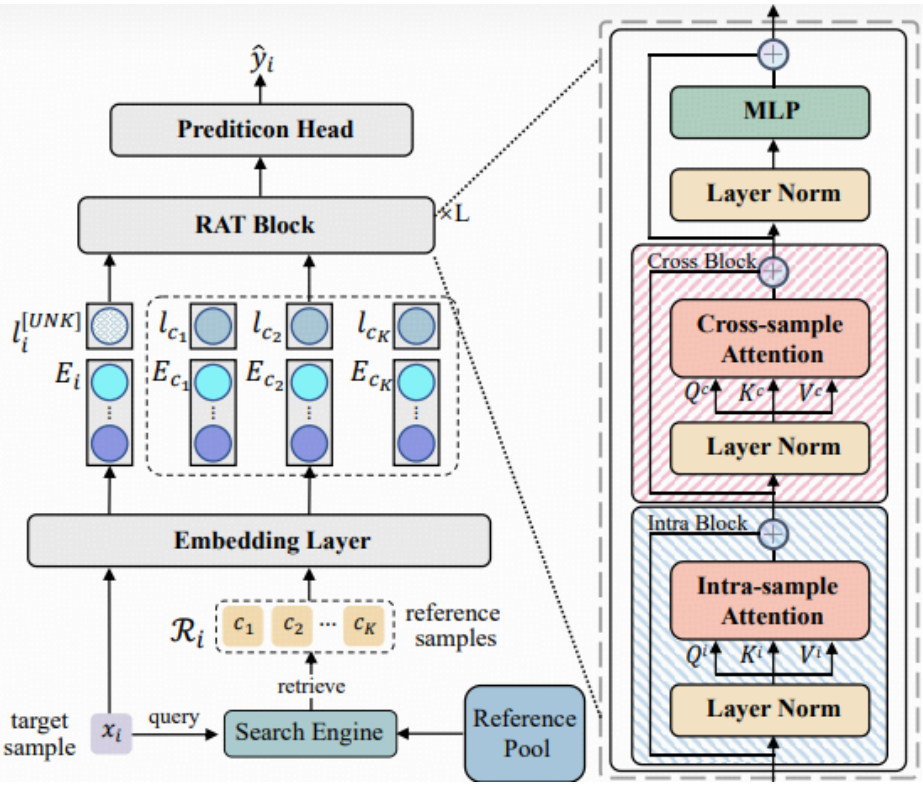
知乎



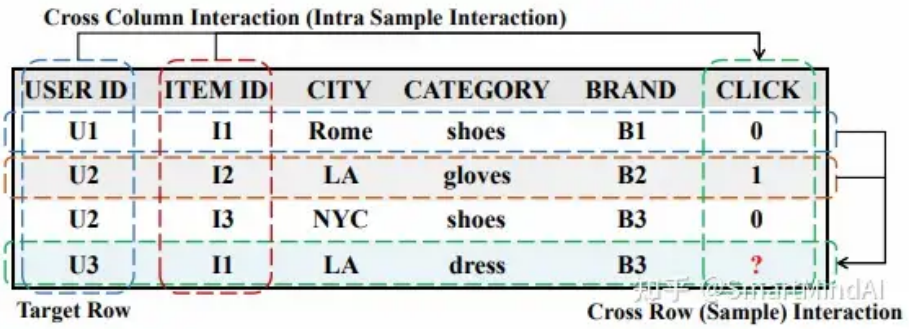
2024华为：RAT-基于Transformer的检索增强型模型提升CTR预估的AUC

