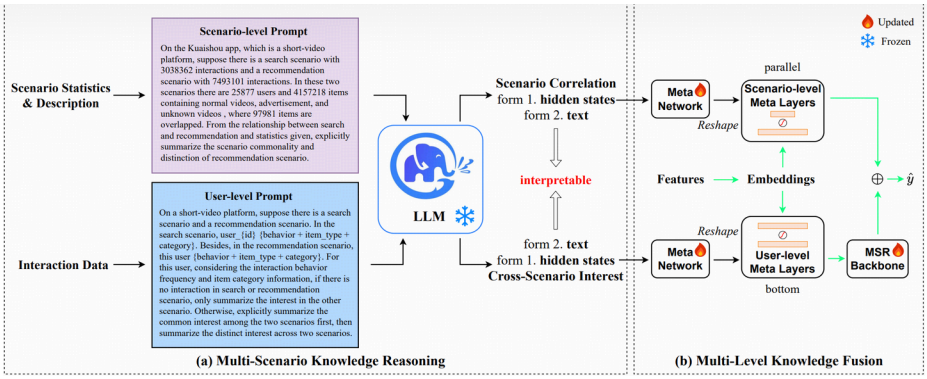


华为2024：LLM4MSR —— 利用场景与用户提示，解锁LLM在多场景推荐的应用

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我

Introduction

近年来，推荐系统（RS）研究聚焦于挖掘用户兴趣，以大量历史交互数据为依据。商业应用中，如移动搜索与推荐服务、电商平台产品分类，催生了多种业务场景（领域）。多场景推荐（MSR）作为一组高效的联合建模方法，旨在提升所有场景推荐的准确性，解决数据稀疏与降低计算成本。

现有MSR方法存在两大挑战：

1. 场景知识整合不足，领域差异仅由单一指示器捕捉，丰富的语义信息，如场景描述，未充分利用，导致场景间相关性未充分建立。
2. 忽视用户个性化偏好，多场景推荐主要依赖多任务学习的参数共享模式与传统推荐系统中的协作信号。

为解决这些问题，我们引入大型语言模型（LLM）：

1. 提出基于LLM的MSR增强方法，整合LLM与元网络，以提升主模型在多场景下的性能为目标。
2. 利用LLM理解跨场景相关性与个性化偏好，元网络作为连接LLM语义空间与推荐系统的桥梁，增强推荐系统的适应性。
3. 通过高效提示设计，利用LLM推断出用户与场景知识的高层次理解，同时解决推理延迟问题。
4. 构建一个更可解释的推荐系统框架，整合LLM与元网络，不仅提升性能，还增强透明度与理解性，使推荐结果更清晰。

Preliminary

Multi-Scenario CTR Prediction

点击通过率（CTR）预测是推荐系统的核心任务，它属于二分类。在多场景推荐系统中，基于历史交互样本$(d, \boldsymbol{x})$，预测样本的点击标签y，其中$y=1$表示点击$y=0$表示未点击。d标识样本领域$\boldsymbol{x}$包含用户和项目属性，假设有M个分类特征$\boldsymbol{x}$通过热编码转换为稠密向量$\boldsymbol{e}$。每个特征m映射到一个嵌入层，计算为 $\boldsymbol{e}_m = \boldsymbol{E}_m \cdot \boldsymbol{x}_m$

其中$\boldsymbol{E}_m$是嵌入表的维度，表示为 $\mathbb{R}^{u_m \times Dim}$ ，u_m表示特征的值计数Dim是嵌入大小。

