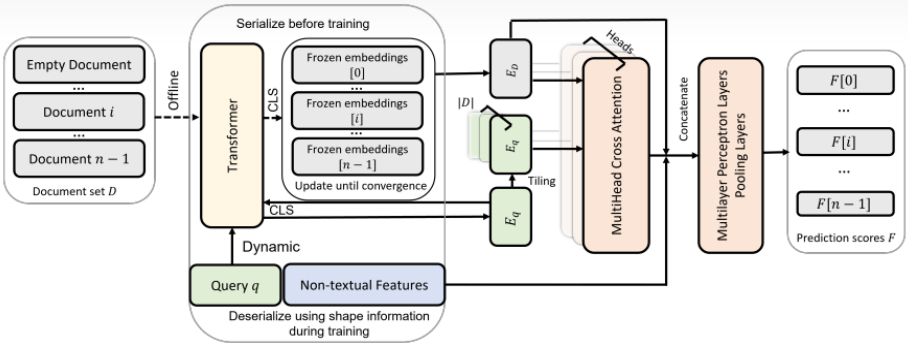

微软2023 - SRANK：引领学习排序新篇章，探索显式与隐式排序框架的奥秘

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

本文提出了一种新的学习排序框架$sRank$，适用于二元相关性检索组件（SR和ACI），并已在生产和推理过程中进行了优化和泛化。同时，为了提高搜索效率，我们采用多阶段检索系统，并在其中使用相对轻量级且效果较差的模型（如相似嵌入空间）。然而，这种方法仍不能解决SR和ACI任务中的检索问题。ACI产品要求快速给出建议，具有低错误容忍度，并在必要时提供智能回复以减少对医生的影响。我们的sRank框架使用双编码器风格的跨注意力架构，在训练和实时预测中学习到一种有效的医生模板选择和文档生成方法。我们提供了训练技术和线性二元交叉熵方法，同时解释了当没有正确匹配项时sRank如何返回结果。

本文的主要贡献在于开发出高效且实用的医生模板选择和文档生成模型。

- 本文提出了一个高效的自训练跨注意力学习排序模型$sRank$。它适用于实时应用、多种损失函数，并且可扩展到可变批次大小。该模型专注于二元相关性应用，但也可用于多级排序或通用对比学习。
- 本文证明了二元相关性的pairwise cross-entropy时间复杂度为$O(n)$，相比通用的RankNet损失函数的$O(n^2)$，在开源学习排序框架TFR-BERT中通过批处理降低至$O(n^2)$。此外，我们利用缓存文档嵌入的方式减少了推理复杂度，在训练过程中实现自我训练和更新嵌入。
- 优化后的排序组件在真实世界的SR和ACI行业中具有广泛应用，并且能够提供高达11.7%至35.5%的准确率提升和下游应用改进。该组件易于扩展到其他行业应用。

Background and Related Work

Classical Learning-to-rank

当评估经典和通用检索系统的关联等级（例如从0到5，其中0对应最不相关，5对应最重要）时，一般偏好于使用列表形式的损失函数而非对偶形式。这是因为神经网络的计算复杂度为$O(n)$，而非$O(n^2)$，所以候选文档集中的每个文档都有一个唯一的正确文档。

假设$f(q, D^q)$用于排序查询q和相关的候选文档集D^q。由于每个候选文档集只有一个正确的答案，即标签为1的文档d^+和剩余所有标签为0的候选文档D^-，所以在LambdaRank或LambdaMart中NDCG差异在其梯度上变为等价。我们提出了一种代表性的方法，即RankNet和MLE损失函数，分别最大化在Equation 2和Equation 3中出现的概率分布$P(d^+)$的对数似然。

$$P_t(d^+) = \frac{1}{|\{D^-\}|} \sum_{d^- \in \{D^-\}} \frac{1}{1+e^{-(f(q,d^+)-f(q,d^-))}}.$$

$$P_t(d^+) = \frac{1}{1+\sum_{d^- \in \{D^-\}} e^{-(f(q,d^+)-f(q,d^-))}}.$$

知乎

Transformer-based Re-ranking

sRank是一种不需要额外负例的批处理方法，它将每个批次中的正例和负例都明确定义。相比之下，其他方法通常只从训练集中提取负例，并且所有批次的大小都相同。sRank在较少的训练资源下获得更好的性能，并能在实时环境中执行。我们提出了一种名为"双编码交叉注意力"的sRank方法， 在实时执行中使用。

