

内容来源: csdn.net

作者昵称: hdubigben

原文链接: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

坐井说天阔---DeepSeek-R1

已于 2025-02-15 13:46:02 修改

编辑



DeepSeek技术社区 文章已被社区收录

加入社区

前言

DeepSeek-R1 这么火，虽然网上很多介绍和解读，但听人家的总不如自己去看看原论文。于是花了大概一周的时间，下班后有进入了研究生的状态---读论文。

DeepSeek这次的目标是探索在没有任何监督数据的情况下训练具有推理 能力的大模型，最初的方向是通过纯强化学习过程进行自我进化来具备推理能力。当然这个目标并未完全实现，最终还是使用到的监督数据进行 微调。

做的第一个尝试就是使用大规模强化学习 (RL) 训练，训练出来的模型称之为: DeepSeek-R1-Zero. 实验结果显示该模型的输出可读性差和语言混合等挑战。

然后考虑是否可以使用少量高质量的数据进行少样本监督训练和强化学习相结合的方式，来提升推理能力，并且提升输出结果质量。这一种思路训练出来的模型便是DeepSeek-R1

顺着这个思路，DeepSeek又进行了下一步的尝试，上一步使用的数据在其他大模型（如阿里的QWen）上进行微调训练，即蒸馏过程。实验结果显示该过程可以显著提升其他没有推理能力的模型也具有良好的推理能力。