应用推荐模型f_d后，输出预测结果 $\hat{y}$： $\hat{y} = f_d(\boldsymbol{e})$

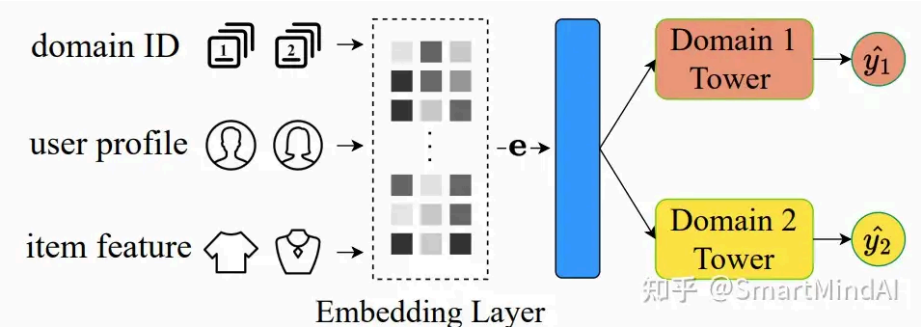
每个特征m的嵌入层通过查找操作计算$\boldsymbol{e}_m$，损失函数采用的是二元交叉熵损失或Logloss，其中$\boldsymbol{E}_m$是嵌入表的维度$\boldsymbol{x}_m$是m-th特征的热编码。损失函数使用二元交叉熵损失或Logloss。

知乎

其中，Θ是学习到的参数集B是批次样本的数量y_i是第i个样本的真实标签$\hat{y}_i$是第i个样本的预测值。

Multi-Scenario Backbone Models

多任务学习（MTL）在推荐系统中引入，旨在通过共享参数提升任务间的协同学习效果，同时提高效率。其目标是利用统一框架学习如点击率和转化率等关联任务，实现任务之间的相互提升。主流策略是设计参数共享机制，包含通用参数与任务特定参数。



多数多场景推荐模型将场景视为任务，通过维护领域共享参数Θ_s和领域特定参数Θ_u，来捕捉场景间的共性和差异。特征嵌入作为输入，经过多任务模型中的共享参数层，如MMoE的专家和STAR的中心网络，提取通用知识。接着，利用场景参数，如门控和塔式结构，生成预测结果。

Large Language Model

大多数使用流行的大语言模型，如GPT和ChatGLM2，采用解码器唯一架构，并通过自我监督训练方式，即对序列中的每个词进行训练，预测其下一个词作为目标，以提高准确性。

$$\mathcal{L}_\theta = -\sum_{i=1}^{N} \log P_\theta \left(u_i \mid u_1, \ldots, u_{i-1} \right)$$

输入文本 **u** 包含 N 个令牌，是 LLM 的参数。进行最大化在先前令牌上条件下的未来令牌的概率分布，等价于最小化上述负对数似然度。

与此同时，它采用了带mask的自注意力或所谓的因果自注意力，阻止访问未来令牌，这使得它能够理解文本数据并生成适当的结果。 具体来说，通过权重矩阵W^Q、W^K和W^V对输入的令牌向量X进行乘法操作，得到查询(Q)、键(K)和值(V)向量。然后计算当前查询向量和键向量的乘积，进一步与值向量相乘并求和：

$$score = Q \cdot K$$

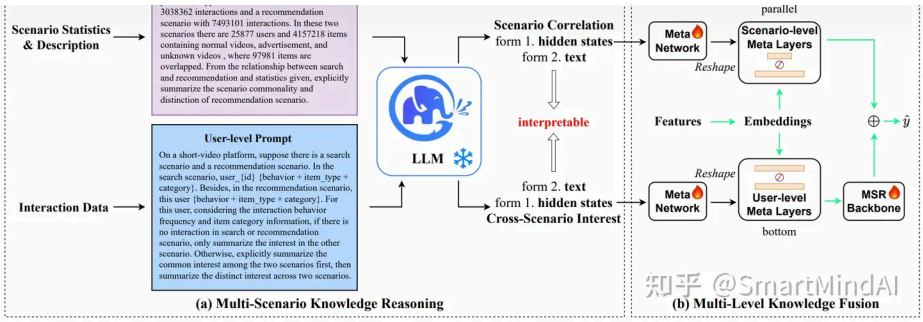
$$output = score \cdot V$$

$$\text{Attenion}(X) = W^V X \cdot \text{Softmax}\left[\frac{(W^K X)^\top (W^Q X)}{\sqrt{d_k}} + M \right]$$

