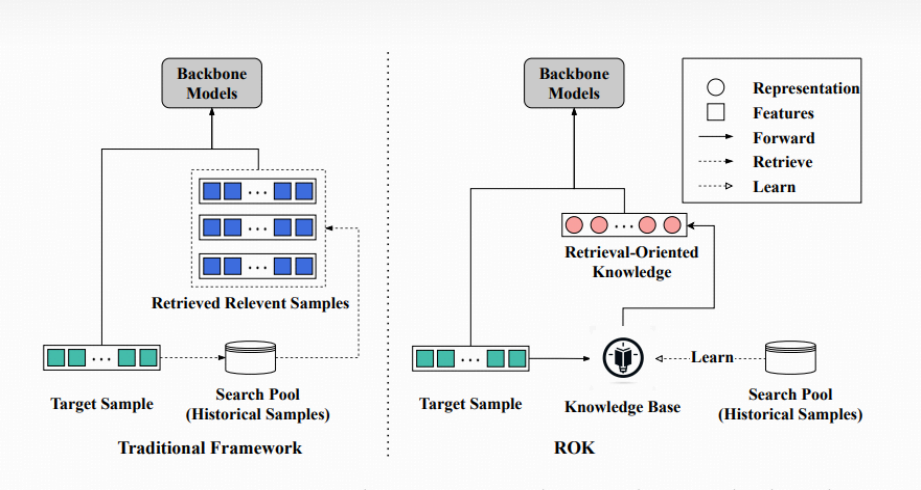
知乎 ··· 赞同 16 分享



2024年华为：采用样本检索技术优化CTR预测效率与精度

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

16 人赞同了该文章

Introduction

点击率预测在推荐系统和网页搜索等场景至关重要，主要通过预测用户点击概率来实现。模型主要分为两类： 1.基于特征交互的方法，依靠特征组合、卷积及注意力等多级特征交互，通过操作符捕捉高阶关联。 2.利用不同架构的用户行为建模方法：RNN、CNN、注意力和内存银行等架构，从用户行为序列中提炼对点击率的指导信息。

这两种方法在各自领域都有优异表现，通过不同途径解析用户行为和特征交互，以提升CTR预测精度。为改善点击率预测，UBR4CTR和SIM利用用户浏览历史去噪，通过哈希$^+$和并行检索优化效率。最近，研究引入了更广的'样本级别检索'，如RIM，不仅局限于相似物品，而是采用k最近邻方法，以样本为中心进行聚合。PET通过构建超图并进行消息传递来优化样本表示，然而，这种方法在处理大规模搜索池（如百万或数十亿）时效率低下，影响了其实时性。DERT通过向量检索优化了检索，但对RIM编码器的依赖、$\log N$的时间复杂度以及样本编码成本限制了其在大规模应用中的适用性。

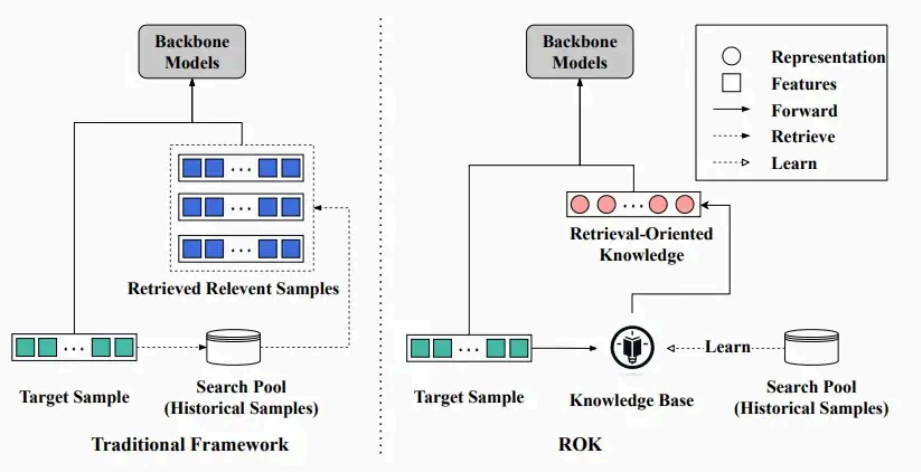


Figure 1: Comparison between traditional sample-level retrieval framework and ROK.

