arxiv.org/abs/2507.00487

MassTool: A Multi-Task Search-Based Tool Retrieval Framework for Large Language Models

MASSTOOL: 在TOOL-AUGMENTED LLM中用于高效工具召回

代码：https://github.com/wxydada/MassTool

Jianghao Lin[1], Xinyuan Wang[1], Xinyi Dai[2], Menghui Zhu[2], Bo Chen[2], Ruiming Tang[2], Yong Yu[1], Weinan Zhang[1]
[1]Shanghai Jiao Tong University, [2]Huawei Noah's Ark Lab
{chiangel,wnzhang}@sjtu.edu.cn

简介

本文提出MassTool，面向Tool-Augmented LLM的工具检索(tool retrieval)框架，专注于在调用外部工具前先筛选候选工具子集。与仅依赖查询与工具描述语义匹配的现有方法不同，MassTool首次提出应该先判断query是否值得llm调用工具，因此将工具检索细分为工具使用检测(Tool Usage Detection)和工具检索两个子任务，然后提出双塔结构用多任务学习联合建模，即dual-step sequential decision-making流程，先判断query是否需要调用工具，再检索合适工具。此外，MassTool从用户意图建模角度更深入细致的分析query。

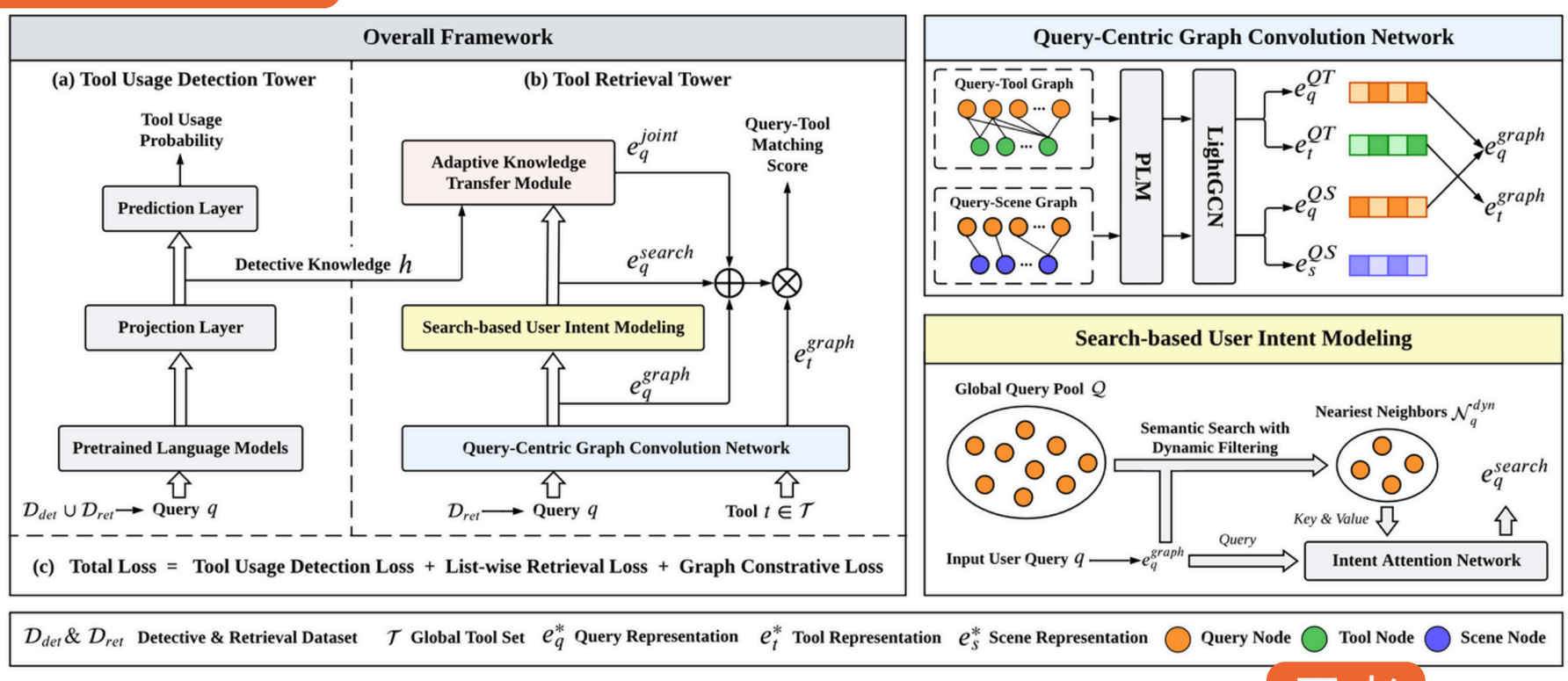
背景

在Tool-Augmented LLM使用时，如果tool是海量的api而非一两个tool，比如之前读过的ToolLLM，通常会有一个工具检索(tool retrieval)阶段，类似于搜索、推荐中两阶段的第一个阶段召回(粗排/粗筛)，先得到一个小点的tool候选集，再让llm从中选择合适的tool进行generation。
目前的工具检索基本上匹配query和tool descrption文本的语义相似度，比如用bert来做，作者觉得太简单了，借鉴搜索引擎的做法，引入用户意图(user intent)。

实验设置

- 数据集：Tool-Lens, ToolBenchG2, ToolBenchG3
- 评价指标：Recall@K和NDCG@K
- 训练框架：BEIR

MassTool框架



Tool Retrieval Tower有点过于复杂了，包含query和tool构成的二部图(用LightGCN建模)、检索query的邻居query增强意图表示(SUIM)、多种信息融合模块(AdaKT)

部分实验结果

Backbone	Framework	ToolLens				ToolBenchG2				ToolBenchG3		
		R@3	R@5	N@3	N@5	R@3	R@5	N@3	N@5	R@3	R@5	N@3
BM25	-	0.2158	0.2688	0.2319	0.2609	0.1706	0.2138	0.1783	0.1988	0.2933	0.3588	0.3220