其中M是掩码矩阵。分词器处理原始文本序列生成词元，然后进行嵌入。这些词元输入至包含多个前向传播和掩码自注意力块的Transformer解码器。生成最大概率的隐藏向量，并以自回归方式解码结果。

Proposed Method

Overall Framework



为解决多场景增强方法在场景知识整合和跨场景个性化兴趣方面的局限，我们提出多场景增强框架LLM4MSR。该框架包括一个锁定参数的LLM和两个可学习的元网络。算法流程概括如下：

步骤一：进行场景相关性的推理以及跨场景兴趣的判断，涉及场景关联与跨场景兴趣的评估。

步骤二：多层次知识的综合融合，以适应性地与主模型结合。

图中展示了 LLM4MSR。具体而言，场景级别的提示是通过场景统计、语义描述和专家知识生成的，而用户级别的提示则是基于历史交互构建的。随后，这些提示被输入到LLM中，以生成场景间的关联和跨场景的个性化兴趣。值得注意的是，仅选取 LLM 输出的最后一个隐藏状态作为代表，并存储，这是一个维度高达 4096 的稠密向量，例如，以 ChatGLM2-6B 为例。最后生成的令牌的状态依赖于所有之前的令牌，包含了整体的信息。最后，层次化的元网络在场景和用户级别生成元⁺层，并通过与 MSR 主干模型的训练，增强其在所有场景上的性能。 综上所述，通过优化方程，可以获得一个简单高效的一体化训练方法，从而获得最终增强的模型。

Multi-Scenario Knowledge Reasoning

主流多场景推荐模型在理解复杂场景关联时受限，实际应用中，领域指标是唯一与场景相关的信息，导致模型场景表示仅能通过隐含的协作信号来学习。为此，我们设计了场景级别的提示，通过统计和语义描述直接提取场景的共性和差异。举例，图中紫色框展示了KuaiSAR-small的交互次数、用户数量、项目数量等信息，以及背景语义信息，揭示了推荐和搜索场景之间的相似性和差异。

现有模型忽视个性化兴趣建模，难以单独捕捉用户在不同场景下的偏好。为此，我们设计用户级别提示，明确探索所有场景下用户的个性化共同与独特兴趣，利用丰富历史交互数据。积极的交互行为相较于消极行为更能精确地反映用户的真实兴趣。

负样本数量通常庞大，包含大量噪声信息，源于数据收集与用户交互的偏见和不确定性。因此，我们筛选并利用了具有积极交互行为的样本（如点击、转发、点赞）构建用户级别提示。

在获得场景级别和用户级别的提示后，输入LLM进行多场景知识推理。通过案例研究⁺，探讨了LLM能获取的多场景知识，以及这些知识如何增强多场景推荐系统的基础模型。此方法旨在提升多场景推荐系统性能，更准确理解用户偏好，提供个性化推荐服务。

Multi-Level Knowledge Fusion

现有方法可能采用直接利用LLM增强信息作为推荐系统RS的额外输入特征，通过线性投影或全连接网络进行维度转换，或者对LLM进行微调。然而，这些方法并不理想。首先，作为普通特征，LLM增强信息在推荐系统中无法发挥预期作用，因为特征数量通常很大。其次，仅进行维度对齐会导致信息损失。最后，微调LLM成本高，且需一致更新。

我们提出的方法LLM4MSR，针对从LLM获得的多场景知识信息以高维隐藏向量形式，采用层次元网络获取信息增益。元网络动态生成元层，适应性地融合场景级和用户级知识，与MSR主干模型中的合作信息相结合。场景级知识扮演通用角色，而用户级信息则个性化且重要，跨越场景界限，包含用户完整档案。

"底部 + 并行"结构，用户级元层位于底层，场景级元层并行实现，以促进不同维度的知识融合和提升。表示LLM输出的最后一个隐藏状态为$\boldsymbol{h}_{LLM}$，元工作将其作为输入，并输出Dim_{mw}和Dim_{mb}维度的元向量mw和mb，通过重塑操作总共获得K个元层结构。