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

11 人赞同了该文章

Introduction

点击率（CTR）预测主要目标是决定用户是否会点击一个内容的二分类任务，在广告和推荐系统等领域有重要应用。



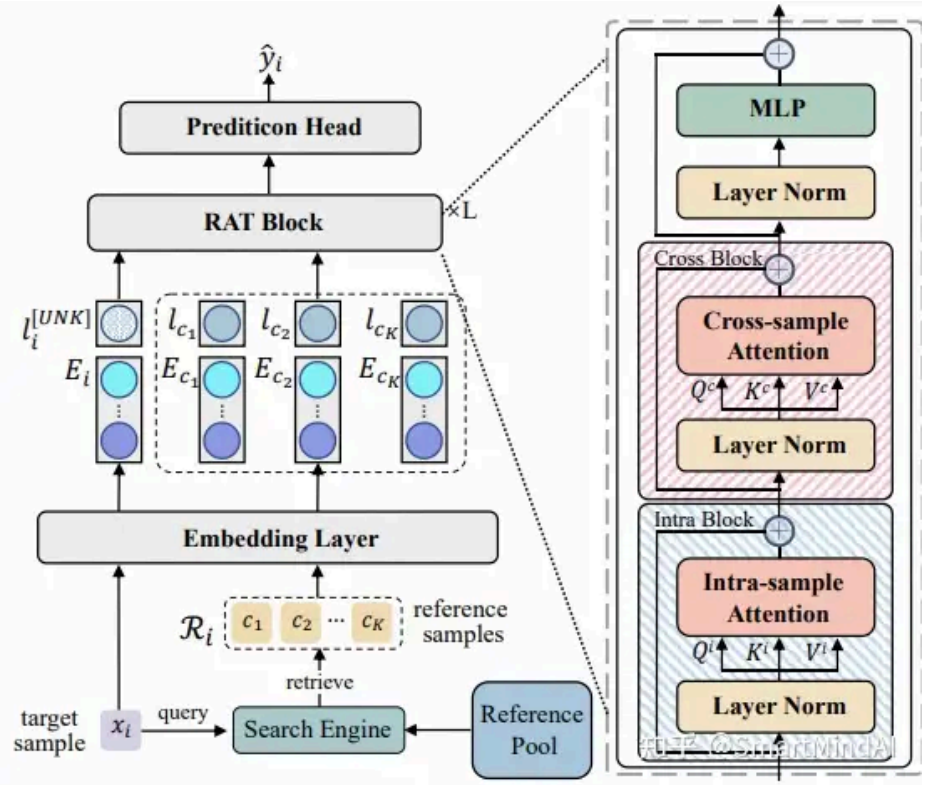
传统方法侧重于样本内部的特征交互，却忽略了作为增强预测的关键上下文的信息。由于特征及其交互的稀疏性，CTR模型需具备强大的捕获和记忆所有交互模式的能力，而这在性能稳健性和模型扩展性方面提出了挑战。

为应对这些挑战，检索增强（RA）学习在近来自然语言处理和计算机视觉领域展现出显著效用。其核心思路是通过检索相似样本来增强模型的预测性能。通过利用外部示例，模型的预测准确性和泛化能力得到了提升。这种方法旨在通过整合周边信息，提升CTR预测的精度与效率。

本文提出了检索增强的Transformer模型，该模型引入了先进的检索技术和优化的交互建模方法，利用Transformer架构增强模型的理解能力，尤其是在复杂交互模式下的处理上，从而提供了更准确和有效的CTR预测结果。

知乎

我们创新性地设计了将内部及跨样本交互融合的检索增强查询Transformer（REtrieval-augmented Transformer, [RAT]），专为点击率预测（CTR）任务服务。其工作原理及模块布局在图进行了直观展示。



Retrieve Similar Samples as Context

对于目标样本的F-字段记录$x_i = [x_i^1; \ldots; x_i^F]$，我们从预留的样本池$\mathcal{P}$中搜索参考上下文下的类似样本。我们使用BM25进行检索，因为它具有无训练性质，并且与先前的工作相吻合。 具体来说，目标样例x_i和候选样例$(x_c, y_c) \in \mathcal{P}$的候选关键字$x_c$的关联得分定义为：

$$s(x_t, x_c) = \sum_{f=1}^{F} \log \frac{N_{\mathcal{P}} - N_{\mathcal{P}}(x_t^f) + 0.5}{N_{\mathcal{P}}(x_t^f) + 0.5} \cdot \mathbb{I}_{\{x_t^f = x_c^f\}},$$

其中 $\mathbb{I}_{\{x_c^f = x_i^f\}}$ 为指示函数[+]，指示x_c^f和x_i^f是否相等。$N_{\mathcal{P}}$为样本池$\mathcal{P}$的总样本数量，而$N_{\mathcal{P}}(x_i^f)$指$\mathcal{P}$中拥有特定特征x_i^f的样本数量。