知乎 @SmartMindAI

本文提出**ROK**框架，解决样本级别检索在推理效率问题，通过两阶段操作：知识导向构建和知识利用。

1. 预训练RIM等检索模型，通过**ROK**的检索仿真实验模块，通过神经网络构建解构-重构知识库，模仿检索聚合表示，避免了直接检索的计算成本，利用神经网络的高效推理。
2. 引入对比正则化确保学习稳定，防止模型失效。

知乎

CTR Prediction

在CTR预测任务中，每个样本$s_t = (x_t, y_t)$包含离散特征c_i^t和目标变量y_t。数据集T由N个样本组成，表示为$\{s_t\}_{t=1}^N$。模型G通过参数θ估计单个样本的点击概率，即

$$\hat{y}_t = G(x_t; \theta)$$

基于检索增强的方法不仅区分于传统CTP的训练集、验证集和测试集，还引入了搜索池T_{pool}。这个池子包含与训练集T_{train}可能重叠，但专为优化目标样本x_t的预测目的设计的相关样本。搜索池的构建标准是找到那些与x_t相关且能提供有价值信息的s_p，即：

$$T_{pool} = \{s_p \in T : s_p \text{与} x_t \text{关联,有助于提高CTR预测}\}$$

通过从搜索池中检索这些样本，模型能够利用额外信息来提升预测准确度。然而，RIM的高依赖性、计算复杂性和样本编码成本可能限制了这种方法在大规模应用中的广泛应用。

$$\hat{y}_t = G\left(x_t, R\left(x_t\right); \theta\right)$$

在CTR预测中，损失函数$L(\theta)$通过最小化预测与真实标签的交叉熵来优化模型，公式如下：

$$L(\theta) = -\sum_{t=1}^{N} (y_t \log \hat{y}_t + (1 - y_t) \log(1 - \hat{y}_t))$$

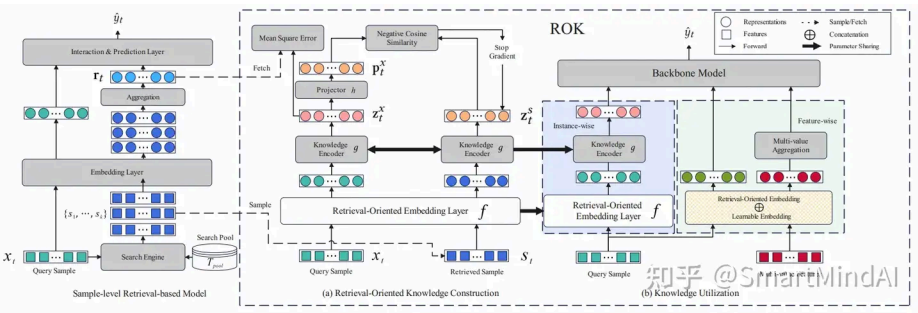
其中N代表样本数量。基于检索的方法中，损失函数旨在利用搜索池T_{pool}中的样本来提升预测，但具体的实现可能依据检索策略的不同。例如，如果基于样本级别（Sample-level），损失函数可能调整为$R(s_p)$，而不仅仅是$\hat{y}_t$。尽管如此，核心目标都是利用检索样本来增强模型对目标样本x_t的预测能力。然而，RIM的局限性（如对RIM编码器的依赖和高计算复杂性）可能制约了这种方法在大规模应用中的广泛采用。

$$\mathcal{L}_t = y_t \log \hat{y}_t + (1 - y_t) \log(1 - \hat{y}_t)$$

Sample-level Retrieval-based Methods

- 目标样本x_t作为查询，从搜索池T_{pool}获取前K个邻近样本$\{s_1, \ldots, s_K\}$。
- 进行特征和标签聚合，如RIM使用注意力聚类，PET构建超图并通过消息传递进行融合，形成知识表示$\mathbf{r}_t$。
- 模型接收到增强的样本表示$\mathbf{r}_t$后，与目标样本的原始表示一起输入预测模块。
- 通过优化参数θ，采用二元交叉熵（BCE）损失函数，使模型预测的点击概率$\hat{y}_t$接近真实标签y_t，提升预测准确性。
- 不论是item-level还是sample-level检索，核心目标都是利用检索样本帮助模型更好地理解目标样本，但具体损失函数会根据策略调整。