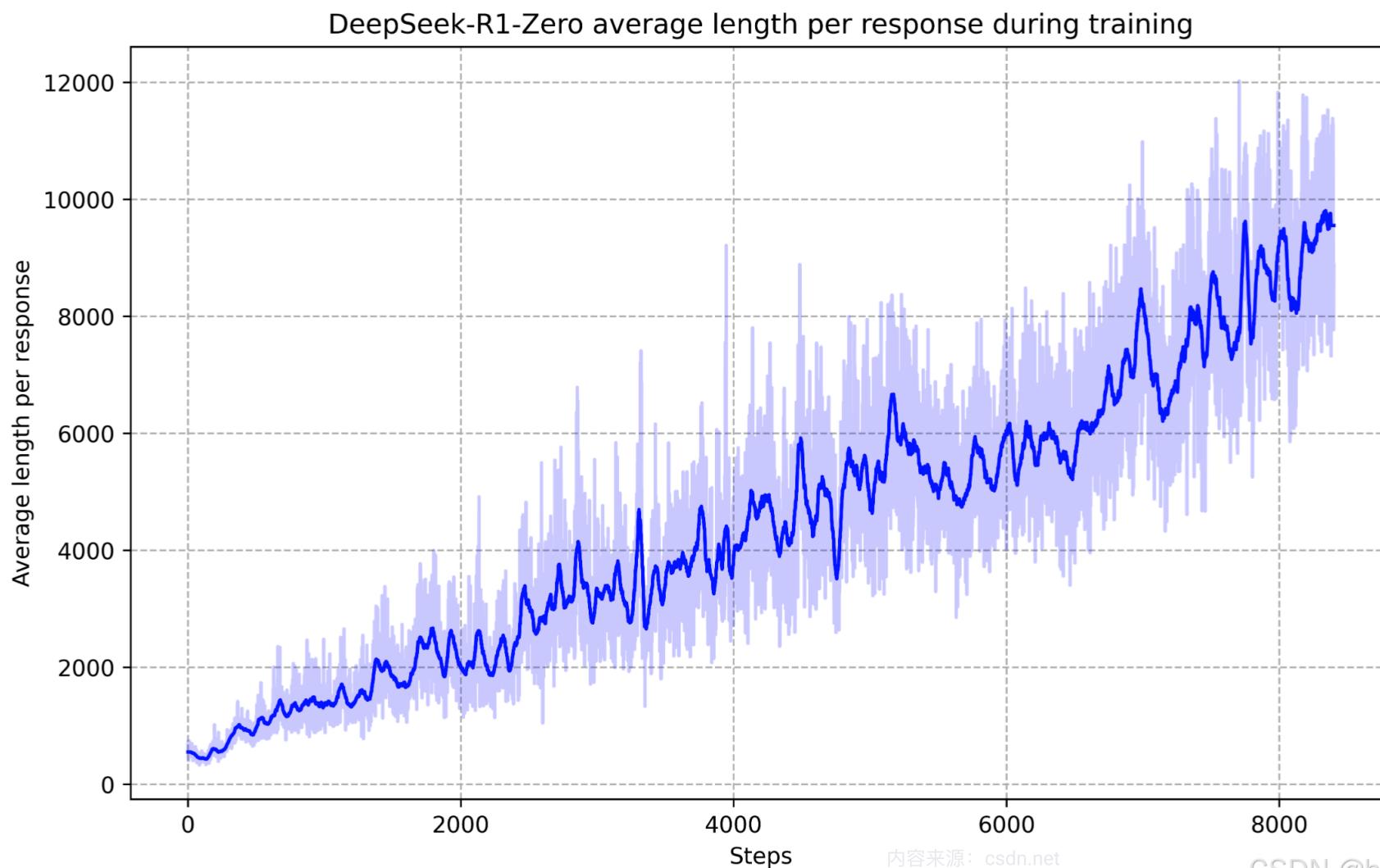
详情如下，水平有限，内容可能有误，仅个人理解，个人记录而已。

DeepSeek-R1-Zero

为了实现目标，首先做了开放研究尝试，只用通过大规模强化学习 (RL) 训练的模型，仅通过 RL 有效学习和泛化的能力，没有监督微调 (SFT)，训练出来的模型称为：DeepSeek-R1-Zero，但实验发现该模型它遇到了可读性差和语言混合等挑战。
内容来源：csdn.net
作者主页：https://blog.csdn.net/weixin_42754905/article/details/145648999

虽然这个尝试不尽人意，但其训练过程还是具有一定参考意义。也是实验过程中发现了两件有意思的事情。

首先训练过程中，发现随着测试时间的增加，DeepSeek-R1-Zero出现了自我进化的现象，会进行自我进行更复杂的行为，如下图所示，比如会主动重新评估先前的步骤，会自己探索解决方案的代替方法。这些行为是与强化学习环境交互自己出现的，并且显著增强了模型的推理能力，从而使得训练出来的模型可以更高效，更准确的处理更具有挑战性的任务



另外一件事是“Aha Moment”，DeepSeek-R1-Zero 自己会分配更多的思考时间来重新评估其初始方法，从而思考是否有更有效，更准确的解决方案。下图给了一个具体的例子。

“Aha Moment”说明了，对于大模型推理能力的训练，不需要明确的教它如何解决问题，只需要简单地为其提供正确的激励措施，然后它就会自主进行思考从而找到更有效的解决策略。“Aha Moment”也在提醒我们，强化学习有可能在AI训练中解锁新的智能水平，为未来开发出更加自主和自适应的大模型铺平道路。

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a+x}} = x$$

First, let's square both sides:

$$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

CSDN @hdubigben

内容来源: csdn.net

作者昵称: hdubigben

原文链接: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

DeepSeek-R1，使用冷启动进行强化学习

DeepSeek-R1-Zero 遇到了可读性差和语言混合等挑战。为了解决DeepSeek-R1-Zero遇到的问题，从两个方向进行了尝试，

1. 是不是可以使用少量高质量的数据作为冷启动 (cold start) 来进一步提高推理性能或者减少训练成本
2. 对于DeepSeek-R1-Zero 遇到的可读性差和语言混合等挑战，如何训练出一个模型。用户友好，可以产生清晰连贯的思维链 (CoT)，也有强大的通用能力。

为了获取高质量的数据，从以下三种方式进行了数据生产，论文中称之为cold start data（冷启动数据）。

1. 用具有长思维链 (CoT) 的少样本 (few-shot) 提示词，
2. 直接让模型来通过思考和验证来生成详细的答案
3. 收集DeepSeek-R1-Zero的输出，然后通过人工标注进行后处理来提炼结果

