arxiv.org/abs/2507.05707

Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions

DualDistill：通过双教师轨迹拼接蒸馏+自蒸馏 提升数学推理能力

Yuanzhe Hu[1*], **Yu Wang**[1*], **Julian McAuley**[1]
[1]UC San Diego
[1]{yuh127,yuw164,jmcauley}@ucsd.edu

Datasets Source Code

简介

本文提出DualDistill蒸馏框架，目的是让学生模型学习在数学任务上何时使用纯文本推理，何时使用工具增强推理(TIR)。简单来说，DualDistill包含两个蒸馏步骤：1) 双教师蒸馏，利用一个纯文本推理教师(Deepseek-R1)，擅长抽象、逻辑复杂的推导和一个TIR教师(OpenHands)，擅长执行高效的数值计算和算法 构建训练集，训练数据的特点是将两个教师的推理轨迹拼接，并包含推理策略切换提示(transition segments)引导学生模型理解何时切换推理方式。然后对学生模型做sft；2) 自蒸馏(Self-Distill)，考虑到学生模型的tool-use比TIR教师差不少，又在学生自身生成的推理轨迹基础上由DeepSeek-R1教师进行验证或纠正，进一步sft学生的策略选择和推理能力。

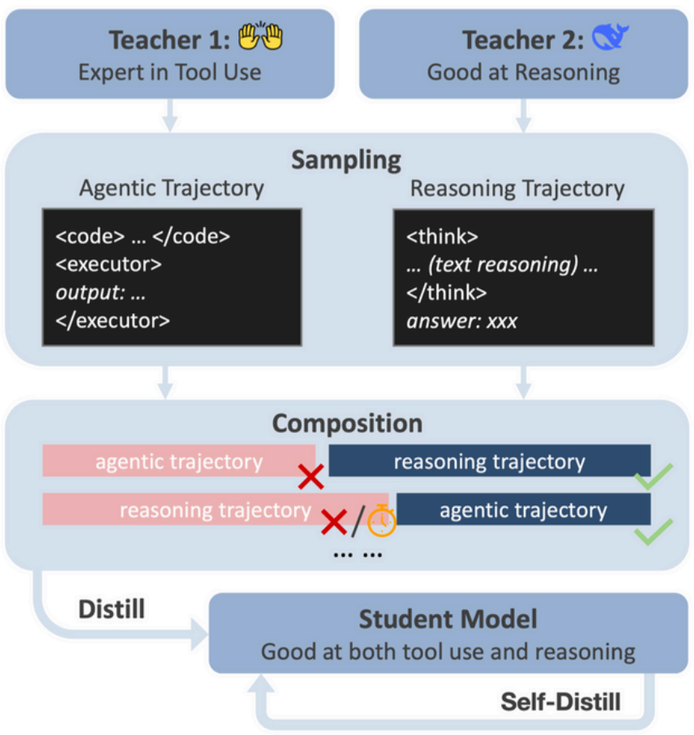
背景

对于数学任务，纯文本推理(text-based reasoning)和工具增强推理(tool-integrated reasoning, TIR)各具优势，TIR能高效完成复杂计算和算法执行，而纯文本推理在抽象推导和逻辑推演中也有独特之处。本文尝试将这两类不同模型的推理能力蒸馏(sft)到同一个学生模型，并且学生模型能够根据数学问题自动切换推理策略。

实验设置

- 任务类型：数学推理，训练集大小：2678
- 训练集的query来自我们读过的**DeepMath-103K**
- 为了让学生模型学会什么时候使用TIR，什么时候使用纯文本推理，作者在构建训练集时，找了两类数学问题: 1) 工具优先子集(Agentic-Favored Subset), 更适合TIR推理的数学题；2) 纯推理优先子集(Pure Reasoning-Favored Subset)，更适合纯文本推理的数学题。两类数据数量相当
- 两个教师模型：OpenHands和DeepSeek-R1，学生模型：Deepseek-R1-Distill-7B

DualDistill流程



Algorithm 1 DUALDISTILL

Input: Teacher policies π_A, π_R; student S_0; training dataset $\mathcal{D} = \{(x_i, a_i)\}_{i=1}^{N}$; thresholds β_1, β_2; sample count K; binary grader $G(\cdot,\cdot)$
Output: Trained student S_2

TEACHER DISTILLATION

Initialize teacher-distillation buffer $\mathcal{T}_1 \leftarrow \varnothing$
for each $(x, a) \in \mathcal{D}$ **do**
  Draw $z \sim \text{Bernoulli}(0.5)$
  $y_1 \sim z\,\pi_A(\cdot|x) + (1-z)\,\pi_R(\cdot|x)$ 其实y1是因为两个教师
  $y_2 \sim (1-z)\,\pi_A(\cdot|x, y_1) + z\,\pi_R(\cdot|x, y_1)$
  $g_1 \leftarrow G(y_1, a),\ g_2 \leftarrow G(y_2, a)$
  switch (g_1, g_2)
    case $(0,1)$: Add $y_1 \oplus t^{-+} \oplus y_2$ to $\mathcal{T}_1$
    case $(1,1)$: Add $y_1 \oplus t^{++} \oplus y_2$ to $\mathcal{T}_1$
    case $(1,0)$: Add y_1 to $\mathcal{T}_1$
  end switch
end for
Balance $\mathcal{T}_1$
Fine-tune S_0 on $\mathcal{T}_1 \rightarrow S_1$

SELF-DISTILLATION

Initialize self-distillation buffer $\mathcal{T}_2 \leftarrow \varnothing$
for each $(x, a) \in \mathcal{D}$ **do** 还是相同的数据集D
  Sample $\{t_j\}_{j=1}^{K} \sim \pi_{S_1}(\cdot|x)$
  $g_j \leftarrow G(t_j, a)$
  $\bar{g} \leftarrow \frac{1}{K}\sum_{j=1}^{K} g_j$
  if $\bar{g} > \beta_1$ **then** verification只用了\pi_R 模型
    Add a correct t_j + verification to $\mathcal{T}_2$
  end if
  if $\bar{g} < \beta_2$ **then**
    Add an incorrect t_j + correction to $\mathcal{T}_2$
  end if
end for
Fine-tune S_1 on $\mathcal{T}_2 \rightarrow S_2$
return S_2

实验

Model	Budget	DeepMath-L	Combinatorics300	MATH500	AIME	AMC	avg.
Qwen2.5-7B-Instruct (w/o tool)	S	17.2	21.8	75.1	8.0	42.9	33.0
	L	17.5	21.8	75.2	8.0	42.9	33.1
Qwen2.5-7B-Instruct (w/ tool)	S	34.7	28.9	70.2	14.7	51.1	39.9
	L	34.7	28.9	70.2	14.7	51.1	39.9
DeepSeek-R1-Distill-7B	S	34.7	34.7	**83.1**	23.3	61.2	47.4
	L	56.3	44.5	89.2	**40.7**	84.8	63.1
Agentic-R1-7B (ours)	S	37.0	36.9	80.0	**28.0**	64.3	49.3
	L	59.3	49.4	82.4	**40.7**	82.2	62.8
Agentic-R1-7B-SD (ours)	S	**40.0**	**38.2**	82.5	27.3	**66.3**	**50.9**
	L	**65.3**	**52.0**	**93.3**	**40.7**	**85.8**	**67.4**

思考

通过设计策略切换提示(transition segments)将两个能力各异的教师推理轨迹拼接去sft 学生模型，竟然能提升不小的推理效果。
对于long/short CoT 推理轨迹数据，是否也可以用拼接方式来做sft呢？说不定也可以提升何时使用long/short CoT的能力，值得做实验。