山西大学学报(自然科学版)49(4):525—534,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2025128

基于大语言模型语义增强的移动应用使用预测方法

康梅英,尚若冰,陈文亮*
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:应用使用预测是智能手机系统优化中的关键技术之一。本文提出一种基于大语言模型语义增强的预测方法,旨在提升模型对用户使用意图的理解能力与分布外泛化能力。通过引入大语言模型自动生成应用功能描述,并结合嵌入模型将文本描述编码为高维语义向量,作为用户历史使用序列的语义表征,输入下游任务模型进行预测。为了更有效地捕捉用户行为中的时序动态与语义依赖,本文设计了一种融合局部激活注意力机制与Mamba-2模块的预测模型。实验结果表明,在真实用户应用使用数据集上,该方法在分布内数据上Top-3准确率由76.41%提升至77.00%,在分布外数据上Top-3准确率由66.06%提升至71.57%,有效提升了模型对未见应用的泛化能力。

关键词:分布外泛化;嵌入模型;Mamba-2
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2026)04-0525-10

Semantic-enhanced Application Usage Prediction *via* Large Language Models

Kang Meiying, Shang Ruobing, Chen Wenliang*
(*School of Computer Science and Technology, Soochow University, Suzhou* 215006*, China*)

Abstract: Application usage prediction is a crucial technique for optimizing smartphone systems. This study proposes a semantic-enhanced prediction method based on large language models to improve the understanding of user intent and generalization to out-of-distribution applications. The method used a large language model to automatically generate functional descriptions of applications. These descriptions were encoded into high-dimensional semantic vectors through an embedding model and used as semantic representations of historical usage sequences. These representations were then fed into the prediction model. To better capture temporal dynamics and semantic dependencies in user behavior, a prediction model was designed by integrating locally activated attention mechanisms with a Mamba-2 module. Experimental results showed that on a real-world application usage dataset, the proposed method increased the Top-3 accuracy from 76.41% to 77.00% on in-distribution data and from 66.06% to 71.57% on out-of-distribution data, significantly improving the model's generalization to unseen applications.

Key words: out-of-distribution generalization; embedding model; Mamba-2

0 引言

随着智能设备的普及与移动应用生态的迅速扩展,用户对智能终端的依赖日益增强[1]。移动应用(Application,APP)使用预测作为提升系统响应效率、优化资源分配与改善用户体验的重要技术手段,近年来受到了广泛关注。通过提前预测用户即将使用的APP,系统能够执行预加载、内存管理和能耗优化等多项系统级优化策略,在操作系统智能化方面具有重要实

收稿日期:2025-11-09;**修回日期**:2025-12-13
基金项目:国家自然科学基金(62036004);江苏高校优势学科建设工程资助项目
作者简介:康梅英(2001—),女,江西赣州人,硕士研究生,研究方向为意图预测。E-mail:mykang@stu.suda.edu.cn
* **通信作者**:陈文亮(Chen Wenliang),E-mail:wlchen@suda.edu.cn
引文格式:康梅英,尚若冰,陈文亮.基于大语言模型语义增强的移动应用使用预测方法[J].山西大学学报(自然科学版),2026,49(4):525-534. DOI:10.13451/j.sxu.ns.2025128.

际意义。

近年来,深度学习的发展推动基于序列建模的APP使用预测方法取得显著进展[2-5]。部分研究引入图表征学习[6-9],挖掘APP间的潜在关联关系提升性能。然而,现有方法在实际部署中仍面临严峻挑战,主要体现在对分布外(Out-of-Distribution, OOD)APP的泛化能力不足。分布外APP是指在训练阶段未曾出现,但在部署阶段实际被用户使用的APP。传统固定类别预测方法难以为其分配有效表示,导致性能下降,且边缘场景中此类问题更为突出。

此外,现有方法普遍依赖结构化标签体系(如APP类别、使用频率等)对APP进行建模,但这类粗粒度标签难以反映APP功能的细粒度差异。例如,“社交类”标签下既包含以信息高效传递和协作办公为目标的工作沟通类APP,也包含强调趣味性和非正式交流的娱乐聊天类APP。尽管归属于同一类别,用户在实际使用中的行为意图与使用场景存在显著差异。由于缺乏对APP语义的深入理解,现有方法在建模用户真实使用意图时存在局限,语义层面的表达缺失成为进一步提升预测性能的主要瓶颈。

针对现有方法在用户意图理解不充分与分布外泛化能力不足两方面的局限,本文提出了一种基于大语言模型(Large Language Model, LLM)语义增强的移动应用使用预测方法AppSE(Application Usage Prediction via Semantic Enhanced),通过引入语义增强机制,为预测模型提供更细粒度、通用性更强的特征表达,从而增强模型对用户行为的建模能力与对新APP的泛化能力。AppSE充分利用LLM强大的自然语言理解和生成能力,为APP生成功能描述文本,通过文本向量表示实现语义增强,在此基础上使用轻量高效预测模型实现高效端侧推理。本文主要贡献如下:

1)提出结合LLM与语义编码机制的特征表示方法,突破传统结构化标签在APP语义建模方面的表达瓶颈,为预测模型提供更具区分性与泛化能力的输入特征。

2)设计融合局部激活注意力机制与Mamba-2模块的预测架构,在保持模型表达能力的同时提高计算效率。

3)在真实用户APP使用数据集上进行系统评估,结果显示所提出方法在预测准确率方面优于现有基准方法,尤其在分布外APP预测任务中展现出良好的适应能力,验证了其在实际部署环境中的应用潜力。