知乎

$$\boldsymbol{W}_l^{(i)} = Reshape(\boldsymbol{h}_{mw})$$

$$\boldsymbol{b}_l^{(i)} = Reshape(\boldsymbol{h}_{mb}), i \in \{1, 2, \ldots, K\}$$

$\boldsymbol{W}_l^{(i)}$和$\boldsymbol{b}_l^{(i)}$是第i个层的权重和偏置矩阵。$\boldsymbol{W}_l^{(i)}$的维度等于所有权重矩阵的总和$\boldsymbol{b}_l^{(i)}$的维度等于所有偏置矩阵的总和。例如，维度为8256的$\boldsymbol{h}_{mw}$可以分解为一个维度为128×64的权重矩阵加上一个维度为64×1的权重矩阵，类似地，偏置矩阵的维度也遵循相同规则。每个层后接一个激活函数σ，如ReLU。

$$\boldsymbol{h}^{(i)} = \sigma(\boldsymbol{W}_l^{(i)} \boldsymbol{h}^{(i-1)} + \boldsymbol{b}_l^{(i)}), i \in \{1, 2, \ldots, K\}$$

在第i个元层的输出$\boldsymbol{h}^{(i)}$中$\boldsymbol{h}^{(0)}$代表初始特征嵌入。接着，将$\boldsymbol{h}^{(0)}$转移至已构建的元层。最后，通过适应性整合并行场景级别的元层输出与MSR主干结构的结果，得到预测结果。

$$\boldsymbol{h} = \text{MSR}(\boldsymbol{h}_u^{(K)})$$

$$\hat{y} = \sigma'(\alpha \cdot \boldsymbol{h}_s^{(K)} + (1 - \alpha) \cdot \boldsymbol{h})$$

在多尺度融合模型中，$\boldsymbol{h}$是输出隐藏向量$\boldsymbol{h}_u^{(K)}$和$\boldsymbol{h}_s^{(K)}$分别代表第K个用户和场景级别的元层输出。最终任务激活函数为σ'，例如，sigmoid 函数用于 CTR 预测任务。α是一个可学习的参数。

Experiments

Datasets

实验在KuaiSAR-small、KuaiSAR和Amazon三个数据集上进行。KuaiSAR-small与KuaiSAR各包含一个搜索场景和一个推荐场景。在Amazon上，我们选择了三个具体的相关推荐场景。表中汇总了数据集的统计信息和描述。点击率指标表示的是标签为0的未点击样本的比例。原始数据按照比例 $\frac{8}{9} : \frac{1}{9} : \frac{1}{9}$ 划分为训练集、验证集和测试集。

Table 1: The statistics of KuaiSAR-small, KuaiSAR, and Amazon, where Rec denotes recommendation.

Dataset	KuaiSAR-small		KuaiSAR		Amazon		
Scenarios	Rec	Search	Rec	Search	Rec#1	Rec#2	Rec#3
Users	25877	25877	25877	25877	24752	24752	24752
Items	2281034	1974165	4046363	2974596	8788	193304	6980
Interactions	7493101	3038362	14605712	4828690	33929	370840	56356
Sparsity	50.49%	87.87%	50.40%	87.88%	21.50%	19.61%	20.67%

Baselines

除了少数论文专门旨在提升多场景推荐之外，先前基于LLM的模型在相同的设置下不适用，尤其无法满足实时推理需求。我们将与两种典型范式进行基准比较，作为基准方法，以明确指出先前基于LLM的模型的局限性。

- 动态网络。它旨在以精细粒度的方式动态生成模型参数。在多场景建模时，利用场景知识产生针对特定场景的关注和塔式模块。
- EPNet是一种在PEPNet中提出的EPNet结构，它接收领域信息输入，生成针对场景的特定EPNet门，用于转换嵌入信息。

