905]

6 Zhang YF, Chen X. Explainable recommendation: A survey and new perspectives. Foundations and Trends® in Information Retrieval, 2020, 14(1): 1–101. [doi: 10.1561/1500000066]

7 Hou YF, Yang N, Wu Y, *et al*. Explainable recommendation with fusion of aspect information. World Wide Web, 2019, 22(1): 221–240. [doi: 10.1007/s11280-018-0558-1]

8 Liu CY, Wu W, Wu SY, *et al*. Social-enhanced explainable recommendation with knowledge graph. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(2): 840–853. [doi: 10.1109/TKDE.2023.3292504]

9 Park SJ, Chae DK, Bae HK, *et al*. Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation. Proceedings of the 15th ACM International Conference on Web Search and Data Mining. Arizona: ACM, 2022. 784–793. [doi: 10.1145/3488560.3498515]

10 Li L, Zhang YF, Chen L. Generate neural template explanations for recommendation. Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020. 755–764. [doi: 10.1145/3340531.3411992]

11 Zhu JH, He YJ, Zhao GS, *et al*. Joint reason generation and rating prediction for explainable recommendation. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5): 4940–4953. [doi: 10.1109/TKDE.2022.3146178]

12 Jin KF, Zhang X, Zhang JY. Learning to generate diverse and authentic reviews via an encoder-decoder model with Transformer and GRU. Proceedings of the 2019 IEEE International Conference on Big Data. Los Angeles: IEEE, 2019. 3180–3189. [doi: 10.1109/BIGDATA47090.2019.9006577]

13 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

14 Liu Z, Ma YP, Schubert M, *et al*. Multimodal contrastive Transformer for explainable recommendation. IEEE Transactions on Computational Social Systems, 2024, 11(2): 2632–2643. [doi: 10.1109/TCSS.2023.3276273]

15 Li L, Zhang YF, Chen L. Personalized Transformer for explainable recommendation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. East Stroudsburg: ACL, 2021. 4947–4957. [doi: 10.18653/V1/2021.ACL-LONG.383]

16 Xu HY, Ye QH, Yan M, *et al*. mPLUG-2: A modularized multi-modal foundation model across text, image and video. Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. 38728–38748.

17 Dong J, Li G, Ma WK, *et al*. Personalized recommendation system based on social tags in the era of Internet of Things. Journal of Intelligent Systems, 2022, 31(1): 681–689. [doi: 10.1515/jisys-2022-0053]

18 Truong QT, Lauw H. Multimodal review generation for recommender systems. Proceedings of the 2019 World Wide Web Conference. San Francisco: ACM, 2019. 1864–1874.

[doi: 10.1145/3308558.3313463]

19 Wu CH, Wu FZ, Qi T, *et al*. MM-Rec: Visiolinguistic model empowered multimodal news recommendation. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022. 2560–2564. [doi: 10.1145/3477495.3531896]

20 Geng SJ, Tan JT, Liu SC, *et al*. VIP5: Towards multimodal foundation models for recommendation. Proceedings of the 2023 Findings of the Association for Computational Linguistics. Singapore: ACL, 2023. 9606–9620. [doi: 10.18653/V1/2023.FINDINGS-EMNLP.644]

21 Wei YW, Liu WQ, Liu F, *et al*. LightGT: A light graph Transformer for multimedia recommendation. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2023. 1508–1517. [doi: 10.1145/3539618.3591716]

22 Chen ZX, Wang XT, Xie X, *et al*. Co-attentive multi-task learning for explainable recommendation. Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019. 2137–2143. [doi: 10.24963/IJCAI.2019/296]

23 Xie FF, Wang YS, Xu K, *et al*. A review-level sentiment information enhanced multitask learning approach for explainable recommendation. IEEE Transactions on Computational Social Systems, 2024, 11(5): 5925–5934. [doi: 10.1109/TCSS.2024.3376728]

24 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021. 8748–8763.

25 Li D, Huang SH, Wei FR, *et al*. Learning to generate product reviews from attributes. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia: ACL, 2017. 623–632. [doi: 10.18653/V1/E17-1059]

26 Zhan HJ, Li L, Li SH, *et al*. Towards explainable recommendation via BERT-guided explanation generator. Proceedings of the 2023 IEEE International Conference on Acoustics. Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/ICASSP49357.2023.10096389]

(校对责编: 张重毅)