1 相关工作

APP使用预测作为智能手机系统优化中的关键任务,旨在基于用户历史行为及上下文信息,预测其下一步可能使用的APP[10]。

早期研究[11-12]多采用基于统计特征的传统方法,如最常使用(Most Frequently Used, MFU)和最近使用(Most Recently Used, MRU)策略,分别依据使用频率与最近访问时间对APP进行排序。这类方法实现简单、计算开销低,常作为后续复杂模型的性能参考基准。在此基础上,研究者进一步引入概率建模方法以提升预测精度。例如,Shin等[13]采用朴素贝叶斯模型融合多种上下文信息,Huang等[14]利用一阶马尔可夫模型建模APP间的转移概率,Liao等[15]则基于最小描述长度准则进行个性化特征选择,迭代挑选最有效特征后,使用最近邻算法进行预测。虽然该类方法具有一定实用性,但在表达复杂行为模式方面能力有限。

随着深度学习的发展,序列建模成为主流方向。Zhao等[16]提出AppUsage2Vec,借鉴Doc2Vec架构,引入注意力机制与双通道神经网络联合建模用户历史行为与个体特征;Xu等[17]基于长短时记忆网络(Long Short-Term Memory, LSTM)融合使用序列与上下文信息,提升了模型的时序建模能力;Zhang等[18]将预测任务转化为点击率排序问题,缓解了类别不平衡带来的影响。与此同时,图表示学习方法被引入以建模用户、APP与上下文之间的高阶关系结构。Wang等[19]提出SGFNN(Sequence-Graph Fusion Neural Network)模型,融合序列建模与图学习方法,通过序列块建模短期行为,通过二部图块与超图块学习用户与APP的相似性及高阶语义关系,展现了图结构与时序建模之间的互补性。

尽管上述方法在用户行为建模方面取得了

显著进展,但仍主要依赖APP标识符、类别标签或人工特征,难以从语义层面反映用户行为背后的动机,限制了模型对分布外场景的泛化能力。Khaokaew等[20]提出MALPE(Mobile App Prediction Leveraging Large Language Model Embeddings)模型,首次将LLM应用于APP使用预测,将用户历史行为序列格式化为自然语言输入,利用LLM推理能力直接输出预测结果,在分布外任务中表现出较强适应性。

然而,MALPE在实际应用中存在明显局限。一方面,LLM模型参数庞大、推理开销高,难以部署于资源受限的移动终端[21],限制边缘场景实用性。另一方面,MALPE依赖预定义输入模板将用户行为序列转为自然语言,模板设计需人工完成,工程成本高且灵活性不足,难以适应多变的任务和数据环境[22]。因此,如何高效、灵活地将其语义理解能力引入到轻量化、可泛化的预测框架中,仍是当前研究的核心挑战。

2 方法

本文整体方法框架如图1所示,主要由两个关键模块构成:语义增强模块与预测模块。

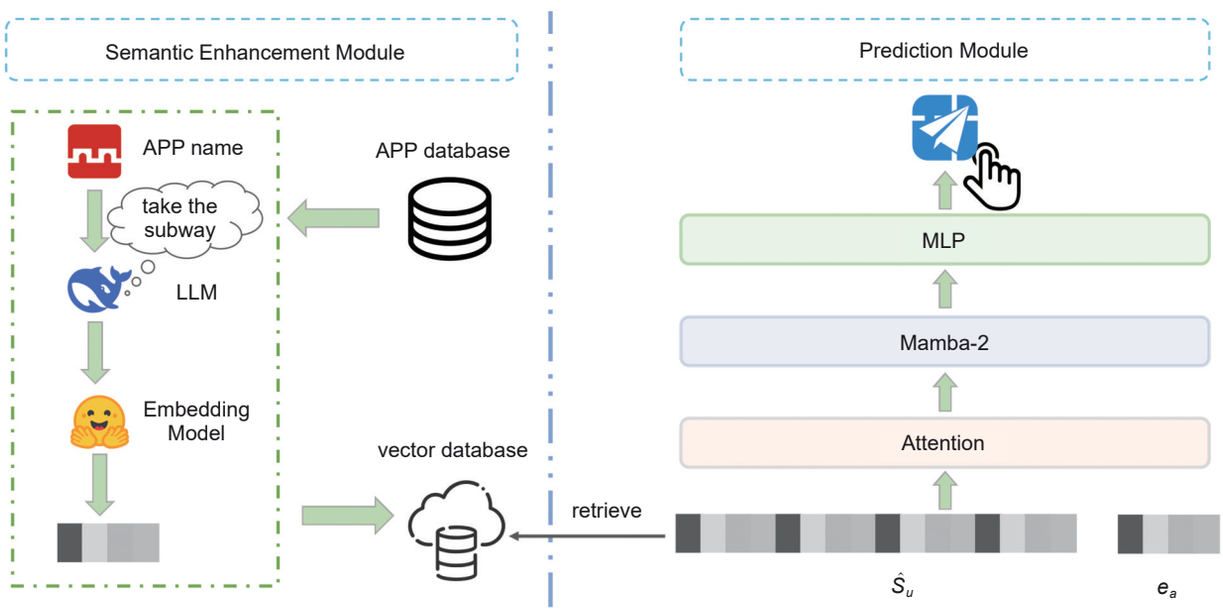


图1 基于大语言模型语义增强的移动应用使用预测方法框架

Fig. 1 A framework for semantic-enhanced mobile application prediction *via* large language models

2.1 问题定义

本文将APP使用预测建模为排序问题,旨在基于用户的历史行为,为其生成一个按点击概率排序的前K个推荐APP(Top-K推荐),用于预测用户下一时刻最有可能使用的APP。

设用户u已安装的APP集合记为$A_u \subseteq A$,其中A表示所有APP的全集。根据用户u最近的n次APP使用记录,构建有序序列$\boldsymbol{S}_u=(a_1,a_2,\cdots,a_n)$,其中$a_i \in A_u$表示用户$u$在时刻$t_i$使用的APP,且满足时间约束$t_1<t_2<\cdots<t_n$。