与rim在CPU上使用Elasticsearch进行检索不同，我们采用了基于GPU的优化实现，以实现更为高效的检索速度。 最终，我们从$\mathcal{P}$中选出K个最高得分的样本。

$$\mathcal{R}_i = \{(x_{c_1}, y_{c_1}), (x_{c_2}, y_{c_2}), \cdots, (x_{c_K}, y_{c_K})\}$$

在处理可能出现的时间戳信息时，我们遵循特定的策略以防止信息泄漏。具体来说，若样本包含时间戳数据，我们会先按照时间序列[+]进行排序。接着，仅允许根据更早时间的时间戳进行查询检索。这样一来，验证和测试环节的安全性得到了保障，因为在这两个阶段，我们始终使用完整训练数据集作为检索参照。这种方式确保符合限制条件且是安全的------在我们的实验设计中，验证集与测试集[+]分别作为整体数据集的最新和次新部分，从而避免了对未来的数据信息的访问。

Construct Retrieval-augmented Input

为了将离散属性转换为维度为D的嵌入向量[+]，我们设计了一个嵌入层。尤其地，我们为检索到样本的标签建立了一个专门的嵌入表，并记作集合

$$\mathcal{E} = \{E^1, E^2, \ldots, E^F, L\}$$

[UNK] 表示）。 对于被检索到的样本(x_{c_k}, y_{c_k})，我们从嵌入表中检索出其特征嵌入和标签嵌入，最终构建得到

$$E_{c_k} = [l_{c_k}; e^1_{c_k}; e^2_{c_k}; \cdots; e^F_{c_k}] \in \mathbb{R}^{(F+1)\times D}$$

同样的方法，目标记录x_i同样给出其包含未知点击标签（ [UNK] ）的嵌入表示

$$E_i = [l_i^{[\mathrm{UNK}]}; e^1_i; e^2_i; \cdots; e^F_i] \in \mathbb{R}^{(F+1)\times D}$$

最后，为了将目标记录x_i与检索到的各个样本进行对比并整合信息，我们通过将E_i与其它所有检索到的样本

$$E_{c_1}, E_{c_2}, \cdots, E_{c_K}$$

按列堆叠的方式构建其检索增强输入。

$$\tilde{E}_i = [E_i; E_{c_1}; E_{c_2}; \cdots; E_{c_K}] \in \mathbb{R}^{(K+1)\times(F+1)\times D}$$

Intra- and Cross-sample Feature Interaction

为了整合样本间以及样本内部的特征交互以提升点击率预测的准确性，直观的做法是将检索得到的所有样本特征⁺进行解堆叠后与目标记录结合，进一步拟用联合注意力机制⁺来捕捉整体特征间的交互模式。然而这种方法存在效率不足的问题，需计算的计算量达到

$$\mathcal{O}((K+1)^2 \cdot (F+1)^2)$$

且实测效能不佳，可能源自于对噪声特征交互的过高敏感度。

为改善这一状况，我们对联合注意力机制进行了构分解，提出了模拿化构建块（[RAT]）。每个块由嵌套的内部层、外部层与多层感知器⁺（MLP）串联组成。第ℓ层的块前向传播的过程可数学化为:

$$\begin{aligned} H_i^\ell &= \mathrm{ISA}(\mathrm{LN}(X_i^\ell)) + X_i^\ell, \\ H_i'^\ell &= \mathrm{CSA}(\mathrm{LN}(H_i^\ell)) + H_i^\ell, \\ X_i^{\ell+1} &= \mathrm{MLP}(\mathrm{LN}(H_i'^\ell)) + H_i'^\ell, \end{aligned}$$

过程如下: - X_i^ℓ表示第ℓ个块的输入端数据。 - $X_i^0 = \tilde{E}_i$为初始输入状态。 - H_i^ℓ和$H_i'^\ell$分别代表块内的隐藏状态，用于捕捉内部样本的交互关系和跨样本间的关联。 - $\mathrm{LN}(\cdot)$是层归一化操作，用于稳定学习过程。 - $\mathrm{ISA}(\cdot)$与$\mathrm{CSA}(\cdot)$分别对应内部样本和跨样本注意力模块，执行多头自我注意力操作，沿着字段轴和样本轴辨别特征间的相关性。

与传统的联合注意力机制相比，我们提出的设计将计算复杂度显著降低至$\mathcal{O}((K+1)^2 + (F+1)^2)$ 以保障较高的计算效率。此外，我们还将通过实验深入探讨不同设计的细节，并展示级联注意力在中的优点。