Application-Specific Requirements

在ACI模板排序任务中，每个医生的历史模板集都是独特的，并且不能与其他医生的模板集共享。此外，医生可能具有不同的模板数量，而模板的大小也可能会有所不同。在这种情况下，传统的重新排序方法可能不是理想的，因为神经网络模型⁺的训练批次大小是固定的。为了应对这个问题，采用一种通用的方法，在训练中对候选文档进行截断或填充，从而获得一致的批次大小。 最后，SR和ACI的任务相关性标签都是二进制⁺的，这意味着最多只能有一个回复或模板显示。在这种情况下，对偶sigmoid交叉熵比列表损失函数更合适。

Methodology

Semantic Cross Attention Ranking

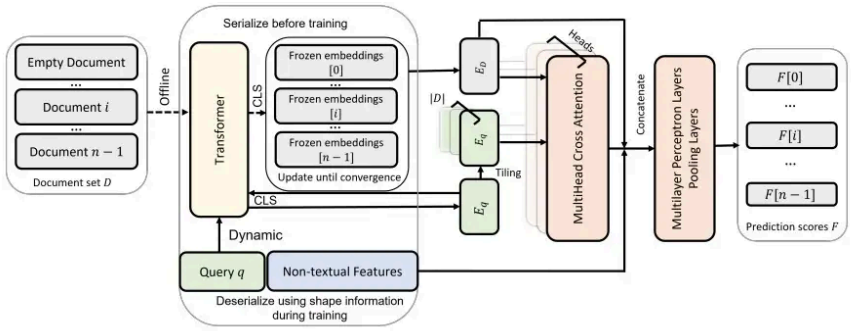

Figure 1: The sRank model takes the serialized record of a query and the frozen embeddings for all its candidates' documents before training, deserializes for Multi-Head Cross Attention, and generates the prediction scores F for the candidate documents.

本文提出了一种新的基于图的架构------多头交叉注意力查询键值架构（$\boldsymbol{E_q}, \boldsymbol{E_D}, \boldsymbol{E_D}$），用于满足基于文本的学习到排序的研究期望。为了避免使用成本⁺较高的方法，例如在查询和候选文档上同时训练嵌入或使用自定义数据预训练Transformer，我们利用学习到排序的训练过程来更新transformer⁺中的权重以生成更具有信息性的候选文档嵌入。此外，本文还讨论了如何将多头交叉注意力应用于冻结的候选文档嵌入$\boldsymbol{E_D}$和动态的每个查询$\boldsymbol{E_q}$的嵌入，以及如何使用ONNX量化来进行推理。

Algorithm 1 Linear pairwise loss for one correct document

Input: labels $Y = (Y_i)_{i=1}^{n}$
Input: prediction scores $F = f(q, (d_i)_{i=1}^{n})$
$P\_DIFF \leftarrow F - F^T$
$L\_DIFF \leftarrow P\_DIFF \cdot Y$
$S \leftarrow exp(L\_DIFF)$
$loss \leftarrow -\frac{1}{n-1} \sum((1-Y) \odot ln\frac{1}{1+S})$
$loss \leftarrow \frac{-ln2+\sum ln(1+S)}{n-1}$ ▷ only one correct document in Y
Return $loss$

Optimized Loss for Binary Relevance

间的预测得分差。L_DIFF是一个大小为$n \times 1$的向量，其中包含了所有候选文档与正确文档之间的线性得分差。当n过大无法放入GPU内存时， 将候选文档分为多个批次，每个批次中包含一个正确的文档。