对于任意候选应用程序$a \in A_u$,预测模型学习条件概率函数$f(\boldsymbol{S}_u,a)\rightarrow p\in[0,1]$,估计用户在下一个时刻$t_{n+1}$使用应用程序$a$的概率,记作:

$$f_u(a)=p(a|\boldsymbol{S}_u)。\tag{1}$$

最终,根据预测概率对所有候选APP进行降序排序,生成用户u的Top-K推荐列表:

$$\boldsymbol{L}_u^K=\operatorname{argsort}_{a\in A_u} f(\boldsymbol{S}_u,a)[1{:}K]。\tag{2}$$

2.2 语义增强模块

2.2.1 APP语义描述生成

LLM凭借其强大的语义理解与泛化能力,为解决复杂场景下的用户意图建模提供了新的思路。诸如豆包[23]、通义千问[24]等已在多种自然语言处理任务中展现出对用户意图的精准建模能力。

为了增强模型对用户意图的理解,本文利用LLM生成APP的功能描述文本。具体方法是向LLM输入提示词,指导其概括APP的核心

功能。所设计的提示词为“你是一个专业的应用程序评价师。请你从应用程序的主要功能角度介绍应用程序,需要简洁易读,在四十个字以内。”同时,提供少量示例以辅助LLM生成高质量描述。

借助LLM生成的APP描述,预测模型能够对用户使用行为背后的潜在动机进行更深层次的语义建模。例如,对于“西安地铁”对应的描述为:“《西安地铁》是一款支持刷码乘车、线路查询及实时到站信息的交通服务应用。”该描述揭示了“西安地铁”APP的功能定位和典型使用场景,有助于预测模型识别用户的出行需求。通过引入APP功能语义信息后,模型能更好地区分出行、娱乐、办公等不同场景,从而提升预测与推荐的精度。

2.2.2 APP语义向量编码

为有效捕捉APP描述中的语义信息,本文采用Embedding模型对描述文本进行编码,生成高维语义向量表示。在预测任务中,每次APP点击被视为一个原子行为。为避免对相同行为重复编码造成的计算冗余,本文选择对每个APP的描述进行编码,而非直接对APP序列进行编码。这种方式在降低计算量的同时保留每个原子行为的完整语义表达。

以T5[25](Text-to-Text Transfer Transformer,T5)模型为例,我们首先对输入的APP描述文本进行预处理,包括分词和规范化处理。处理后的文本被传递给T5模型进行前向传播,模型输出每个位置在最后一层的隐藏状态向量$\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_n\}$,其中$\boldsymbol{h}_i\in\mathbb{R}^d$表示第$i$个位置的语义表示。为了消除填充部分对语义向量计算的干扰,我们引入掩蔽标识向量$\boldsymbol{m}_i\in\{0,1\}$指示有效位置,并采用掩蔽均值池化策略计算最终的语义向量$\boldsymbol{e}_a$:

$$\boldsymbol{e}_a=\mathrm{LayerNorm}\left(\frac{\sum_{i=1}^{n}(\boldsymbol{h}_i\cdot\boldsymbol{m}_i)}{\sum_{i=1}^{n}\boldsymbol{m}_i}\right),\qquad(3)$$

其中LayerNorm表示层归一化操作,用于提升数值稳定性与表示质量,最终得到的语义向量$\boldsymbol{e}_a\in\mathbb{R}^d$用于表征应用程序$a$的语义特征。最终,所有APP的语义向量构成一个可检索的语义知识库$\boldsymbol{E}=\{\boldsymbol{e}_a|a\in A\}$,其中$\boldsymbol{E}$存储在离线数据库中,供后续任务检索使用。

2.3 预测模块

为有效建模用户历史行为序列与候选应用程序之间的交互关系,提升模型对序列的建模能力,本文设计了融合局部激活注意力机制[26]与Mamba-2[27]模块的预测模型。

2.3.1 输入表示

给定用户u的历史行为序列$\boldsymbol{S}_u=(a_1,a_2,\cdots,a_n)$和候选应用程序$a\in A_u$,模型首先从离线构建的语义知识库$\boldsymbol{E}=\{\boldsymbol{e}_a|a\in A\}$中检索相应的语义向量作为嵌入表示。具体地,用户历史序列的嵌入表示记为

$$\hat{\boldsymbol{S}}_u=\left(\boldsymbol{e}_{a_1},\boldsymbol{e}_{a_2},\cdots,\boldsymbol{e}_{a_n}\right)\in\mathbb{R}^{n\times d},\qquad(4)$$

其中$\boldsymbol{e}_{a_i}\in\mathbb{R}^d$表示序列的第$i$个APP的嵌入,候选应用程序$a$的嵌入向量为$\boldsymbol{e}_a\in\mathbb{R}^d$。

2.3.2 局部激活注意力机制

为了捕捉候选应用程序与用户历史行为序列之间的非线性交互关系,本文引入局部激活注意力机制。该机制以候选应用程序嵌入$\boldsymbol{e}_a$与历史行为嵌入序列$\hat{\boldsymbol{S}}_u$为输入,通过计算局部相关性得分,选择性地激活对当前候选应用程序相关的兴趣区域,从而提高预测的准确性。

在相关性得分计算阶段,首先构建候选应用程序嵌入、历史行为嵌入、差异特征与乘积特征四类语义特征,并在特征维度上拼接为输入张量:

$$\boldsymbol{z}_{\mathrm{con}}=\left[\boldsymbol{e}_a^{\oplus};\ \hat{\boldsymbol{S}}_u;\ \boldsymbol{e}_a^{\oplus}-\hat{\boldsymbol{S}}_u;\ \boldsymbol{e}_a^{\oplus}\odot\hat{\boldsymbol{S}}_u\right]\in\mathbb{R}^{n\times 4d},\qquad(5)$$