Overview of ROK



我们提出ROK（Retrieval-Oriented Knowledge）框架。它包含两个阶段：检索导向知识构造和知识利用。

1. 检索导向知识构造：

$\mathbf{z}_t^x$，仿效预训练模型的$\mathbf{r}_t$。

- 采用均方误差⁺作为知识模仿损失，同时加入对比性正则化来稳定学习。

2. 知识利用：

- 将检索增强的插件式表示融入基础的CTR模型，既可以实例级，也可以特征级，既保留了检索能力，又减少了在线运行时的计算成本。 接下来，我们将详细探讨知识库的设计和两阶段的训练策略。

Structure Design of Knowledge Base

Retrieval-Oriented Knowledge Construction

检索仿真实验和对比正则化。首先，通过实验模拟预训练模型的聚合检索知识，这涉及到对RIM（一个）进行样本级的训练，使其适应各种场景。其次，我们设计了一个特殊的嵌入层和知识编码器，它们集成在知识库中，能直接将输入样本转化为聚合表示$\mathbf{z}_t^x$，这种方式模仿了预训练模型⁺的处理过程。 为了确保学习的稳定性并防止模式塌陷，我们引入了对比性正则化，这有助于优化模型对不同检索结果的泛化能力。这两个模块共同工作的目标是提升模型在保持高效推理的同时，充分利用检索增强的优势。

Retrieval Imitation

在重构样本x_t的表示$\mathbf{z}_t^x$后，我们用均方误差(MSE)作为损失来训练知识库以逼近聚合表示$\mathbf{r}_t$，公式如下

$$\mathcal{L}_{imit} = \mathrm{MSE}\left(\mathbf{z}_t^x, \mathbf{r}_t\right)$$

这种知识库的构建和训练方法确实可以看作是知识蒸馏⁺，它从预训练模型的聚合知识中汲取并内化，形成具有检索能力的学习资源。通过模仿预训练模型的聚合表示，知识库能够更好地理解和处理新的样本，从而提升CTR预测的准确性。对比性正则化的引入则保证了学习过程的稳定性和泛化能力，避免了过拟合，使得知识库能适应不同情况下的样本。

Contrastive Regularization

在这个过程中，这种新颖的数据增强方法，增强了对比学习的局部视角。采用自由负样本策略，而非仅依赖全局特征。如图(a)所示，它从搜索池选取最相关的临近样本s_t，然后对目标样本x_t和s_t进行输入，得到重构后的表示$\mathbf{z}_t^x$和$\mathbf{z}_t^s$。

为克服过拟合，引入了投影器h，这个投影器将原始表示转换为项目表示$\mathbf{p}_t^x$和$\mathbf{p}_t^s$。这样的设计确保了在无负样本情况下也能保持学习的稳定性，通过对比性正则化，模型能更深入地提取样本的局部特征，从而提升知识库的泛化能力和对目标样本的预测准确性。

$$\begin{aligned} v * -0.5 \mathbf{p}_t^x &\triangleq h\left(\mathbf{z}_t^x\right) \triangleq h\left(g\left(f\left(x_t\right)\right)\right), \\ \mathbf{p}_t^s &\triangleq h\left(\mathbf{z}_t^s\right) \triangleq h\left(g\left(f\left(s_t\right)\right)\right) . v * -0.5 \end{aligned}$$

投影到的项目表示$\mathbf{p}_t^x$与重构样本$\mathbf{z}_t^s$之间的余弦相似性⁺定义为：

$$sim(\mathbf{p}_t^x, \mathbf{z}_t^s) = \frac{\mathbf{p}_t^x \cdot \mathbf{z}_t^s}{\|\mathbf{p}_t^x\|_2 \|\mathbf{z}_t^s\|_2}$$

这里，点积⁺·衡量两者方向上的交集，欧氏范数$\|\cdot\|_2$是向量长度，相似性值介于-1（完全不相关）至1（完全相同）。该指标用于评估两向量在空间中的相对位置关系，用于正则化和对比学习中的决策，帮助保持学习过程的稳定性和避免模式塌陷。

$$\mathcal{D}\left(\mathbf{p}_t^x, \mathbf{z}_t^s\right) = \frac{\mathbf{p}_t^x}{\|\mathbf{p}_t^x\|_2} \cdot \frac{\mathbf{z}_t^s}{\|\mathbf{z}_t^s\|_2}, v * -0.5$$

其中$\|\cdot\|_2$是指l_2范数，用来量化向量的欧氏距离⁺。在损失函数计算中，为了避免重构向量$\mathbf{z}_t^s$直接参与优化导致的空间偏移，我们不会考虑它的梯度。为克服这个问题，我们使用对称损失，灵感来源于Jensen-Shannon散度，其形式为：

