arxiv.org/abs/2309.17452

ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving

开源代码：github.com/microsoft/ToRA

Zhibin Gou[1,2]*, **Zhihong Shao**[1,2]*, **Yeyun Gong**[2]†, **Yelong Shen**[3]
Yujiu Yang[1]†, **Minlie Huang**[1]†, **Nan Duan**[2], **Weizhu Chen**[3]
[1]Tsinghua University [2]Microsoft Research [3]Microsoft Azure AI
{gzb22,szh19}@mails.tsinghua.edu.cn
{yegong,yeshe,nanduan,wzchen}@microsoft.com

简介

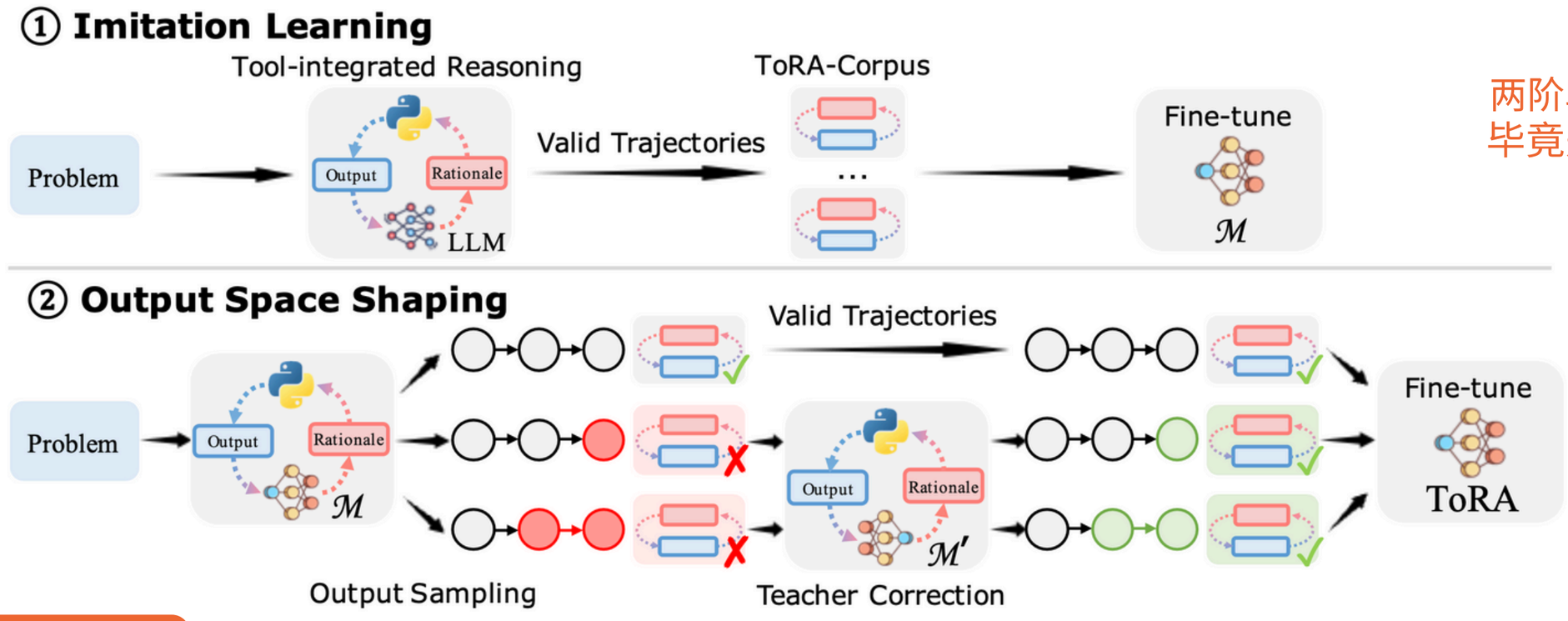
本文为数学推理任务设计了ToRA(Tool-integrated Reasoning Agents)方法，比较早地将自然语言推理与外部工具调用(Python解释器)结合起来，模拟人类“边思考、边计算”的解题方式。

首先用GPT生成TIR训练集，解题过程是迭代式的“语言+工具调用”格式，然后对CodeLlama做sft。重点是，作者认为光用sft去模仿还不够，因为每道题的解法可能有很多种，为了增强多样性(diversity)，又设计了output space shaping阶段，简单说就是让sft模型自己尝试生成多种解题路径，找出其中对的，也把那些错误的用一个更大的教师模型进行改正，这样训练集更大更丰富了，再做sft得到ToRA模型。