其中$\boldsymbol{e}_a^{\oplus}\in\mathbb{R}^{n\times d}$表示将候选应用程序嵌入向量$\boldsymbol{e}_a$沿时间维度复制$n$次后所得的矩阵,$\odot$表示逐元素乘法,$[\cdot;\cdot]$表示在特征维度上的拼接操作。

随后将$\boldsymbol{z}_{con}$输入前馈神经网络,计算每个历史行为的局部相关性得分:

$$\boldsymbol{s}_{\mathrm{local}}=\boldsymbol{W}^s\cdot\mathrm{ReLU}\left(\boldsymbol{W}^h\cdot\boldsymbol{z}_{\mathrm{con}}+\boldsymbol{b}^h\right)+\boldsymbol{b}^s,\qquad(6)$$

其中$\boldsymbol{W}^h\in\mathbb{R}^{h\times 4d}$和$\boldsymbol{W}^s\in\mathbb{R}^{1\times h}$分别为可学习的权重矩阵,$\boldsymbol{b}^h\in\mathbb{R}^h$与$\boldsymbol{b}^s\in\mathbb{R}$为对应的偏置向量,$\mathrm{ReLU}(\cdot)$为激活函数。

注意力权重基于$\boldsymbol{s}_{\mathrm{local}}$进行归一化计算,其中缩放因子$\sqrt{d}$用于缓解梯度消失,掩码向量$\boldsymbol{m}\in\{0,1\}^n$用于抑制填充位置对注意力分布的干扰:

$$\boldsymbol{\alpha}=\mathrm{softmax}\left(\frac{\boldsymbol{s}_{\mathrm{local}}}{\sqrt{d}}\odot\boldsymbol{m}\right)\in\mathbb{R}^n。\qquad(7)$$

最终,利用$\boldsymbol{\alpha}$对历史行为嵌入进行加权,得到与候选应用程序语义相关的历史表示向量:

$$\boldsymbol{X}=\boldsymbol{\alpha}\odot\hat{\boldsymbol{S}}_u=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\mathbb{R}^{n\times d},\quad(8)$$

该表示有效整合了与候选应用高度相关的历史语义信息,有助于提升预测性能。

2.3.3 Mamba-2模块

为提升模型对用户动态意图的建模能力,本文引入基于结构化状态空间模型(Structured State Space Models, SSMs)范式的Mamba-2[27]模块。针对APP使用预测中用户行为的时序特性,Mamba-2通过半可分矩阵的低秩分解与选择性记忆机制,增强对长期行为模式和意图迁移的感知能力。相较于传统Transformer架构,Mamba-2采用线性变换机制,将序列建模的计算复杂度从二次降为线性,显著降低了计算开销。

本文将局部激活注意力机制输出的加权历史行为嵌入序列$\boldsymbol{X}$输入至Mamba-2模块。首先通过线性变换矩阵$\boldsymbol{W}_{\text{proj}}$投影为多个中间分支:

$$[\boldsymbol{z},\boldsymbol{x},\boldsymbol{B},\boldsymbol{C},\Delta\boldsymbol{t}]=\boldsymbol{W}_{\text{proj}}\boldsymbol{X},\quad(9)$$

其中$\boldsymbol{z}\in\mathbb{R}^{n\times d}$为门控分支,用于控制信息流强度;$\boldsymbol{x}\in\mathbb{R}^{n\times d}$为主分支,承载核心序列信息;$\boldsymbol{B},\boldsymbol{C}\in\mathbb{R}^{n\times g\cdot s}$为状态空间控制参数;$\Delta\boldsymbol{t}\in\mathbb{R}^{n\times h}$为每个注意力头对应的时间步调节参数。参数$d,g,s,h$分别控制维度、分组数、状态维度和注意力头数。

为强化局部上下文建模且保持因果性,模型对$[\boldsymbol{x},\boldsymbol{B},\boldsymbol{C}]$应用分组因果卷积,卷积核大小$d_{\text{conv}}$决定上下文窗口,分组卷积确保各通道独立处理。

在状态空间建模阶段,采用结构化动态系统对隐状态进行更新:

$$\boldsymbol{s}_t=\boldsymbol{W}_t^A\cdot\boldsymbol{s}_{t-1}+\boldsymbol{B}'_t\cdot\boldsymbol{x}'_t,\quad\boldsymbol{y}_t=\boldsymbol{D}\cdot\boldsymbol{s}_t,\quad(10)$$

其中$\boldsymbol{s}_t\in\mathbb{R}^{g\times s}$为每个分组下的隐状态向量(初始状态为0);$\boldsymbol{W}_t^A=-\exp(\log\boldsymbol{W}^A)$为基于可学习权重矩阵$\boldsymbol{W}^A\in\mathbb{R}^h$生成的负衰减因子,确保其为负值以实现有效的记忆控制;$\boldsymbol{D}\in\mathbb{R}^h$为可学习的残差调节向量,用于平衡历史信息的保留与新信息的引入。

为增强时间建模的灵活性,引入可学习的时间步长调节机制。通过可学习偏移量$\Delta\boldsymbol{t}_{\text{bias}}$调整时间间隔得到步长:

$$\widetilde{\Delta\boldsymbol{t}}=\text{softplus}(\Delta\boldsymbol{t}+\Delta\boldsymbol{t}_{\text{bias}}),\quad(11)$$

其中softplus激活函数用于确保步长为正数,从而保持时间建模的物理合理性。

随后输出主分支结果$\boldsymbol{Y}\in\mathbb{R}^{n\times d}$,与门控分支$\boldsymbol{z}$融合后经归一化处理:

$$\hat{\boldsymbol{y}}_t=\text{RMSNorm}\big(\boldsymbol{Y}\odot\sigma(\boldsymbol{z}),\epsilon=1\times10^{-5}\big),\quad(12)$$