有了冷启动数据之后便开始微调 DeepSeek-V3-Base 模型。结果显示可读性大大增强。而且相对于DeepSeek-R1-Zero， 使用人类设计标注的冷启动数据的训练模式性能更好。

Reasoning -oriented Reinforcement Learning（面向推理的强化学习）

为了进一步增强它在推理密集型任务中推理能力。比如编程，逻辑推理等任务（这些任务有定义明确的问题和解决方案）。参考DeepSeek-R1-Zero训练过程，对微调后的DeepSeek-V3-Base模型进行了大规模的强化学习训练，

同样，在实验中发现，在CoT过程中已经出现了语言混合的问题，尤其是涉及到多种语言时更容易出错。为了缓解语言混合问题，训练期间引入了语言一致性奖励 (language consistency reward)，其计算方式是目标语言单词在 CoT 中的比例。虽然这种语言对齐会导致性能略有下降，但这种奖励与人类的偏好一致，使其更具可读性。

最后，将推理任务的准确性 (the accuracy of reasoning tasks) 和语言一致性的奖励结合起来，直接相加形成最终的奖励。然后，我们在微调模型上进行强化学习训练，直到它在推理任务上实现收敛。

Rejection Sampling and Supervised Fine-Tuning（抑制采样和监督微调）

在有了一定推理能力之后，为了增加其在写作，角色扮演和其他通用任务的能力。利用结果检查点 (resulting checkpoint) 来收集SFT (Supervised Fine-Tuning) 数据来用于下一轮训练以增强其通用能力。

内容来源：csdn.net

作者名称：wangben

原文链接：https://blog.csdn.net/weixin_42754905/article/details/145648999
作者主页：https://blog.csdn.net/weixin_42754905

为了增量通用能力，收集了两类数据：Reasoning data -- 推理数据，Non-Reasoning data 非推理数据。

为了获取Reasoning data -- 推理数据，我们通过在结果检查点（resulting checkpoint）通过抑制采样（rejection sampling）来获取推理提示（reasoning prompts）和推理轨迹（reasoning trajectories）。其中一些数据通过使用生成奖励模型，将真实数据和模型预测输入到DeepSeek-V3中进行判断是否合格。

经过滤之后。我们总共收集了大约600k个推理相关的训练样本。

对于非推理数据，如写作、翻译，我们采用DeepSeek-V3管道，并复用DeepSeek-V3的SFT数据集的一部分。我们总共收集了大约200k个与推理无关的训练样本。

我们使用上述约800k样本的精选数据集对DeepSeek-V3-Base进行了两个epoch的微调。

Reinforcement Learning for all Scenarios (适用于所有场景的强化学习)

为了进一步使模型与人类偏好保持一致，保证模型的有用性和无害性，进一步完善模型的推理能力。我们实现了一个二级强化学习阶段。

使用奖励信号（reward signal）和多种提示分布（Diverse Prompt Distributions）的组合来训练模型。对于推理数据，跟DeepSeek-R1-Zero一样，利用基于规则的奖励来执行推理过程。

DeepSeek-R1 Evaluation (DeepSeek-R1评估)

内容来源: csdn.net

作者昵称: hdubigben

原文链接: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

Benchmark (Metric)		Claude-3.5-Sonnet-1022	GPT-4o-0513	DeepSeek-V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
English	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
	MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	92.9
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	84.0
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0	30.1
Code	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	87.6
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	92.3
	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	65.9
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6	96.3
Math	Codeforces (Rating)	717	759	1134	1820	2061	2029
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7	53.3
	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	79.8
Chinese	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	78.8
	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
C-Eval (EM)	C-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-	63.7

作者主页：https://blog.csdn.net/weixin_42754905/article/details/145648995

内容来源：codeforces
作者昵称：hdzhang
原文链接：https://blog.csdn.net/weixin_42754905/article/details/145648999

