

赞同 29

分享

蚂蚁2024：跨越长度界限：LLM通过长序列建模提升ctr效果



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

已关注

29 人赞同了该文章

Introduction

点击率⁺预测，尤其是LLMs，因其强大的语义理解和广泛的知识，在广告、搜索和推荐等多个场景中扮演关键角色。比如，M6-Rec利用LLM如M6进行交互重建，实现包括CTR在内的推荐任务。CTRL和FLIP通过整合LLM的语义，结合ID基模型，通过对比学习和掩码语言模型提高预测精度。KAR构建了三层框架，通过LLM的推理和事实知识，通过嵌入促进知识在CTR模型中的传递，展示了LLMs在该领域的显著提升潜力和广泛应用前景。

在CTR建模中，集成长时段的用户行为通常能显著提升模型效能，但将长文本信息融入LLM模型却导致训练和运行时间激增，不适宜大规模部署。为此，现有的LLM-CTR策略常被迫接受小规模模型和较短行为序列，以克服效率问题。此外，专为推荐设计的LLMs，如序列推荐，对长序列问题关注有限。

因此，如何优化LLMs以有效处理长用户行为，这是当前CTR预测领域亟待解决的挑战。我们设计并提出了Behavior Aggregated Hierarchical Encoding (BAHE)来解决LLMs在处理长用户行为序列时的效率问题。BAHE通过将用户行为的编码与交互过程分离，创新性地利用LLM的浅层网络提取基础行为特征，存储在离线数据库，避免重复计算，同时保留了详尽的交互细节。深层的LLM层则处理复杂的交互，生成全面用户表示，简化了学习过程，降低了计算成本⁺。