Use cases and Experiments

在离线评估时，我们使用的主要度量指标是top-one准确率。因为我们的模型被设计为仅能回复一条信息或选择一个医生模板，在SR任务中最多回复一条信息，在ACI任务中最多选择一个医生模板。我们提出的方法相较于MLM损失提升了2-7%。

Smart Reply for Customer Support

智能回复任务是通过高效CPU分类器，从22种Microsoft产品中选择最适合的产品，以在运行时有效地对到来的支持消息交互进行分类，并使用学习排序模型生成最佳回复。使用训练数据和测试数据*集，Smart Reply任务展示了其统计信息。

sRank提高了离线时top-1精度11.7%，并将Smart Reply提供给全球代理进行A/B测试。sRank使Smart回复的CTR绝对提高了42.5%。在A/B测试期间，使用sRank的Smart回复使代理满意度提高13.4%，并使代理编写回复的速度快了38.7%。

Cleaned customer-agent message pairs	1.3 Million
Maximum input tokens	512
Canned reply templates	200
Supported Microsoft products	22
Canned reply templates per product	3-26
Data set size with augmentation	10 Million

Table 1: SR data statistics

Top-one accuracy gains	11.7
Click-through rate (CTR) uplift	42.5
Agent satisfaction improvement	13.4
Time reduction for composing agent messages	38.7

Table 2: SR offline and online metric gains (%)

Template Ranking in ACI

我们通过使用Big Bird RoBERTa生成对话和模板嵌入，在临床环境智能中进行了模板排序的任务。我们会从医生提供的模板中选择最佳模板作为参考。我们会根据这个选择的模板，以及相应的医学会诊脚本，来生成医疗笔记。我们选择了骨科临床记录的物理检查部分作为目标，因为这部分常常需要使用模板。我们使用的基准系统是一个DPR排序模型，它 在满足时间限制的情况下完成这个任务。

知乎

Total number of medical templates	6118
Medical templates per physician	8-39

Table 3: ACI template modeling configuration

ROUGE-L of baseline relative to ROUGE-L with oracle templates	46.0
ROUGE-L of sRank relative to ROUGE-L with oracle templates	92.0
Top-one accuracy gain over DPR	35.5
Top-one accuracy gain (<25% new templates)	41.5
Top-one accuracy gain (25-75% new templates)	40.6
Top-one accuracy gain (>75% new templates)	20.7

Table 4: ACI sRank metric gains (%)

当使用\"oracle\"模板指导医疗笔记生成时，医疗笔记的质量得到显著提高。然而，如何在运行时预测正确的模板是面临的一个挑战。无模板指导的情况下，笔记生成仅能达到使用指导的ROUGE-L的46%。DPR模型不能预测模板的足够准确性，使用其模板导致的ROUGE-L下降率为2%。相比之下，sRank模型 更准确地预测模板，并使端到端的ROUGE-L比没有模板使用增加了92%的ORACLE-ROUGE-L。这表明sRank具有引导更高质量医疗笔记生成的有效能力。此外，表格显示了sRank在ACI指标方面的排序增强。对于排序增强，尽管sRank模型实现了35.5%的更高top-one精度，但同时也减少了7.5%的推理时间。为了评估sRank对抗模板编辑的鲁棒性，我们还验证了它在包含各种频率编辑模板的测试会话中仍然能够实现准确度提高。

Conclusion

本文介绍了跨注意力学习到排序的sRank模型在微软工业任务中的应用，并对其进行了优化和改进，提升性能。sRank具有高质量和快速性，可能适用于其他需要从候选集中选择最佳选项的行业排序任务。

原文《Explicit and Implicit Semantic Ranking Framework》

编辑于 2024-02-19 14:13 · IP 属地北京

微软（Microsoft） 工业级推荐系统 ctr预估