其中$\sigma(\cdot)$表示Swish激活函数。最终通过输出映射矩阵$\boldsymbol{W}_{\text{out}}\in\mathbb{R}^{d\times d'}$还原为原始维度:

$$\boldsymbol{H}=\boldsymbol{W}_{\text{out}}\hat{\boldsymbol{y}}_t\in\mathbb{R}^{n\times d}。\quad(13)$$

$\boldsymbol{H}$表示在候选应用程序语义条件下,用户历史行为的动态建模结果。最终选取序列中最后一个有效时间步表示$\boldsymbol{h}^m$作为后续决策的用户状态表示。

2.3.4 全连接层和分类输出

在APP使用预测中,全连接层(Multi-Layer Perceptron, MLP)作为模型的最终决策模块,根据用户历史行为序列预测候选应用程序的点击概率。具体地,将候选应用程序的嵌入向量$\boldsymbol{e}_a$与用户历史序列的最终表示$\boldsymbol{h}^m$进行拼接,输入多层感知机进行非线性变换,输出该候选应用程序在下一时刻被用户使用的概率:

$$\hat{p}=\text{MLP}\big([\boldsymbol{e}_a;\boldsymbol{h}^m]\big)\in(0,1),\quad(14)$$

其中$\hat{p}$表示在历史序列条件下用户使用候选应用程序a的预测概率。对于每个样本,将所有候选应用程序的概率排序,选取概率最高的Top-K项作为最终的推荐结果。

3 实验设计

3.1 数据集

本文实验使用来自中国手机厂商在两个不同时间段采集的真实世界大规模APP使用日志数据[18]。该数据采集均基于真实移动终端设备。为保护用户隐私,所有用户标识信息均已进行匿名化处理,数据中不包含任何可识别个人身份的内容。数据的基本统计信息如表1所示。

表1 数据集概览

Table 1 Overview of the datasets

数据集版本	时期	用户数量	APP数量
2022-data	2022/03/02—2022/04/27	50 000	76 460
2023-data	2023/02/05—2023/05/05	200 000	180 671

本文选取用户最近使用的16个APP作为其历史行为序列,并分别从两个数据集中各抽

取10 000名用户数据用于模型训练与测试。训练集使用2022年4月1日至4月7日的数据,4月8日的数据作为标准测试集;同时构建一个分布外测试集,该集采用2023年3月1日至3月7日的数据,并确保预测应用中仅包含模型未训练过的样本,以评估其在分布外环境下的泛化能力。

3.2 基准模型

为全面评估所提方法的有效性,本文选取以下具有代表性的基准模型进行对比实验:(1)MFU:基于用户历史中使用频率最高的APP进行预测,反映用户长期使用偏好。(2)MRU:依据用户最近一次使用的APP进行预测,强调用户近期行为特征。(3)AppUsage2Vec[16]:融合注意力机制与时间上下文建模的深度学习方法,可有效挖掘历史序列中的关键行为。(4)SGFNN[19]:通过构建用户-APP图结构,并结合图神经网络与Transformer编码器联合建模,捕捉行为序列与语义之间的多层次关联。(5)CTR-RAD(Click-through Rate Ranking-based Application Usage Prediction)[18]:将APP使用预测任务转化为点击率预估问题,融合多种行为和上下文特征,提升低频APP预测能力。(6)MAPLE[20]:将用户历史行为序列转化为类自然语言格式输入,利用LLM的推理能力进行APP预测。

3.3 评价指标

本文采用Top-K准确率作为模型性能评估指标,该指标在APP使用预测任务中被广泛采用,用于衡量推荐列表中是否包含用户实际下一次使用的APP。Top-K准确率的计算公式如下所示:

$$A_{\text{Top-}K}=\frac{N\left(a_{n+1}\in L_u^K\right)}{N}, \tag{15}$$

其中N表示测试样本总数,$N(\cdot)$表示满足条件的样本数量。a_{n+1}表示用户实际在下一时刻使用的目标APP。当$a_{n+1}\in L_u^K$时,表示该样本预测正确。

4 实验结果

4.1 APP使用预测

本文所有实验均独立重复五次,结果取平均值以确保可靠性。表2展示了本文方法与多个基准模型在标准数据集和分布外数据集上的性能对比结果。总体来看,本文方法在整体评估指标上均优于现有基准模型,展现出显著的性能优势。

表2 对比实验结果(%)

Table 2 Results of comparative experiments (%)

数据	模型	Top-1	Top-3	Top-5	Top-10
标准数据	MFU	38.77	69.55	78.44	80.59
	MRU	22.31	64.50	72.65	79.77
	AppUsage2Vec	36.88	70.20	79.00	86.52
	SGFNN	46.25	73.14	80.49	87.28
	CTR-RAD	40.24	75.21	83.58	<u>89.70</u>
	AppSE	47.35	76.35	<u>83.88</u>	89.61
	MAPLE(10%)	<u>47.91</u>	<u>76.41</u>	83.59	89.00
	AppSE(10%)	**48.14**	**77.00**	**84.31**	**89.90**
分布外数据	MFU	27.27	61.71	<u>73.06</u>	<u>76.14</u>
	MRU	25.79	64.82	71.23	77.29
	AppUsage2Vec	0.00	0.00	0.00	0.00
	SGFNN	0.00	0.00	0.00	0.00
	CTR-RAD	0.02	5.79	17.88	41.71
	MAPLE	<u>45.45</u>	<u>66.06</u>	71.05	75.45
	AppSE	**45.85**	**71.57**	**77.99**	**81.96**

注:表中数值表示模型在APP使用预测任务中的Top-K准确率(百分比)。Top-K准确率用于衡量推荐列表前K个结果中是否包含用户实际下一次使用的APP。表中加粗部分为最优结果,下划线部分为次优结果。

Note: The values represent the Top-K accuracy (percentage) in the APP usage prediction task. Top-K accuracy measures whether the user's actual next APP appears within the Top-K recommended list. Bold entries indicate the best performance, and underlined entries indicate the second-best performance.