其中τ是温度参数，调节相似性分数的平滑程度。这个损失策略旨在最大化正样本（同目标的$\mathbf{z}_t^x$和$\mathbf{z}_t^x$）间的相似度，同时最小化负样本（不同目标的$\mathbf{z}_t^x$和$\mathbf{z}_t^s$）之间的相似性，以此促进模型学习更精确且泛化能力强的表示。

$$\mathcal{L}_{contra} = -\left(\frac{1}{2}\mathcal{D}\left(\mathbf{p}_t^x, \mathbf{z}_t^s\right) + \frac{1}{2}\mathcal{D}\left(\mathbf{p}_t^s, \mathbf{z}_t^x\right)\right). v * -0.5$$

在基于检索的知识构建过程中，（均方误差）损失与对比性正则化结合起来，形成总损失：

$$L_{\text{total}} = L_{\text{MSE}} + L_{contrastive}$$

MSE损失确保知识库能接近聚合表示$\mathbf{r}_t$，而对比性正则化（由JS散度驱动的对称损失）通过强化正样本间相似度和减少负样本间相似性，促进了表示的精确性和泛化能力。这样，能避免因目标直接优化导致的空间偏差，从而学习出既精细又泛化的知识表示。

$$\mathcal{L} = (1-\alpha)\cdot\mathcal{L}_{imit} + \alpha\cdot\mathcal{L}_{contra}, v * -0.5$$

其中，参数α控制了MSE损失和对比正则化的重要性。当α大于0时，MSE起主导作用，强调模仿聚合表示；而对比正则化则通过JS散度⁺调整负样本的差异，以提升泛化能力。构建完成后，得到的知识库便进入了后续的利用阶段。

Knowledge Utilization

在知识利用阶段，我们利用经过增强的检索表示，将其注入到不同的基础CTR模型中。为了提升预测效果，我们提出两种策略： 1. 联合学习：这种方法结合了知识库中的表示和预训练模型的部分，通过轻量级微调，增强了对用户行为的深度理解和预测。 2. 在线更新：另一种方法则是动态地利用知识库，允许在实际交互中根据新数据不断调整和优化模型，以保持对最新趋势的学习和适应。

- **基于特征的检索增强**通过将检索导向的嵌入层f与基线模型的原始特征向量融合，生成样本的检索增强特征。这种方法将检索层的附加信息加入到模型中，形成一个融合的嵌入层，以提升CTR预测。为保证参数对比的公正性，尽管嵌入层参数丰富，但我们将两者特征维度都调整为$d/2$，确保比较的准确性。这样做的目的是整合两者优势，增强模型在特定任务上的表现.
- **实例级检索增强**通过为每个样本创建个性化的检索增强实例，如RIM的聚合表示和DIEN的兴趣状态，利用知识库生成$\mathbf{z}_t^x$。这种方法结合了基线模型⁺（如RIM和DIEN）的专属特征，与样本本身的特征嵌入相结合。这种聚合表示由独特的模型视角提供，然后与这些特征输入到交互和多层感知器⁺（MLP）中，以提升模型对个体样本的理解和识别能力。这样做是为了确保模型的性能在保持公平性的同时，能针对每个实例进行精细化处理。

其中ϕ的角色是提取基线模型特有的单例表示。它通过解析模型的输出，获取每个样本独有的特定信息，这些信息源于模型内部的特征计算，如使用CNN或RNN等机制。将这些独特表示与样本的特征嵌入合并，生成增强的实例级聚合表示$\mathbf{z}_t^x$。这样做旨在使模型能深入理解每个样本的个体特性。这种策略提升了CTR预测的精确性，因为模型能够更精准地识别和处理每个实例。

Experimental Settings

Datasets.

实验在三个主流平台的数据集------天猫(Tmall)，淘宝(Taobao)和支付宝(Alipay)上进行了。 按照RIM的方法，数据被整理成这样的结构：最古老的数据作为检索池，最新数据作为验证集，而样本时间点位于两者之间的分配给了训练集。检索型方法，如DIEN，会从这个搜索池中查找邻近样本。针对这类问题，还会从搜索池中提取序列特征，比如用户的行为数据。

知乎

Dataset	Users #	Items #	Samples #	Fields #	Features #
Tmall	424,170	1,090,390	54,925,331	9	1,529,676
Taobao	987,994	4,162,024	100,150,807	4	5,159,462
Alipay	498,308	2,200,291	35,179,371	6	3,327,205

知乎 @SmartMindAI

Evaluation Metrics.