Proposed Method

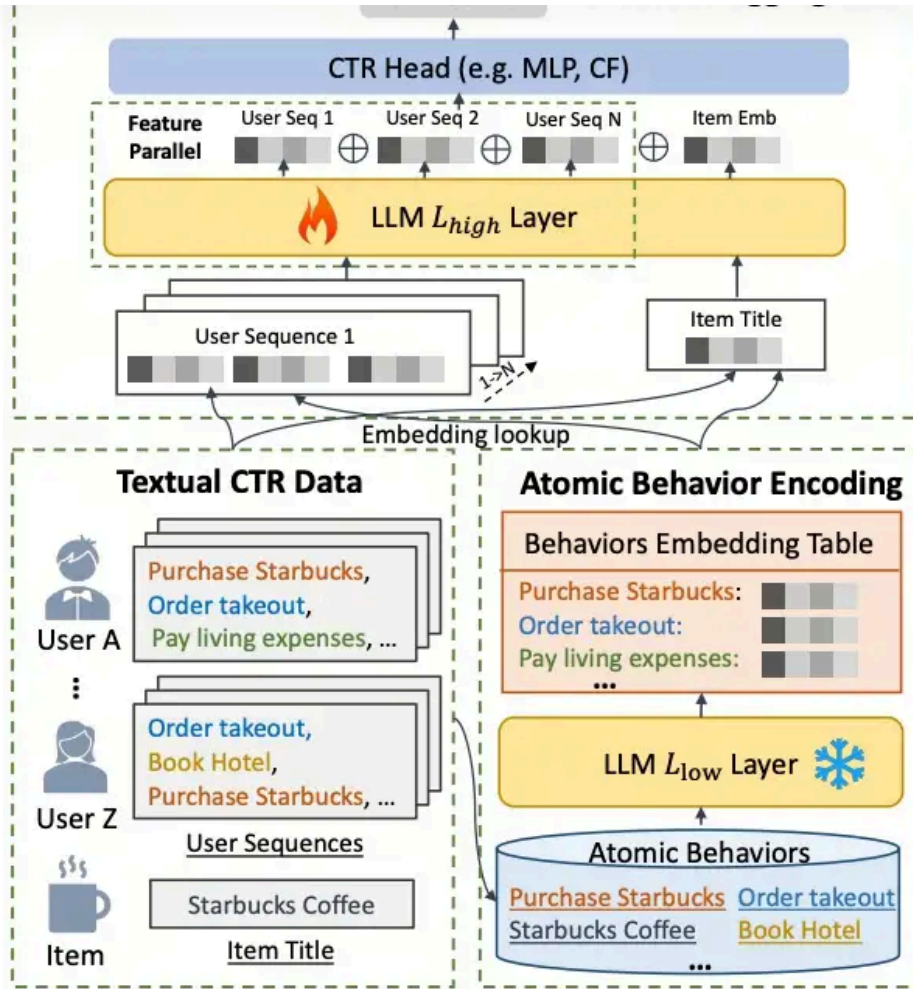


Figure 1: Architecture of the proposed BAHE method.

Problem Definition

在LLM的CTR预测任务中，我们使用用户文本行为序列来预测用户对特定商品的点击意愿。每个用户 u 有长度为 l_u 的多阶段行为序列 s_u ，由 N 个不同类型的 M 个行为 a_{umn} 组成，每个行为由 K 个令牌构成。这些行为序列覆盖了用户活动的多种领域，如点击、收藏或添加购物车。

商品集合 I 由文本特征 t_i 组成，这些特征通常包括商品标题，而商品的文本长度 l_i 远小于用户行为序列的长度 l_u 。我们定义行为集 H 为所有用户行为中的所有独一无二的原子行为的集合，它们是构建用户行为序列的基础元素。通过处理这样的长用户行为序列，我们的目标是通过LLM模型有效地提取用户行为模式，同时降低计算复杂性⁺。

$$H = \{a_{umn} | u \in U, 0 < m < M, 0 < n < N\}$$

我们使用LLM模型对用户 u 的文本行为序列 s_u 和商品文本特征 t_i 进行学习，目标是预测用户对商品 i 是否会点击，公式如下：

$$L = - \sum_{u \in U, i \in I} \log P(\text{click}_i | u, s_u, t_i; \theta)$$

其中 click_i 是二元标签

$$P(\text{click}_i | u, s_u, t_i; \theta)$$

是通过模型计算的点击概率 θ 是模型参数。该损失函数⁺通过最小化预测错误来提升模型在用户点击判断上的准确性。通过这种方式，我们期望LLM能从用户的行为序列中提取有效的模式，同时控制计算复杂性，实现CTR预测任务。

Behavior Aggregated Hierarchical Encoding

传统的LLM在处理用户行为序列时，尤其是当用户序列顺序不同（如 s_i 和 s_j ）导致信息重复时，效率问题突出，因为模型可能过多地关注了冗余信息。BAHE方法通过行为表示的分离，避免了这种重复，不论顺序如何。它通过聚合每个行为的原子特征，仅保留重要信息，减少了内存和计算负担，同时保持长序列处理的效率。

BAHE的目的是优化计算资源利用，通过LLM对行为序列进行深层次理解，生成全面用户表示，而非纠缠在低维行为编码上。在处理大量用户数据时，BAHE显著缩短了训练时间，使得模型能快速更新和调度，这既保证了预测精度，又体现了其在提高效率方面的优越性。因此，BAHE已成为解决LLM在CTR建模中效率瓶颈的有效解决方案。

BAHE通过行为聚合层次编码，它不依赖于行为顺序，而是对行为进行抽象和聚合，消除对特定顺序的依赖。这样，无论行为顺序如何变化，BAHE都能保持行为含义的稳定，降低了模型对新序列结构调整的必要性。因此，BAHE的这种松散耦合设计使得LLM能更灵活地处理动态行为序列，减少了更新的频率和成本，提高了模型的适应性和效率。

我们通过BAHE（Behavior Aggregated Hierarchical Encoding）方法解决了传统LLM在处理行为序列时的问题。BAHE通过行为级别的抽象和聚合，摒弃了对行为顺序的紧密依赖，实现了松散耦合。底层的令牌不再直接参与计算，而是从离线数据库按行为检索，这减少了冗余并提高了处理效率。