本文方法仅以APP使用序列作为输入。为保留SGFNN的原始设计,实验中额外引入用户过去一天对不同APP的偏好信息,将其作为图结构的输入特征以提升模型表达能力。在此配置下,SGFNN在Top-1准确率上表现较为突出,但Top-5准确率仍略逊于仅基于序列建模的CTR-RAD方法。针对MAPLE,考虑到其两阶段推理需求,实验中补充引入APP类型及用户完整APP列表作为辅助输入特征,以增强模型对用户行为上下文的建模能力。

此外,AppUsage2Vec与SGFNN均属于有限类别预测模型,难以对训练阶段未出现的分布外APP进行有效预测。相比之下,CTR-RAD通过将APP预测任务转化为点击率排序问题,

具备一定的分布外APP预测能力。受训练资源限制,本文仅使用10%的用户数据对MAPLE进行训练和测试。MAPLE模型依托LLM的语义理解与推理能力,性能表现不俗,但仍略逊于本文模型,尤其在分布外任务的Top-5指标上低了6.94%。各基线方法与本文方法的推理时间对比见表3,其中MAPLE的推理时间是本文方法的300余倍。

表3 推理时间

Table 3 Inference time

模型	AppUsage2Vec	SGFNN	CTR-RAD	MAPLE	w/o SE	AppSE
平均推理时间/ms	8.89	12.12	1.65	419.53	**1.33**	7.22

注:表中加粗部分为最优结果,下划线部分为次优结果。

4.2 消融实验

为进一步验证所提方法中各模块的有效性,本文设计了消融实验,评估不同模块对最终预测性能的贡献。本文的语义增强部分包括APP描述生成模块和APP语义编码模块;下游任务模型由Attention机制与Mamba-2模块组成。文中符号约定如下:SE(Semantic Enhanced)表示完整的语义增强模块,ADG(Application Description Generation)表示APP描述生成模块。

如表4所示,消融实验结果表明,SE模块为分布外应用提供了丰富且有效的语义向量表示,使模型能够更精准地识别和区分这些新应用,从而显著提升了分布外APP的预测性能,Top-1准确率提升约44.9%。与此同时,借助LLM生成的APP功能描述,预测模型能够更深入地捕捉用户使用行为背后的潜在动机。

表4 消融实验结果

Table 4 Results of ablation study

模块	标准数据				分布外数据			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
w/o SE	44.98	75.38	83.51	89.67	0.95	11.79	27.21	48.99
w/o ADG	46.50	75.57	83.32	89.48	42.32	70.07	76.62	80.83
w/o Attention	43.63	72.48	80.74	87.84	26.73	55.50	66.23	76.13
w/o Mamba-2	39.29	69.60	81.57	89.38	12.75	66.00	76.06	80.35
AppSE	**48.14**	**77.00**	**84.31**	**89.90**	**45.85**	**71.57**	**77.99**	**81.96**

注:同表3。

在下游建模部分,Attention与Mamba-2模块在不同指标上展现出互补特性:Mamba-2模块在Top-1准确率上表现更优,适合识别用户最有可能使用的APP;Attention机制则在Top-5准确率上更具优势,能更好地捕捉潜在多样的用户意图。将两者联合建模进一步提升了整体性能,体现出两种结构之间的协同效应。

为进一步验证ADG模块在语义表达方面的有效性,本文采用PCA降维结合t-SNE方法对生成的APP语义向量进行了低维空间的可视化分析(见图2)。结果显示,语义增强后的APP语义向量在低维空间中呈现出显著的聚类特征。图中功能相似的APP聚集成簇,充分体现了该语义表示对APP间功能关联性的有效捕捉。相比之下,仅基于APP名称生成的语义向量分布较为分散,缺乏明显的聚类结构,难以反映APP功能的内在关联性。

4.3 对比实验

4.3.1 LLM模型对比实验

为评估不同LLM在APP描述生成任务中的表现,本文选取了当前主流的中文语言模型,包括 DeepSeek-R1-Distill-Qwen-1.5B[28] 、DeepSeek-R1-Distill-Qwen-7B[28]、豆包(Doubao)[23]和通义千问(Qwen)[24],实验结果如表5所示。

从下游模型的预测性能来看,各模型生成的描述在语义嵌入质量上存在差异,其中豆包的表现最优。结合人工对生成内容的观察发现,参数规模较小的DeepSeek-R1-Distill-Qwen-1.5B表现出较明显的特殊性:其生成的应用程序功能描述往往未严格遵循提示词要求,输出

内容偏多且信息量冗余,人工可读性较差。但该模型在下游预测性能上反而表现较好。这一现象可能源于Embedding模型对语义表示的强鲁棒性——即便输入描述存在形式上的冗余或不规范,仍能有效提取核心信息,从而保证了下游任务的性能稳定。

4.3.2 Embedding模型对比实验

本文在APP语义编码模块中选取了多种代表性架构的Embedding模型进行对比实验,包括Encoder架构的BERT[29]、Decoder架构的GPT-2[30]、Encoder-Decoder架构的T5[25]、专门针对文本嵌入任务优化的模型Text2Vec[31]以及基于LLM的Qwen3[32]。实验结果如表6所示,T5、和Text2Vec在APP描述的语义表示方面表现突出,生成的嵌入更有利于下游任务的建模;相比之下,BERT在本任务中的表现相对较弱,可能由于其预训练目标与当前任务的适配性不足所致。