Overall Performance

我们在KuaiSAR-small、KuaiSAR和Amazon数据集上的AUC整体性能在表中显示。

知乎

STAR	0.7225	0.6089	0.7387	0.6268	0.6156	0.6320	0.6225
STAR_DN	0.7241	0.6116	0.7404	0.6270	0.6306	0.6350	0.6355
STAR_EP	0.7230	0.6082	0.7388	0.6266	0.6211	0.6302	0.6378
STAR_ours	**0.7276**★	**0.6181**★	**0.7408**★	**0.6332**★	**0.8720**★	**0.6364**★	**0.7543**★
OMoE	0.7241	0.6163	0.7394	0.6310	0.5995	0.6043	0.6252
OMoE_DN	0.7249	0.6183	0.7402	0.6319	0.6027	0.6171	0.6289
OMoE_EP	0.7246	0.6179	0.7392	0.6330	0.6143	0.6176	0.6312
OMoE_ours	**0.7265**★	**0.6186**★	**0.7413**★	**0.6332**★	**0.8182**★	**0.6180**★	**0.6823**★
MMoE	0.7235	0.6150	0.7392	0.6313	0.5907	0.5999	0.6032
MMoE_DN	0.7246	0.6164	0.7394	0.6330	0.6310	0.6201	0.6371
MMoE_EP	0.7245	**0.6178**	0.7397	0.6326	0.6218	0.6221	0.6315
MMoE_ours	**0.7264**★	0.6166★	**0.7410**★	**0.6341**★	**0.7860**★	**0.6222**★	**0.6854**★
PLE	0.7249	0.6149	0.7396	0.6313	0.6059	0.6127	0.5995
PLE_DN	0.7258	0.6165	0.7400	0.6333	0.6066	0.6195	0.6162
PLE_EP	0.7252	**0.6184**	0.7402	0.6339	0.6173	0.6229	0.6263
PLE_ours	**0.7269**★	**0.6184**★	**0.7408**★	**0.6343**★	**0.8265**★	**0.6250**★	**0.7074**★
AITM	0.7236	0.6154	0.7398	0.6329	0.6039	0.6126	0.6046
AITM_DN	0.7243	0.6172	0.7398	0.6318	0.6057	0.6151	0.6166
AITM_EP	0.7237	0.6177	0.7389	0.6337	0.6064	0.6163	0.6238
AITM_ours	**0.7273**★	**0.6178**★	**0.7407**★	**0.6349**★	**0.8196**★	**0.6178**★	**0.7089**★
Shared Bottom	0.7228	0.6169	0.7389	0.6321	0.6073	0.6077	0.6268
Shared Bottom_DN	0.7250	0.6176	0.7396	0.6323	0.6081	0.6253	0.6294
Shared Bottom_EP	0.7243	**0.6182**	0.7392	0.6331	0.6268	0.6174	0.6233
Shared Bottom_ours	**0.7269**★	**0.6182**★	**0.7400**★	**0.6341**★	**0.8323**★	**0.6262**★	**0.7182**★

首先，动态网络和EPNet在大多数情况下能够提升不同场景下的推荐准确度，但在KuaiSAR中的AITM_DN和所有数据集上的STAR_EP存在例外。相比之下，提出的，尤其是在Amazon的Rec#1和Rec#3上。它在所有场景下也始终优于基本方法，除了在KuaiSAR-small的搜索场景中基于MMoE。其优越性的原因在于，

一方面，多场景主模型和基本方法都包含了不足的场景知识，即仅包含领域ID，缺乏个性化建模。另一方面，，然后整合元网络以明确增强多场景建模能力。

此外，我们观察到，与在Amazon的Rec#1和Rec#3上显著的提升相比，#2上带来了相对较小的改进（大约0.7%）。原因是Rec#1和Rec#3相对较为稀疏，每个用户平均交互次数为1.4次和2.3次，如表所示。因此，在这些两个场景中，LLM捕捉到的用户偏好信息较少，对Rec#2的提升相对较小。

原文《LLM4MSR: An LLM-Enhanced Paradigm for Multi-Scenario Recommendation》

发布于 2024-07-30 13:20 · IP 属地北京

LLM 推荐系统 场景

理性发言，友善互动

还没有评论，发表第一个评论吧