BAHE的工作流程是先利用LLM对行为的聚合特征进行高效学习，而非处理单个令牌，这避免了不必要的计算。在处理用户序列时，仅针对每个用户进行行为交互建模，而不是每次行为都解析整个序列。这种方法使得模型能更专注于理解行为模式，而非纠缠于低维细节，从而在处理长序列和大规模数据时，既能保持高精度，又能显著节省计算资源。

Atomic Behavior Encoding (ABE)

$$E_{a_i} = F_p(LLM_{Low}(a_i)) \quad E = \{a_i : E_{a_i} | a_i \in H\}$$

在BAHE中，每个行为 a_i ，比如由 K 个令牌组成的

$$a_i = [a_{i1}, a_{i2}, \dots, a_{iK}]$$

通过 LLM_{Low} 获得低维隐藏状态

$$LLM_{Low}(a_i) \in \mathbb{R}^d$$

池化函数 F_p （如维度为 d ）用于提取行为的关键特征，将这些 K 维特征压缩成单个行为级的 d 维嵌入向量，即

$$E[a_i] = F_p(LLM_{Low}(a_i))$$

行为嵌入表 E 是一个大小为 $|H| \times d$ 的矩阵，其中 $|H|$ 是行为数。BAHE通过这种方式将编码从行为的令牌细节转移到行为级别，消除了冗余，嵌入长度显著减少，从 K 减少到 1 。这种转换简化了计算，同时保持了对行为模式核心信息的把握。尤其在处理大量用户行为数据时，BAHE既能保证预测准确性，又能显著降低资源消耗，提升模型处理效率。

Behavior Aggregation (BA)

BAHE通过 E 获取原子行为的嵌入，如 $E_{a_i} = E[a_i]$ ，然后根据用户 n 的序列 s_{un} ，按行为顺序收集这些嵌入。这种方法不聚焦于单个令牌，而是通过行为级聚合，构建序列的综合表示，减少了维度。这种策略使得BAHE在处理长序列时，既能保持高精度，又能通过减少计算复杂度，有效应对大数据环境下的效率挑战。

$$E(s_{un}) = [E(a_{un1}) \oplus E(a_{un2}) \oplus \dots \oplus E(a_{unM})]$$

留了行为序列的语义信息⁺，又能保证预测的准确性，整体上提升了系统在面对大量行为数据⁺时的效率。

$$Q_{un} = F_d(F_p(LLM_{L_{high}}(E(s_{un}))))$$

对于序列 Q_{un} ，BAHE通过 $LLM_{L_{high}}$ 生成的高维全局表示，经过 F_d 的降维操作，将其维度从 d 降至 \hat{d} 。这个步骤旨在精简模型的输入，减少计算复杂性，同时尽量保留关键信息，确保模型的稳定性和预测精准度。即使维度降低，由于 F_d 的选择，重要学习信号的损失被控制在可接受范围内。

Feature Parallel (FP)

面对用户序列数量随增大的挑战，BAHE通过并行处理每个用户序列，利用 $LLM_{L_{high}}$ 进行单个计算，避免了全局复杂性的指数增长。它生成每个用户序列的最终表示，然后串联起来形成全局用户表示。这种策略保证了对用户行为模式的敏感度，同时通过分散计算，大大提升了处理大规模数据的效率。因此，BAHE既能维持模型性能，又能有效地管理和控制计算资源。

$$Q_u = [Q_{u1} \oplus Q_{u2} \oplus \dots \oplus Q_{uN}]$$

BAHE同样运用上述策略，为每个商品生成表示 Q_i 。它通过 $LLM_{L_{high}}$ 与商品行为的原子特征交互，通过类似的行为嵌入处理方式，构建出商品的全局表示。这种并行处理方式使得BAHE能够高效地处理大量商品，同时维持对商品特性的准确理解和预测，即使在用户数量增加的情况下也能有效管理和控制计算资源。