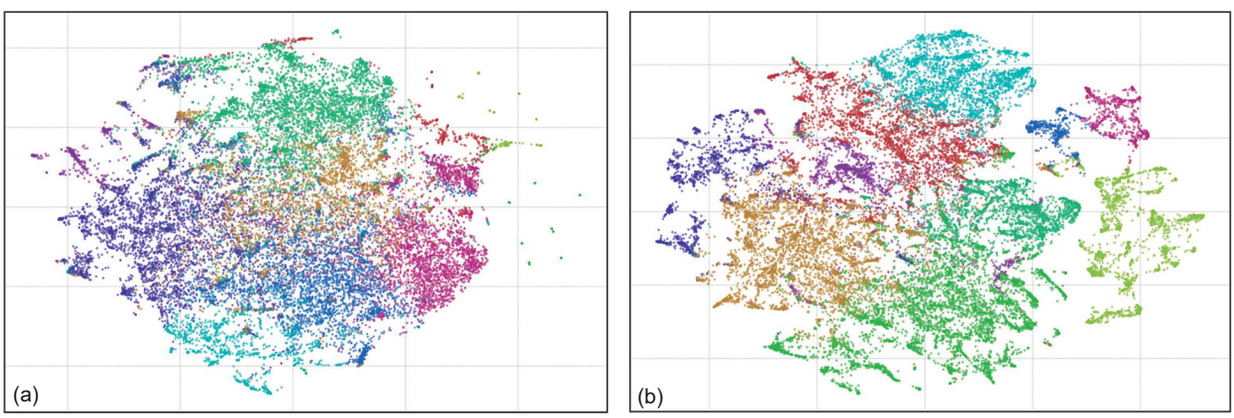


图2 语义增强前的APP语义向量(a)与语义增强后的APP语义向量(b)可视化对比

Fig. 2 Visualization comparison of APP embedding vectors before (a) and after (b) semantic enhancement

表5 LLM实验结果

Table 5 Experimental results on large language models

模块	标准数据				分布外数据			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
DeepSeek-1.5B	47.09	76.47	83.93	89.61	43.43	71.14	77.63	80.02
DeepSeek-7B	47.23	76.37	83.94	89.69	42.94	70.79	77.26	80.35
Qwen-Max	46.32	75.16	82.67	88.07	42.55	69.98	77.16	81.27
Doubao	**47.88**	**76.87**	**84.26**	**89.94**	**45.85**	**71.57**	**77.99**	**81.96**

注:同表3。

表6 Embedding模型实验结果

Table 6 Experimental results of embedding models

模块	标准数据				分布外数据			
	Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
BERT	44.60	74.61	82.76	88.81	37.86	64.23	73.09	78.94
GPT—2	46.70	76.11	83.77	89.52	43.63	69.77	76.95	80.53
Text2Vec	47.31	76.56	84.06	89.74	42.42	71.22	77.69	80.57
Qwen3	46.97	76.37	84.01	89.74	41.79	70.22	77.24	81.64
T5	**47.88**	**76.87**	**84.26**	**89.94**	**45.85**	**71.57**	**77.99**	**81.96**

注:同表3。

5 总结与展望

本文围绕移动APP使用预测任务,提出了一种基于LLM语义增强的APP使用预测方法AppSE。该方法充分挖掘LLM在理解APP功能和用户意图的能力,将其作为外部知识源引

入用户行为建模过程中,从而提升模型在分布外问题上的泛化能力。在真实数据集上验证了引入语言模型对提升预测准确性和鲁棒性的显著作用。

尽管所提方法在准确性与通用性方面已展现出良好表现,后续仍存在若干值得深入研究的问题:(1)在实际部署场景中,需进一步提升模型推理效率,降低模型在终端侧的计算资源开销;(2)针对LLM生成的描述内容与实际APP功能存在偏差的问题,未来可设计自动化校验与纠错机制,以提高语义增强模块的准确性与稳定性。

参考文献:

[1] Tila C. Number of Apps Available in Leading App Stores as of August 2024 and June 2025[EB/OL]. (2025-06-27) [2025-08-10]. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/.

[2] Moreira G S, Jo H, Jeong J. NAP: Natural App Processing for Predictive User Contexts in Mobile Smartphones [J]. *Applied Sci*, 2020, **10**(19): 6657. DOI: 10.3390/app10196657.

[3] Suleiman B, Lu K, Chan H W, *et al*. DeepPatterns: Predicting Mobile Apps Usage from Spatio-temporal and Contextual Features[M]//Service-Oriented Computing. Cham: Springer International Publishing, 2021: 811-818. DOI: 10.1007/978-3-030-91431-8_58.

[4] Katsarou K, Yu G, Beierle F. WhatsNextApp: LSTM-based Next-app Prediction with App Usage Sequences [J]. *IEEE Access*, 2022, **10**: 18233-18247. DOI: 10.1109/ACCESS.2022.3150874.

[5] Yang K, Zhao X, Zou J H, *et al*. ATPP: A Mobile App Prediction System Based on Deep Marked Temporal Point Processes[C]//2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS). Pafos: IEEE, 2021: 83-91. DOI: 10.1109/DCOSS52077.2021.00028.

[6] Chen X L, Wang Y, He J Y, *et al*. Cap: Context-aware App Usage Prediction with Heterogeneous Graph Embedding[J]. *Proc ACM Interact Mob Wearable Ubiquitous Technol*, 2019, **3**(1): 1-25. DOI: 10.1145/3314391.

[7] Yu Y, Xia T, Wang H D, *et al*. Semantic-aware Spatio-temporal App Usage Representation *via* Graph Convolutional Network[J]. *Proc ACM Interact Mob Wearable Ubiquitous Technol*, 2020, **4**(3): 1-24. DOI: 10.1145/3411817.