Experiments

Datasets

我们使用三个广泛认知的基准数据集⁺对[RAT]进行了性能评估，包括ML-tag、KKBox和Tmall。

Dataset	#Samples	#Fields	Missing_ratio	Positive_ratio
ML-tag	2,006,859	3	0%	33.33%
KKBox	7,377,418	19	6.08%	50.35%
Tmall	54,925,331	9	0.36%	50%

知乎

我们采用了两个常见的评估指标进行模型性能评估:**AUC**（ROC曲线下的面积）和**Logloss**（交叉熵损失）。这两大指标分别用于衡量模型的区分能力和预测精确度，是业界广为使用的评估标准。其中，AUC值越接近1表示模型分类效果越好，而较小的Logloss值则表示模型的预测质量越高。

Baselines

我们选择了两种类型的基线模型进行对比: (i) 传统的模型:包括DeepFM, xDeepFM, DCNv2, AOANet，这些模型在数据驱动的推荐系统中表现良好，但在检索增强方面较为传统，没有利用检索结果来提升模型性能。 (ii) 检索增强模型：RIM, PET。

为了保证公平比较，我们使用BM25进行设置一致性的对齐。

Comparison with State-of-the-arts

Overall Performance

我们在表中报告了模型的性能。

Model	ML-tag		KKbox		Tmall		$\Delta AUC\uparrow$
	AUC↑	Logloss↓	AUC↑	Logloss↓	AUC↑	Logloss↓	
DeepFM	0.9685	0.2130	0.8429	0.4895	0.9401	0.3341	-0.24%
DCNv2	0.9691	0.2147	0.8303	0.5073	0.9391	0.3377	-0.75%
AOANet	0.9694	0.2105	0.8313	0.5046	0.9391	0.3369	-0.70%
xDeepFM	0.9697	0.2409	0.8475	0.4846	0.9406	0.3513	0.00%
PET	0.9692	0.2602	0.8316	0.5044	0.9520	0.3175	-0.24%
RIM	0.9711	0.1832	0.8465	0.4907	0.9534	0.3131	0.46%
RAT	**0.9809**	**0.1421**	**0.8500**	**0.4812**	**0.9589**	**0.3091**	**1.13%**

请注意，对于点击率预测，AUC提高到小数点后三位被认为是可接受的，因为即使是如此微小的改进，如果统计上显著，也可以带来显著的收入增加。根据表中的结果，检索增强（RA）模型通常优于传统模型，这表明结合跨样本信息对于增强点击率预测是有效的。尽管PET的总体性能不优于xDeepFM，主要原因是其不足以捕捉内部样本特征交互作用，但它在与其他传统模型相比时表现出优势。此外，[RAT]超越了最先进的RA模型RIM，这表明Transformer在建模细致入微的内部和跨样本交互作用方面的强大能力。

Performance on Long-tail Data

受到检索机制在长尾识别领域的成功案例启发，我们提出探讨[RAT]是否能通过增强点击率预测能力，来处理长尾数据集的挑战。以ML-tag为例，我们特别观察了用户样本中低频交互部分（即排名10%和20%的用户子集）的表现。通过分析，我们可以见到检索增强模型（RA模型），相较于传统的预测模型，表现出显著提高的效能。这一结果明确显示了引入检索上下文对推断稀疏特征和它们之间的交互性具有积极影响。

Type	Model	Tail 10% Users		Tail 20% Users	
		AUC↑	Logloss↓	AUC↑	Logloss↓
Non-RA	DeepFM	0.9487	0.2714	0.9616	0.2319
	DCNv2	0.9445	0.2748	0.9578	0.2388
	AOANet	0.9479	0.2764	0.9617	0.2344
	xDeepFM	0.9482	0.2866	0.9611	0.2436
RA	PET	0.9496	0.2746	0.9631	0.2296
	RIM	0.9543	0.2574	0.9654	0.2200
	RAT	**0.9583**	**0.2250**	**0.9727**	**0.1799**

知乎

的能力和潜力，这尤其对于解决特征稀疏性和冷启动问题具有高度的实际意义。

原文《RAT: Retrieval-Augmented Transformer for Click-Through Rate Prediction》

发布于 2024-06-27 10:47 · IP 属地北京

华为 Transformer 搜索引擎