CTR Modeling

BAHE通过组合用户表示 Q_u 和商品表示 Q_i ，通过模型 F_θ 进行点击率预测，公式为：

$y = F_\theta(Q_u, Q_i)$ 这里的 F_θ 利用了用户和商品的联合特征，利用LLM的语义理解和行为序列信息，来预测用户对商品的点击概率。这种策略使得BAHE在保证计算效率的同时，能够提供更精细的CTR预测，体现了对LLM潜在能力的有效利用。

$$y = F_\theta(Q_u \oplus Q_i)$$

BAHE具有通用性，不依赖于特定的CTR模型，能适应各种嵌入模型。为证明这一点，BAHE以简单而实用的深度神经网络（DNN）为例进行操作。它通过优化与LLM相关的损失函数，来构建和学习基于LLM的点击率模型。训练完成后，BAHE能利用 Q_u 和 Q_i 的嵌入，结合LLM提供的语义信息，显著提升下游模型的预测精度，体现其对LLM潜在价值的充分利用。

Complexity Analysis

BAHE对传统模型的时间复杂度⁺分析表明，原始情况下为 $O(L(NMK)^2)$ ，其中 N 代表用户数量 M 代表行为数量 H 代表行为中的原子动作数 K 代表每个动作的令牌数，且 L 代表LLM层数。

BAHE通过优化策略将过程分为两步：首先进行原子行为的低维编码，这部分的复杂度是 $O(L_{low}(HK^2))$ ；其次，行为特征并行聚合，其复杂度为 $O(L_{high}(NM^2))$ ，其中 $L = L_{low} + L_{high}$ 。

BAHE带来的效率提升体现在：

1. 编码阶段节省：由于只关注 H 个关键行为，而非冗余处理所有 N^2M^2 个行为，因此节省了计算，具体为 $O(L_{low}((N^2M^2 - H)K^2))$ 。
2. 意图理解和聚合阶段：BAHE通过并行处理行为特征，同样减少了不必要的计算，但这个部分的具体节省没有直接给出，需要根据实际减少的特征数量来估算。

总体上，BAHE通过这两步优化，有效地降低了时间和空间复杂度⁺，特别是在处理大量用户和行为数据时，提高了模型运行效率。

Experiment

在处理一个包含5000万次CTR数据（每日日志总计），来自实际业务环境的大型数据集，BAHE的应用得以验证。数据集包含了6种类型的行为特征，如用户消费记录、搜索历史和小程序+活动，以及商品的标题信息，每个用户行为由50个动作组成，每个动作平均有5个令牌，总计1000万个原子行为。这些数据按照日志时间被划分为训练、验证和测试集+，目标是预测用户点击行为。

通过这种方式，我们能够客观地衡量BAHE在处理如此大规模数据集时的性能和效率提升。它能有效地处理用户行为序列，利用LLM的语义理解 and 行为序列信息，进行高精度的CTR预测。同时，BAHE的并行处理策略优化了计算复杂度，尤其是在处理海量数据时，确保了系统的高效运行。

Baseline Methods

为验证LLM在CTR建模中的优势，我们选取了LLM-CTR作为基线，这是一种常见的基于LLM的模型，它将用户行为序列合并成一个长序列进行处理。然而，BAHE因其模型无关性，能够广泛应用于各类LLM模型。

实验中，我们不仅评估了BAHE，也展示了使用LLM-CTR模型（具体为DNN）的下游CTR性能。通过比较两者，我们清晰地展示了BAHE在面对大规模用户行为数据时，无论是理解上下文信息还是预测CTR，都展现出了显著优势。这归功于BAHE能更有效地利用LLM的语义知识，通过并行处理和降维操作，提高计算效率，同时保持预测的准确性。因此，BAHE证实了LLM潜在的增益，特别是在处理复杂行为数据集时。

Evaluation Metrics

在评估CTR性能时，我们采用Area Under the Curve (AUC)这一标准，这是衡量预测准确性的通用准则。此外，我们关注的是计算资源的效率，通过GPU小时数(GPU-h)来量化训练所需的时间和内存消耗。同时，我们通过考察DNN（下游CTR模型的AUC_d），揭示了LLM对这类模型潜在的改进效果，这进一步强化了BAHE在利用LLM优势和有效管理计算资源方面的有效性。

