Meta生成式推荐器：基于万亿参数顺序变换器的推荐系统

原创 方方 方方的算法花园 2024年11月14日 08:54 北京

/论文概况\

论文标题： Actions Speak Louder than Words: Trillion-Parameter Sequential Transducersfor Generative Recommendations 《行动胜于言语：用于生成式推荐的万亿参数顺序变换器》
论文链接： https://arxiv.org/pdf/2402.17152
作者所在机构： Meta AI
一句话概括： 将推荐问题重新表述为生成式建模框架内的顺序转导任务，设计了新架构 HSTU，实验表明其在多个数据集和实际应用中表现优异，提升了推荐系统的性能和效率，为推荐系统发展提供了新方向。

/论文挑战\

1.特征结构缺失： 推荐系统中的特征缺乏明确结构，而异构特征（如高基数 id、交叉特征等）在工业级模型中作用关键。

2.动态词汇挑战： 推荐系统使用的是十亿级别的动态词汇，与语言模型中的静态词汇不同，这给训练和推理带来了挑战，且在目标感知方式下考虑大量候选时推理成本高。

3.计算成本瓶颈： 推荐系统需处理的用户行为数据量巨大，计算成本成为大规模顺序模型的主要瓶颈，相比之下，GPT - 3 的训练规模虽大，但推荐系统每日处理的用户行为数量级更高。

/论文贡献点\

1.提出生成式推荐器（GRs）： 将推荐问题重新表述为顺序转导任务，统一了 DLRMs 中的异构特征空间，使排名和检索任务能以生成式方式训练，提高了训练效率和模型性能。

2.设计层次顺序转导单元（HSTU）： 修改了注意力机制以适应大且非平稳的词汇，利用推荐数据集特性提高计算速度，在长序列上比 FlashAttention2 - based Transformers 快 5.3x 至 15.2x。

3.提出 M - FALCON 算法： 通过微批处理充分分摊计算成本，在相同推理预算下，使模型复杂度提高 285 倍，吞吐量提高 1.50x - 2.99x。

4.验证技术有效性： 在合成数据集、公共数据集和大型互联网平台上进行实验，证明 GRs 在离线和在线评估中均显著优于 DLRMs，且模型质量随训练计算量呈幂律扩展，为推荐系统的发展提供了新方向。

/从 DLRMs 到 GRs\

将推荐问题从传统的深度学习推荐模型（DLRMs）重新表述为生成式推荐器（GRs）中的顺序转导任务，具体内容如下：

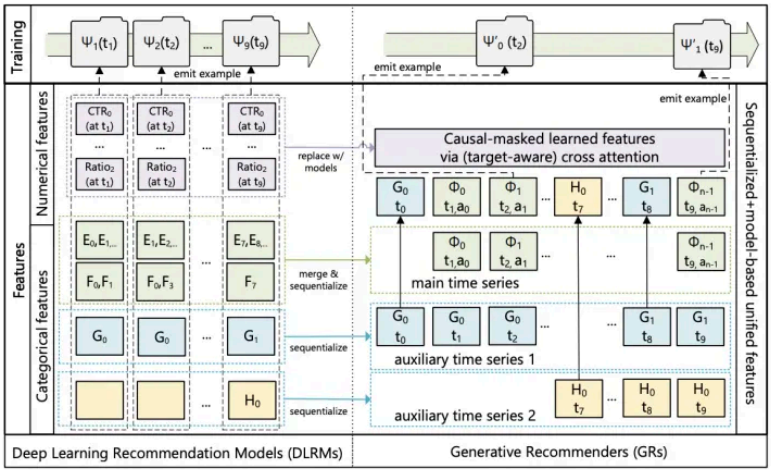

Figure 2. Comparison of features and training procedures: DLRMs vs GRs. E, F, G, H denote categorical features. Φ_i represents the i-th item in the merged main time series. $\Psi_k(t_j)$ denotes training example k emitted at time t_j. Full notations can be found in Appendix A.