我们使用标准的评估工具，如AUC（曲线下面积）和log-loss（点-wise概率预测），来度量性能。这两种指标分别体现模型的排名精准度和预测准确性。对于每个指标，我们将顶级方法进行对比，通过星号()来标识显著差异。

$$\text{Rel.\ Impr.} = \left(\frac{\text{AUC(measured model)}-0.5}{\text{AUC(base model)}-0.5} - 1\right) \times 100\%, v * -0.5$$

在实验中，基础模型*在每个数据集上被视为基准，而研究关注的是ROK（检索导向知识增强）方法如何提升这些基线模型的表现。通过Rel. Impr.这个指标，我们量化了当应用ROK后，模型相对于基线模型M_1的实际提升M_2就是包含了ROK技术的模型。这样，我们能够直接评估ROK对模型性能的增量贡献，明确显示出其在特定数据集上的实际效果。

Overall Performance Comparison: RQ1

Category	Model	Tmall			Taobao			Alipay		
		AUC	LL	Rel. Impr	AUC	LL	Rel. Impr	AUC	LL	Rel. Impr
Traditional Models	GBDT	0.8319	0.5103	-13.55%	0.6134	0.6797	-54.75%	0.6747	0.9062	-34.00%
	DeepFM	0.8581	0.4695	-6.72%	0.671	0.6497	-31.76%	0.6971	0.6271	-25.54%
	FATE	0.8553	0.4737	-7.45%	0.6762	0.6497	-29.69%	0.7356	0.6199	-10.99%
	HPMN	0.8526	0.4976	-8.15%	0.7599	0.5911	3.71%	0.7681	0.5976	1.28%
	MIMN	0.8457	0.5008	-9.95%	0.7533	0.6002	1.08%	0.7667	0.5998	0.76%
	DIN	0.8796	0.4292	-1.12%	0.7433	0.6086	-2.91%	0.7647	0.6044	0.00%
	DIEN	0.8839	0.4272	0.00%	0.7506	0.6082	0.00%	0.7485	0.6019	-6.12%
Retrieval-based Models	SIM (Item-level)	0.8857	0.4520	0.47%	0.7825	0.5795	12.73%	0.7600	0.6089	-1.78%
	UBR (Item-level)	0.8975	0.4368	3.54%	0.8169	0.5432	26.46%	0.7952	0.5747	11.52%
	RIM (Teacher model)	0.9151	0.3697	8.13%	**0.8567***	**0.4546***	42.34%	0.8005	0.5736	13.52%
	DERT° (Sample-level)	0.9200	0.3585	9.40%	0.8647	0.4486	45.52%	0.8087	0.5219	16.62%
Our Model	ROK	**0.9226***	**0.3546***	10.08%	0.8382	0.5098	34.96%	**0.8093***	**0.5304***	16.85%

知乎 @SmartMindAI

在实验中，提升了基线模型在天猫、淘宝和支付宝数据集上的AUC分别提升了10.08%、34.96%和16.85%，证明了其卓越性能。 item-level retrieval-based methods（如SIM和UBR）这类检索模型，并且在性能上与RIM这样的研究前沿接近，尽管其基础是知识蒸馏和对比学习。选择中等表现的DIN和DIEN作为对照，是为了最大化提升可能，未来可以考虑使用更优秀的教师模型来进一步优化ROK。

Compatibility Analysis: RQ2

Model	Tmall				Taobao				Alipay			
	AUC	Rel. Impr.	LL	Rel. Impr.	AUC	Rel. Impr.	LL	Rel. Impr.	AUC	Rel. Impr.	LL	Rel. Impr.
DeepFM	0.8585	4.21%	0.4803	9.31%	0.6710	3.52%	0.6497	2.51%	0.6971	5.59%	0.6271	4.80%
DeepFM+ROK	**0.8946***	-	**0.4356***	-	**0.6946***	-	**0.6334***	-	**0.7361***	-	**0.5970***	-
DIN	0.8796	4.26%	0.4292	5.59%	0.7433	8.85%	0.6086	6.82%	0.7647	5.83%	0.6044	12.24%
DIN+ROK	**0.9171***	-	**0.4052***	-	**0.8091***	-	**0.5671***	-	**0.8093***	-	**0.5304***	-
DIEN	0.8839	4.38%	0.4272	20.50%	0.7506	11.67%	0.6082	16.18%	0.7485	5.28%	0.6019	7.61%
DIEN+ROK	**0.9226***	-	**0.3546***	-	**0.8382***	-	**0.5098***	-	**0.7884***	-	**0.5561***	-

