

2023.11.07 周报

November 7, 2023

1 EntailmentBank 实验

1.1 整理 related work, 比较现有方法及创新点

1.2 Results

- 在推理开始前, 利用语言模型的 deductive reasoning 能力, 一次性生成全部的推论
- 针对 task2, 在推理前加一个 filter, 筛选不相关前提
- 针对 task3, 在推理模块中加入推测缺失知识的要求
- 调整参数 + 多次实验; 结果如 Tabel 1所示

Task type	Method	Leaves		Steps		Intermediates		Overall
		F1	AllCorrect	F1	AllCorrect	F1	AllCorrect	AllCorrect
Task1	EntailmentWriter (T5-large)	98.7	86.2	50.5	37.7	67.6	36.2	33.5
	EntailmentWriter (T5-11B)	99.0	89.4	51.5	38.2	71.2	38.5	35.3
	IGRG	97.6	89.4	50.2	36.8	62.1	31.8	32.4
	MetGen	100.0	100.0	57.7	41.9	70.8	39.2	36.5
	NLProofS	97.8	90.1	55.6	42.3	72.4	40.6	38.9
	REFT (T5-large)	100.0	100.0	54.6	40.7	66.9	36.3	34.8
	Ours	100.0	100.0	45.5	38.5	68.4	55.3	38.5
Task2	EntailmentWriter(T5-large)	84.3	35.6	35.5	22.9	61.8	28.5	20.9
	EntailmentWriter (T5-11B)	89.1	48.8	41.4	27.7	66.2	31.5	25.6
	MetGen	82.7	46.1	41.3	29.6	61.4	32.4	27.7
	NLProofS	90.3	58.8	47.2	34.4	70.2	37.8	33.3
	IGRG	69.9	23.8	30.5	22.4	47.7	26.5	21.8
	REFT	81.0	39.0	38.5	28.4	56.3	28.6	25.7
	Ours	71.9	45.2	48.9	38.7	55.9	47.3	23.4
Task3	EntailmentWriter	39.9	3.8	7.4	2.9	35.9	7.1	2.9
	MetGen	34.8	8.7	9.8	8.6	36.6	20.4	8.6
	NLProofS	43.2	8.2	11.2	6.9	42.9	17.3	6.9
	IGRG	46.6	10.0	11.3	8.2	38.8	36.5	11.8
	REFT	46.2	11.4	15.2	9.6	41.4	17.6	9.4
	FAME	43.4	13.8	16.6	12.4	40.6	19.9	11.9
	Ours	51.1	11.1	18.4	11.2	40.1	37.3	11.1

Table 1: Results on EntailmentBank

2 继续阅读 AI4Math 的论文

总结出将机器学习方法与定理证明结合的以下几种分类范畴:

- ATP 和 ITP:
纯 ATP/利用 machine learning 代替 ITP 中的 human, 将其转化成一个完全 automated 的过程
- 使用的深度学习方法
传统 ML/DL 方法 vs LLM:
traditional(svm 等)/ sequence model (rnn,lstm) / gnn / RL /LLM generative model
- 根据应用 DL 的不同模块/任务, 如: premise selection, tactic prediction, proof generation, RL for proof guidance, error repair, pre-train method/pre-train synthesized data 等

3 计划

- 对于 Entailment Bank 任务，考虑引入错误修正模块，作为一个新的创新点，尝试这部分的代码和 baseline(T5-Large,NLProofs)
- 根据 Section2 中的分类方法，继续阅读并整理 AI4Math 论文