1.统一DLRMs中的异构特征空间

（1）分类（稀疏）特征：将用户相关的分类特征（如喜欢的项目、关注的创作者等）顺序化，选择最长时间序列（通常为用户交互的项目）作为主时间序列，对其他缓慢变化的特征（如人口统计信息或关注的创作者）进行压缩并合并到主时间序列中。

（2）数值（密集）特征：由于其频繁变化且从计算和存储角度难以完全顺序化，考虑到分类特征已被顺序化和编码，在GRs中通过足够表达力的顺序转导架构和目标感知公式，可在增加序列长度时有效捕获数值特征，从而去除数值特征。

2. 将排名和检索重新表述为顺序转导任务

（1）排名任务：在GRs中，排名任务面临挑战，因为工业推荐系统常需"目标感知"公式，通过交错项目和动作，将排名任务表述为$p(a_{i+1}|\Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{i+1})$，并应用小神经网络将输出转换为多任务预测，从而实现对所有*nc*次交互的目标感知交叉注意力计算。

（2）检索任务：标准的检索任务在因果自回归设置下也被定义为顺序转导任务，其输入为一系列用户与内容的交互对，输出为根据交互情况确定的内容序列，若交互为正，则输出对应内容，否则为0。

3.生成式训练

工业推荐器通常在流设置中训练，传统自注意力架构的计算需求在处理长序列时成本过高。为解决此问题，论文提出从传统印象级训练转向生成式训练，通过按用户采样率调整训练成本，将计算复杂度从$O(N^3d + N^2d^2)$降低到$O(N^2d + Nd^2)$，使编码器成本在多个目标上分摊，提高了训练效率。

/HSTU\

生成式推荐的高性能自注意力编码器Hierarchical Sequential Transduction Unit（HSTU），包括其设计动机、结构、优势以及相关算法，具体内容如下：

1.HSTU设计动机

（1）为了使生成式推荐系统（GRs）能在工业规模下有效处理大规模、非平稳词汇表，需要新的编码器设计。

（2）传统推荐系统存在特征缺乏明确结构、使用大规模动态词汇表、计算成本高且难以有效训练等问题，而GRs通过将用户行为视为生成建模的新模态，采用新的特征空间和训练方式，有望克服这些挑战。