[8] Fang X, Yang H, Ding D, *et al*. Enhancing App Usage Prediction Accuracy with GCN-transformer Model and Meta-path Context[J]. *IEEE Access*, 2024, **12**: 53031-53044. DOI: 10.1109/ACCESS.2024.3372397.

[9] Huang Z H, Li T, Wang X, *et al*. Predicting Mobile App Usage with Context-aware Dynamic Hypergraphs[J]. *IEEE Trans Mob Comput*, 2025, **24**(6): 5511-5524. DOI: 10.1109/TMC.2025.3532992.

[10] Li T, Xia T, Wang H D, *et al*. Smartphone App Usage Analysis: Datasets, Methods, and Applications[J]. *IEEE Commun Surv Tutor*, 2022, **24**(2): 937-966. DOI: 10.1109/COMST.2022.3163176.

[11] Kamisaka D, Muramatsu S, Yokoyama H, *et al*. Operation Prediction for Context-aware User Interfaces of Mobile Phones[C]//2009 Ninth Annual International Symposium on Applications and the Internet. Bellevue: IEEE, 2009: 16-22. DOI: 10.1109/SAINT.2009.12.

[12] Leroux P, Roobroeck K, Dhoedt B, *et al*. Mobile Application Usage Prediction Through Context-based Learning[J]. *J Ambient Intell Smart Environ*, 2013, **5**(2): 213-235. DOI: 10.3233/ais-130199.

[13] Shin C, Hong J H, Dey A K. Understanding and Prediction of Mobile Application Usage for Smart Phones[C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing. Pittsburgh: ACM, 2012: 173-182. DOI: 10.1145/2370216.2370243.

[14] Huang K, Zhang C H, Ma X X, *et al*. Predicting Mobile Application Usage Using Contextual Information[C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing. Pittsburgh: ACM, 2012: 1059-1065. DOI: 10.1145/2370216.2370442.

[15] Liao Z X, Li S C, Peng W C, *et al*. On the Feature Discovery for App Usage Prediction in Smartphones[C]//2013 IEEE 13th International Conference on Data Mining. Dallas: IEEE, 2013: 1127-1132. DOI: 10.1109/ICDM.2013.130.

[16] Zhao S, Luo Z L, Jiang Z W, *et al*. AppUsage2Vec: Modeling Smartphone App Usage for Prediction[C]//2019 IEEE 35th International Conference on Data Engineering (ICDE). Macau: IEEE, 2019: 1322-1333. DOI: 10.1109/ICDE.2019.00120.

[17] Xu S J, Li W Z, Zhang X, *et al*. Predicting and Recommending the Next Smartphone Apps Based on Recurrent Neural Network[J]. *CCF Trans Pervasive Comput Interact*, 2020, **2**(4): 314-328. DOI: 10.1007/s42486-020-00045-z.

[18] Zhang Y Q, Kang M Y, Li X C, *et al*. Optimizing Smartphone App Usage Prediction: A Click-through Rate

Ranking Approach[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona: ACM, 2024: 6281−6290. DOI: 10.1145/3637528.3671567.

[19] Wang Y Z, Jiang R H, Liu H C, *et al*. Sequence-graph Fusion Neural Network for User Mobile App Behavior Prediction[M]//Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track. Cham: Springer Nature Switzerland, 2023: 105−121. DOI: 10.1007/978-3-031-43427-3_7.

[20] Khaokaew Y, Xue H, Salim F D. Maple[J]. *Proc ACM Interact Mob Wearable Ubiquitous Technol*, 2024, **8**(1): 1−25. DOI: 10.1145/3643514.

[21] Wang R, Gao Z Y, Zhang L Y, *et al*. Empowering Large Language Models to Edge Intelligence: A Survey of Edge Efficient LLMS and Techniques[J]. *Comput Sci Rev*, 2025, **57**: 100755. DOI: 10.1016/j.cosrev.2025.100755.

[22] Liu P F, Yuan W Z, Fu J L, *et al*. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing[J]. *ACM Comput Surv*, 2023, **55**(9): 1−35. DOI: 10.1145/3560815.

[23] Chen J Z, Fan T T, Liu X, *et al*. Seed1.**5**−Thinking: Advancing Superb Reasoning Models with Reinforcement Learning[PP/OL]. (2025−04−23) [2025−08−10]. https://arxiv.org/abs/2504.13914.

[24] Bai J Z, Bai S, Chu Y F, *et al*. Qwen Technical Report[PP/OL]. (2023−09−28) [2025−08−10]. DOI: 10.48550/arXiv.2309.16609.

[25] Raffel C, Shazeer N, Roberts A, *et al*. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer[J]. *J Mach Learn Res*, 2020, **21**(140): 1-67.

[26] Zhou G R, Zhu X Q, Song C R, *et al*. Deep Interest Network for Click-through Rate Prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1059−1068. DOI: 10.1145/3219819.3219823.

[27] Dao T, Gu A. Transformers Are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: PMLR, 2024: 10041−10071. DOI: 10.5555/3692070.3692469.

[28] Guo D Y, Yang D J, Zhang H W, *et al*. Deepseek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[PP/OL]. (2025−01−29)[2025−08−10]. https://arxiv.org/abs/2501.12948.

[29] 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. Minneapolis: ACL, 2019: 4171−4186. DOI: 10.18653/v1/N19-1423.

[30] Radford A, Wu J, Child R, *et al*. Language Models Are Unsupervised Multitask Learners[R]. San Francisco: OpenAI, 2019: 160025533. https://api. semanticscholar. org/CorpusID:160025533.

[31] XU M. Text2vec: A Tool for Text to Vector[CP/OL]. (2023−09−20) [2025−08−10]. https://github. com/shibing624/text2vec.

[32] Zhang Y Z, Li M X, Long D K, *et al*. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models[PP/OL]. (2025−06−11) [2025−08−10]. https://arxiv.org/abs/2506.05176.