arxiv.org/abs/2507.02592

WebSailor: Navigating Super-human Reasoning for Web Agent

Kuan Li*, Zhongwang Zhang*, Huifeng Yin*(✉), Liwen Zhang*, Litu Ou*, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang(✉), Ming Yan, Pengjun Xie, Fei Huang, Jingren Zhou

Tongyi Lab, Alibaba Group

要想Deep Search能力强，训练集必须高难度

https://github.com/Alibaba-NLP/WebAgent

简介

本文提出WebSailor，目的是训练出具备超强Deep Search能力的开源Web Agent，作者认为要想Deep Search能力强，训练集的难度必须高，为此，作者首先创建了answer既有难度同时query模棱两可的SailorFog-QA训练集，简单来说，利用真实互联网随机游走构建复杂的知识图谱，通过子图采样与模糊化描述生成需要多步推理的(query, answer)数据。有了训练集，作者采用两阶段训练法：sft和RLVR，为了高效的RL训练，作者基于DAPO进行改进提出了Duplicating Sampling Policy Optimization (DUPO) 算法。

背景

Deep (Re)Search类型的论文读了很多了，背景意义就不再提了，只说下作者的insight: 目前开源社区deep search agent能力之所以上不去，很大原因是训练用的QA数据太简单。为此作者先构建了一个高难度的QA训练集，再基于数据集用两阶段sft和RLVR训练。

实验设置

典型的TIR工作

两个tool：搜索引擎和网页访问

- 实验对象：Qwen2.5 系列，框架：Megatron(sft)和verl(rl)，强化学习算法：DUPO
- RLVR格式的Reward设计: format reward和answer reward两项，用LLM-as-Judge计算answer reward

$$R_i = 0.1 * R_i^{format} + 0.9 * R_i^{answer}$$

WebSailor描述

SailorFog-QA数据集的创建：

1. 构造基于模糊实体(fuzzy entity)的知识图谱：先从Wikipedia找一批模糊实体(fuzzy entity)，然后去互联网检索信息，抽取相关联实体和关系，得到一个小知识图谱，然后随机采样实体节点，继续Internet检索得到相关实体和关系来扩充知识图谱。以上步骤迭代多次，就会得到一个大规模的以模糊实体为基础的知识图谱。
2. 从知识图谱采样子图并模糊描述：基于知识图谱进行子图采样，然后从子图中构建query和answer，为了让query更难，故意将一些时间啊人命啊模糊化，人类真坏啊:(
3. 创建reasoning trajectory：因为要做sft，所以需要reasoning trajectory，如果直接让已有的推理模型生成，轨迹中会包含大量啰嗦的，具有不必要风格的thinking内容，为此，作者先让推理模型生成轨迹，再把thinking部分去掉，保留tool调用和返回结果，用另一个llm来专门根据tool调用和返回结果生成thinking，这样得到精简的reasoning trajectory。

两阶段训练：

4. 拒绝采样FT: 说白了就是先对训练数据过滤再sft，这一步是为了让llm具备TIR能力
5. DUPO算法：GRPO变体或者说DAPO变体，训练前先把那些rollout全对的query去掉，太简单了对训练没帮助也浪费资源，训练过程中，把batch中那些rollout全对或全错的query也删掉，为了补充到batch size，采样复制batch内其他query的方式，这样做比DAPO训练效率提升2-3倍。

思考

之前我们读过WebDancer，本文属于同组的续作，这一次作者继续构造高复杂度的(query, answer)，高难度的训练集有没有意义？必须有，高考状元不是刷小学题刷出来的。

就是不清楚训练集是否会开源，其次本文将llm上下文长度限制在32k，考虑到训练成本，看来地主家也缺粮啊。

@机器爱学习