2. HSTU结构与原理

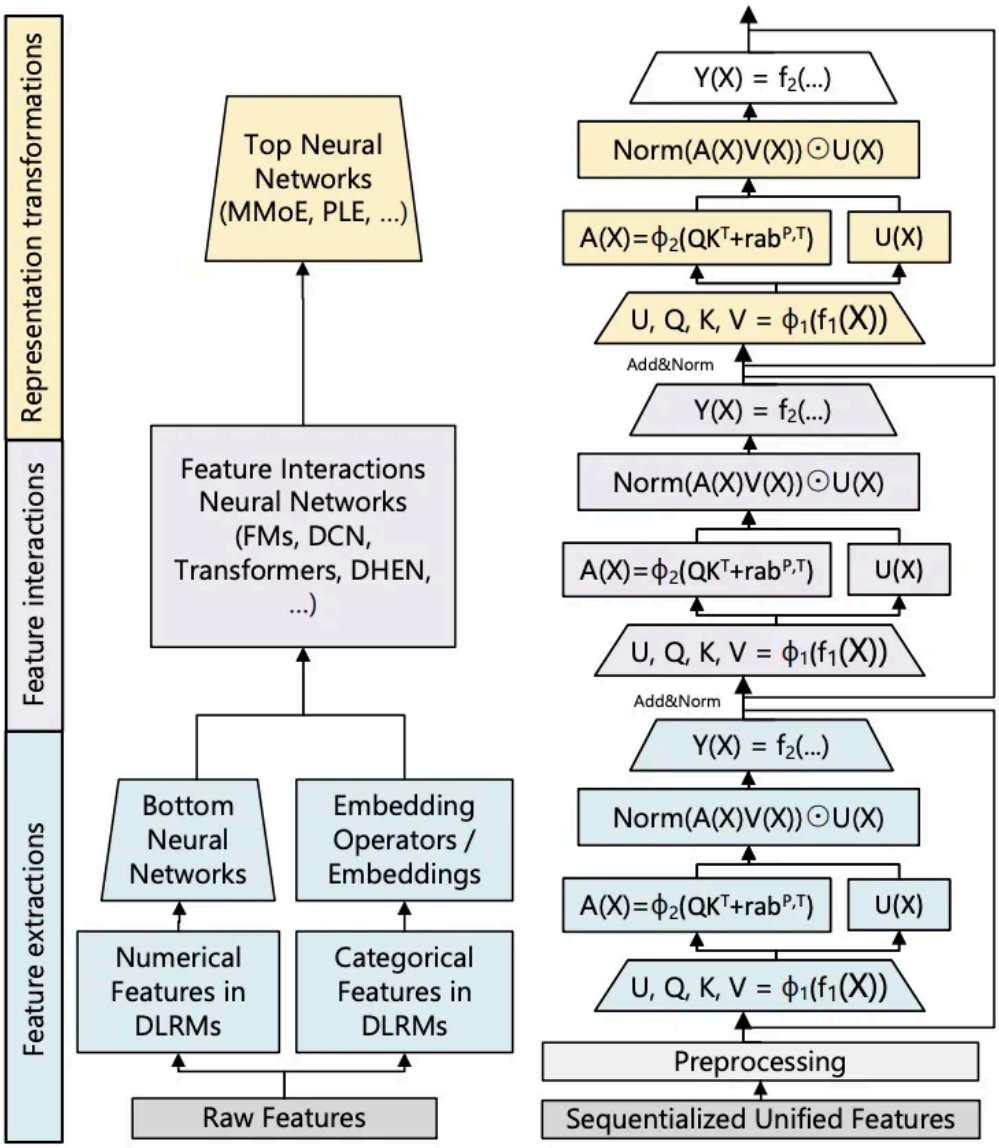

Figure 3. Comparison of key model components: DLRMs vs GRs. The complete DLRM setup (Mudigere et al., 2022) is shown on the left side and a simplified HSTU is shown on the right.

（1）整体架构：HSTU由多个相同层通过残差连接堆叠而成，每层包含点式投影、空间聚合和点式变换三个子层，通过这种设计简化了传统推荐模型（DLRMs）中的异构模块，实现了高效计算。

（2）点式聚合注意力机制：采用新的点式聚合（归一化）注意力机制，与传统softmax注意力机制不同，它通过计算点式池化后的归一化因子，更好地捕捉用户偏好强度，同时在处理非平稳词汇表时表现更优，这在合成数据实验中得到验证，HSTU相对标准Transformer在Hit Rate@10指标上有显著提升。

（3）利用和增加稀疏性：推荐系统中用户历史序列长度分布往往偏斜，可利用这种稀疏性提高编码器效率。HSTU开发了高效的GPU注意力内核，通过融合计算将注意力计算转化为分组矩阵乘法（GEMMs），减少内存访问，使自注意力计算变为内存受限，提升了2 - 5倍吞吐量。此外，还通过随机长度（SL）算法进一步增加稀疏性，根据用户历史序列长度有选择地缩短序列，在不影响模型质量的前提下显著降低计算成本，实验表明在合适的α值下，SL能在高稀疏度下保持模型性能。

（4）最小化激活内存使用：与Transformers相比，HSTU采用简化且完全融合的设计，减少了注意力外部的线性层数量，将计算融合为单个算子，降低了激活内存使用，使构建更深网络成为可能。同时，通过采用行式AdamW优化器和优化存储方式，减少了大规模原子id表示词汇表时的内存压力。

3. M - FALCON算法提升推理效率

（1）针对推荐系统在推理时需处理大量候选集的问题，提出M - FALCON算法。该算法通过修改注意力掩码，在一次前向传播中并行处理多个候选，将交叉注意力成本从$O(b_m n^2 d)$降低到$O((n+b_m)^2 d)$（当bm相对n较小时），并可将候选集划分为微批次，利用编码器级KV缓存进一步减少成本或降低尾延迟。

（2）实验表明，M - FALCON算法使HSTU - 基于的生成式推荐系统在推理时能处理更复杂模型，在处理1024/16384个候选时，吞吐量比传统DLRMs高1.50x - 2.99x，尽管GR模型计算复杂度是DLRMs的285倍。

LLM与推荐 15　LLM论文阅读 13

LLM与推荐 · 目录

上一篇 · RecRanker：指令调优LLM用于 top-k 推荐排序