知乎 @SmartMindAI

(1) 如与DeepFM、DIN和DIEN结合时，平均来看，无论是AUC提升4.30%、8.01%还是对log-loss的减少11.80%、8.51%、8.22%，在所有数据集上都表现出检索导向知识的有效增强。 (2) 无论它们专注于特征交互或行为建模，都能轻松融入。这说明，强调了其兼容性和模型无关性。这使得我们能够在不同情境中自由选择模型，并利用，不受单一模型限制。

Industry Application

通过实证研究，我们观察到，尽管可能涉及复杂的超参数调整以找到最佳α值以平衡检索和对比式正则化，但。这表明，尽管可能需要一些计算投入，但其在提升模型性能的同时，能保持快速响应和适应性，体现了其实用价值。

知乎

Phase	Tmall	Taobao	Alipay
Phase 1 (Retrieval)	156 (40)	116 (30)	24 (4)
Phase 2	201	163	68
Total	357	279	92
DIEN without ROK	327	330	79

一是，由于参数冻结和嵌入维度减半，加快了收敛。二是尽管第一阶段和第二阶段与无，但RIM的检索部分耗费了大部分时间。通过优化批量大小和利用缓存，我们成功减少了检索时间。未来，增加对主干模型以外知识库的定期更新可能进一步提升整体效率。

Inference Efficiency

在实际的在线服务场景，如广告平台和推荐系统，高推理效率是关键。利用知识库，我们在处理大规模数据时将样本检索的时间复杂度从$\mathcal{O}(N \log N)$降低至近乎常数的$\mathcal{O}(1)$，通过直接跳过传统检索-聚合步骤，显著提升了效率。这样的优化对于快速响应用户需求，尤其在大数据环境下显得尤为必要。在实际的后端模型如DeepFM、DIEN和DIN与，我们进行了详尽的效率评估。结果显示，基于检索的方法如RIM，其推理时间远高于这些基线，特别是在处理大量数据时，RIM的延迟可能高出DeepFM、DIEN和DIN1-2个数量级，这与推荐系统对即时响应的高要求相悖，使其在工业应用中变得不可行。而，不仅保持了高性能，而且推理速度几乎无增，这在时间复杂度上大幅降低至常数$\mathcal{O}(1)$。

Table 6: Comparison of AUC and inference speed (ms per sample) across models on Tmall, Taobao, and Alipay.

Model	Tmall		Taobao		Alipay	
	AUC	Inference Speed	AUC	Inference Speed	AUC	Inference Speed
DeepFM	0.8585	1.34	0.6710	1.36	0.6971	1.19
DeepFM+ROK	0.8946	1.46	0.6946	1.40	0.7361	1.41
DIEN	0.8839	5.44	0.7506	4.89	0.7485	3.69
DIEN+ROK	0.9226	5.51	0.8382	5.04	0.7884	3.73
DIN	0.8796	1.28	0.7433	1.32	0.7647	1.43
DIN+ROK	0.9171	1.46	0.8091	1.36	0.8093	1.56
UBR (Retrieval)	0.8975	20.71 (17.67)	0.8169	56.45 (53.32)	0.7952	30.32 (27.31)
RIM (Retrieval)	0.9151	174.81 (173.27)	0.8567	206.22 (204.78)	0.8005	113.95 (112.37)
DERT (Retrieval)	0.9200	17.78 (16.19)	0.8647	19.53 (16.97)	0.8087	17.95 (16.31)

相较于RIM，，显著减少了每样本的推理时间，例如在Tmall数据集上，DIEN+3倍，尽管在性能上略有优势。因此，，实现了在在线服务部署的高可行性和效率，成为理想的选项。，能够无缝融入现有的推荐系统，无需额外数据传输，简化了过程，大大提升了效率。它兼顾了预测精度和快速推理，使其成为现代在线服务领域内高效且引人注目的解决方案。

原文《Retrieval-Oriented Knowledge for Click-Through Rate Prediction》

发布于 2024-06-04 17:52 · IP 属地北京

ctr 华为 深度学习（Deep Learning）

理性发言，友善互动