Performance

Offline Performance

表格1比较了BAHE与基线方法。

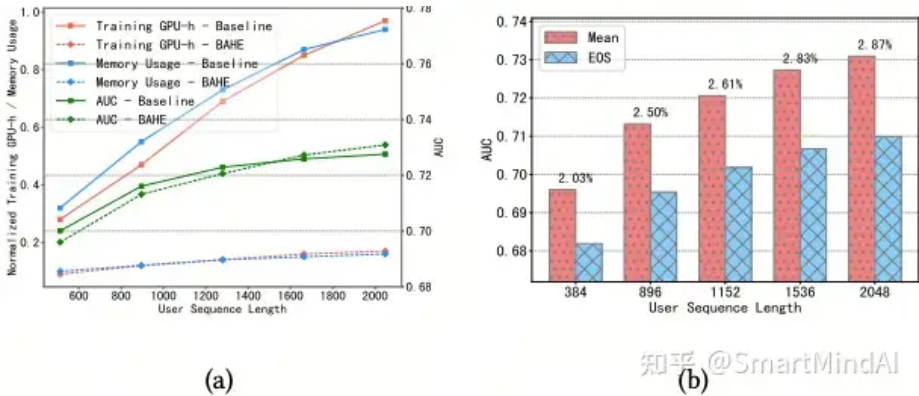
Table 1: Performance of BAHE at different text lengths. "AUC" represents the modeling performance of LLM, while "AUC_d" indicates the performance when transferring LLM representations to downstream models.

Model	Text Length=1024			Text Length=2048			
	AUC	GPU-h	Mem(G)	AUC	GPU-h	Mem(G)	AUC _d
DNN	-	-	-	-	-	-	0.7219
LLM-CTR	0.7161	448	43.8	0.7276	928	75.4	0.7323
+ FP	0.7143	420	36.8	0.7326	864	67.8	0.7369
+ ABE	0.7150	256	23.0	0.7332	416	38.0	0.7372
+ BA(BAHE)	0.7132	116	9.8	0.7309	164	12.6	0.7352

Online Deployment

在实际的电子商务广告CTR预测应用中，BAHE成功部署并进行了两周的A/B测试。相较于基线方法，BAHE展现出显著优势。BAHE能在一天内处理5000万条数据的训练，远超出基线模型+每周的处理上限，这大大提升了处理能力。这使得基于LLM的模型能更充分地发挥其潜力，进而导致在线点击率增长了9.65%，广告CPM（每千次展示费用）上升了2.41%。这些实证数据明确证实了BAHE在提升效率和模型效能方面的显著贡献。

Comparison at Different Sequence Lengths



通过图示研究，我们探究了BAHE与基线方法在不同序列长度下的性能差异。结果显示，随着用户行为序列的增长，AUC评分普遍提升，这证实了更长的文本确实能增强LLM的CTR预测能力。此外，BAHE在处理长序列时表现出更强的优势，这强调了其在处理复杂行为数据时的优越性及对LLM潜在增益的充分利用。

Comparison of Different Pooling Methods

实验数据显示，相较于使用LLM的最后一个隐状态（EOS），平均池化⁺的表现更好。这暗示在生成型LLM中，全局的表示方式（即平均池化）更能有效地捕捉和利用信息，而非仅仅依赖于单个隐状态。这可能是由于全局表示能综合多个时间步的上下文，从而在理解和预测上提供了更全面和准确的视角。


Conclusion

针对LLM在处理海量用户文本序列时的计算效率问题，本文创新提出BAHE策略。BAHE通过聚焦于原子行为的编码，以提升行为表示的复用性，进而增进跨用户行为表示的普遍性。它巧妙地利用LLM的分层编码结构，将行为理解和交互的建模从计算中解耦，这显著提升了效率。实证研究⁺表明，BAHE不仅能大幅提升处理速度（约5倍），还在CTR预测上表现出显著提升，为LLM在实际场景中的广泛应用提供了理论支持和实践建议。






原文《Breaking the Length Barrier: LLM-Enhanced CTR Prediction in Long Textual User Behaviors》

发布于 2024-05-11 11:31 · IP 属地北京


阿里巴巴集团 序列推荐 工业级推荐系统



理性发言，友善互动



发布



还没有评论，发表第一个评论吧