ANCE	Raw	0.2083	0.2656	0.2145	0.2457	0.2083	0.2656	0.2145	0.2457	0.2155	0.2638	0.2344
	QTA	0.7718	0.9051	0.7892	0.8652	0.5545	0.6729	0.5822	0.6383	0.6408	0.7541	0.6855
	MMRR	0.7591	0.8992	0.7726	0.8524	0.5674	0.6839	0.5947	0.6524	0.6226	0.7422	0.6703
	APIRetriever	0.8062	0.9417	0.8235	0.9015	0.5858	0.6720	0.5858	0.6375	0.6511	0.7663	0.6927
	COLT	0.9215	0.9778	0.9278	0.9610	0.7076	0.8059	0.7076	0.7798	0.7337	0.8397	0.7795
	MassTool (Ours)	**0.9648***	**0.9847***	**0.9670***	**0.9785***	**0.7927***	**0.8678***	**0.8124***	**0.8429***	**0.7840***	**0.8662***	**0.8259***
	Rel.Imprv.	4.69%	0.71%	4.23%	1.82%	12.02%	7.68%	10.32%	8.09%	6.86%	3.16%	5.95%
TAS-B	Raw	0.1910	0.2371	0.1981	0.2233	0.1910	0.2371	0.1981	0.2233	0.2532	0.3115	0.2780
	QTA	0.7731	0.9031	0.7883	0.8623	0.5736	0.6872	0.6033	0.6561	0.6497	0.7637	0.6964
	MMRR	0.7607	0.8893	0.7785	0.8518	0.5786	0.6982	0.6061	0.6626	0.6419	0.7602	0.6844
	APIRetriever	0.8126	0.9406	0.8254	0.8994	0.6278	0.6749	0.5896	0.6421	0.6604	0.7764	0.7041
	COLT	0.9149	0.9691	0.9248	0.9563	0.7164	0.8112	0.7460	0.7874	0.7449	0.8458	0.7903
	MassTool (Ours)	**0.9523***	**0.9812***	**0.9577***	**0.9744***	**0.7958***	**0.8684***	**0.8164***	**0.8455***	**0.7923***	**0.8675***	**0.8338***
	Rel.Imprv.	4.08%	1.25%	3.56%	1.89%	11.08%	7.05%	9.44%	7.38%	6.36%	2.57%	5.50%

思考

针对海量api的tool use场景，第一阶段先工具检索是非常有必要的，只不过本文的设计过于复杂了，不清楚扩展性如何，比如基于训练集query和金标准tool构造二部图，在真实应用时新的query如何加到图中？
现在RLVR + TIR一般只有几个tool，还用不到检索，而且通过RLVR训练的llm已经具备一定的是否、何时调用tool能力。

@机器爱学习