背景

大多数LLM在做数学推理时，只是用自然语言一步步写解题过程，并没有真正调用外部计算工具。虽然也有少数方法尝试使用工具（如生成代码来求解），但它们往往缺乏语言推理过程。ToRA是较早真正将“语言推理 + 工具调用（tool use）”融合起来用于数学任务的方案。

ToRA



两阶段SFT，NO RL
毕竟是23年的工作

部分实验结果

Model	Size	Tools	ZS	GSM8k	MATH	GSM-Hard	SVAMP	TabMWP	ASDiv	MAWPS	AVG
Used for training?				✓	✓	✗	✗	✗	✗	✗	
Proprietary Models											
GPT-4	-	✗	✗	92.0	42.5	64.7	93.1	67.1	91.3	97.6	78.3
GPT-4 (PAL)	-	✓	✗	94.2	51.8	77.6	94.8	95.9	92.6	97.7	86.4
ChatGPT	-	✗	✗	80.8	35.5	55.9	83.0	69.1	87.3	94.6	72.3
ChatGPT (PAL)	-	✓	✗	78.6	38.7	67.6	77.8	79.9	81.0	89.4	73.3
Claude-2	-	✗	✗	85.2	32.5	-	-	-	-	-	-
PaLM-2	540B	✗	✗	80.7	34.3	-	-	-	-	-	-
Open-Source Models											
LLaMA-2	7B	✗	✗	13.3	4.1	7.8	38.0	31.1	50.7	60.9	29.4
LLaMA-2 SFT	7B	✗	✓	41.3	7.2	16.1	31.9	27.8	47.4	60.0	33.1
LLaMA-2 RFT	7B	✗	✓	51.2	-	-	-	-	-	-	-
Platypus-2	7B	✗	✗	14.4	5.4	8.6	36.7	26.5	47.9	58.4	28.3
WizardMath	7B	✗	✓	54.9	10.7	20.6	57.3	38.1	59.1	73.7	44.9
CodeLLaMA (PAL)	7B	✓	✗	34.0	16.6	33.6	59.0	47.3	61.4	79.6	47.4
Toolformer†	7B	✓	✓	-	-	-	29.4	-	40.4	44.0	-
ToRA	7B	✓	✓	68.8	40.1	54.6	68.2	42.4	73.9	88.8	62.4
ToRA-Code	7B	✓	✓	**72.6**	**44.6**	**56.0**	**70.4**	**51.6**	**78.7**	**91.3**	**66.5** (+19)

Inference流程

和现在RLVR 做TIR的inference一致

Algorithm 1 Inference of Tool-Integrated Reasoning

Require: problem q, model $\mathcal{G}$, prompt p, external tools $\mathcal{E}$, stop condition $Stop(\cdot)$, maximum iteration rounds n

$\tau_0 \leftarrow$ "" ▷ Trajectory Initialization
for $i \leftarrow 1$ to n **do**
  $r_i \sim \mathbb{P}_{\mathcal{G}}(\cdot|p \oplus q \oplus \tau_{i-1})$ ▷ Rationale Generation (Eq. 1)
  if $Stop(r_i)$ **then** ▷ Stopping Criteria
    return $\tau_{i-1} \oplus r_i$
  end if
  $a_i \sim \mathbb{P}_{\mathcal{G}}(\cdot|p \oplus q \oplus \tau_{i-1} \oplus r_i)$ ▷ Program Generation (Eq. 2)
  $o_i \leftarrow \mathcal{E}(a_i)$ ▷ Tool Execution
  $\tau_i \leftarrow \tau_{i-1} \oplus r_i \oplus a_i \oplus o_i$ ▷ Trajectory Update (Eq. 3)
end for
return τ_n

思考

ToRA是23年的工作，但是本文的reasoning path已经和现在RLVR类型工作一致了，都是由类似(thinking, tool参数,tool结果)的多组三元组构成，或许都可以用WebDancer说的ReAct格式进行解释。那个时候还没有用RLHF/RLVR做TIR的，要么是prompt要么是sft，作者在做完sft之后，思考了一个问题：训练集是(prompt, response)组成的，一个prompt只有一个response，但是数学题，解题过程可以不唯一，为了增加多样性，设计了output space shaping阶段，其实就是扩展数据，结果是一个prompt会对应多个不同的response。但是，回忆下机器学习的内容，训练一个分类模型，如果训练集一个X对应多个y，这可不是好的训练集，所以第一个问题：一个prompt对应多个response是否合理？如果不合理/合理，那么为什么加了shaping能提升效果，llm到底学到了什么？如果像增强output 多样性，看起来RL要比sft合理多了。

@机器爱学习