Table 4 | Comparison between DeepSeek-R1 and other representative models

DeepSeek-R1 表现出优于 DeepSeek-V3 的性能。这种改进主要归因于 STEM 相关问题的准确性提高，其中通过大规模强化学习实现了显著的收益。

此外，DeepSeek-R1 在 FRAMES 方面表现出色，这是一项长期依赖上下文的 QA 任务，展示了其强大的文档分析能力。这凸显了推理模型在 AI 驱动的搜索和数据分析任务中的潜力。在 factual benchmark SimpleQA，DeepSeek-R1 的性能优于 DeepSeek-V3，展示了其处理基于事实的查询的能力。

在此基准上，OpenAI-o1 超过 GPT-4o 也观察到类似的趋势。然而，DeepSeek-R1 在 Chinese SimpleQA 基准测试中的表现比 DeepSeek-V3 差，主要是因为它倾向于在安全 RL 之后拒绝回答某些查询。在没有安全 RL 的情况下，DeepSeek-R1 可以达到超过 70% 的准确率。

Distillation: Empower Small Models with Reasoning Capability (蒸馏：为小模型赋予推理能力)

此外也将收集到的 800k 样本直接对 Qwen 和 Llama 等模型进行了微调，这种简答的蒸馏方式显著增强了模型的推理能力。蒸馏模型评估结果如下，蒸馏法具有强大的潜力

内容来源: csdn.net

作者昵称: hdubigben

原文链接: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

CSDN @hdubigben

在蒸馏过程中，只使用了SFT (Supervised Fine-Tuning) 数据进行微调。为了验证大规模 RL 训练实现与蒸馏相当的性能，在这一节又进行了一步实验，在Qwen-32B模型上进行大规模的0样本强化训练，结果显示，不如使用SFT (Supervised Fine-Tuning) 数据进行微调的结果。

内容来源: csdn.net

作者昵称: hdubigben

原文链接: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

CSDN @hdubigben

总结和展望

作者在文中的结论有一定的余地，得出来两个结论。

首先将更强大的模型蒸馏成更小的模型会产生极好的结果，但小模型进行依赖于本大规模 RL 训练需要巨大的计算能力，甚至可能无法达到蒸馏的性能。

其次，虽然蒸馏策略既经济又有效，但超越intelligence界限可能仍然需要更强大的基础模型和更大规模的强化学习。

对未来有以下三点展望。

- 通用能力：目前，DeepSeek-R1 在函数调用、多轮次、复杂角色扮演和 JSON 输出等任务方面的能力不如 DeepSeek-V3。展望未来，我们计划探索可以利用 CoT 来增强这些领域的任务多长时间。
- 语言混合：DeepSeek-R1 目前针对中文和英文进行了优化，这可能会导致在处理其他语言的查询时出现语言混合问题。例如，DeepSeek-R1 可能会使用英语进行推理和响应，即使查询使用的是英语或中文以外的语言。我们的目标是在将来的更新中解决此限制。
- 提示工程：在评估 DeepSeek-R1 时，我们观察到它对提示很敏感。Few-shot 提示始终会降低其性能。因此，我们建议用户直接描述问题并使用zero-shot设置指定输出格式以获得最佳结果。

作者昵称: hdubigben

原文地址: https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页: https://blog.csdn.net/weixin_42754905

ps:

膜拜大佬。DeepSeek-R1虽然不是第一个推理大模型，但R1的出现第一次让推理模型能被大多数人用上，而且是免费用上，并且进行了开源。此外，蒸馏小模型的尝试，也提供了一种新的思路--通过少样本在通用大模型基础上进行微调，同样可以实现具有推理能力的推理模型。这对于企业训练微调自己的推理模型，提供了技术可行性和经济可行性。

内容来源：csdn.net

作者昵称：hdubigben

原文链接：https://blog.csdn.net/weixin_42754905/article/details/145648999

作者主页：https://blog.csdn.net/weixin_42754905