

DeepSeek-R1 论文解读

原创 just do it now 已于 2025-01-21 20:42:46 修改 阅读量1.3w 收藏 34 点赞数 29

文章标签： 人工智能 自然语言处理 深度学习

松山湖开发者村综合... 文章已被社区收录

<https://huggingface.co/deepseek-ai/DeepSeek-R1>

标题: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

作者: DeepSeek-AI

摘要: 本文介绍了第一代推理模型DeepSeek-R1-Zero和DeepSeek-R1。DeepSeek-R1-Zero是通过大规模 **强化学习** (RL) 训练的模型, 无需监督微调 (SFT) 作为初步步骤, 展示了卓越的推理能力。然而, 它面临可读性差和语言混合等挑战。为了解决这些问题并进一步提升推理性能, 作者引入了DeepSeek-R1, 该模型在RL之前结合了多阶段训练和冷启动数据。DeepSeek-R1在推理任务上的表现与OpenAI的o1-1217相当。为了支持研究社区, 作者开源了DeepSeek-R1-Zero、DeepSeek-R1以及基于Qwen和Llama的六个密集模型 (1.5B、7B、8B、14B、32B、70B)。

本文的核心贡献包括:

1. **后训练**: 直接在基础模型上应用大规模强化学习, 开发了DeepSeek-R1-Zero, 展示了LLMs通过纯 **RL** 自我进化的潜力。
2. **蒸馏**: 展示了将大模型的推理模式蒸馏到小模型中的有效性, 显著提升了小模型的推理能力。

一、方法

1.1 概述

作者展示了通过大规模强化学习 (RL) 可以显著提升模型的推理能力, 即使不使用监督 **微调** (SFT) 作为冷启动。进一步地, 加入少量冷启动数据提升性能。本文介绍了以下内容:

1. **DeepSeek-R1-Zero**: 直接在基础模型上应用RL, 无需任何SFT数据。
2. **DeepSeek-R1**: 从经过数千个长链推理 (CoT) 示例微调的检查点开始应用RL。
3. **蒸馏**: 将DeepSeek-R1的推理能力蒸馏到小型密集模型中。

1.2 DeepSeek-R1-Zero: 在基础模型上进行强化学习

DeepSeek-R1-Zero通过纯RL过程展示了强大的推理能力。作者采用了Group Relative Policy Optimization (GRPO) 算法, 通过从旧策略中采样一组基线, 从而优化策略模型。奖励模型主要由准确性奖励和格式奖励组成, 确保模型在推理过程中生成正确的答案并遵循指定的格式。

DeepSeek-R1-Zero 的构建过程

DeepSeek-R1-Zero 是一个通过**纯强化学习** (Reinforcement Learning, RL) **训练的模型**, **无需监督微调** (Supervised Fine-Tuning, SFT) **作为步骤**。它的构建过程主要围绕如何通过RL激励模型自我进化, 从而提升推理能力。以下是DeepSeek-R1-Zero构建过程的详细介绍:

1. 基础模型选择

DeepSeek-R1-Zero 的基础模型是 **DeepSeek-V3-Base**。这个模型是一个预训练的大型语言模型 (LLM), 具备较强的语言理解和生成能力。选择这一点, 是因为它已经具备了处理复杂任务的基础能力, 适合通过RL进一步优化推理能力。

2. 强化学习框架

DeepSeek-R1-Zero 使用了 **Group Relative Policy Optimization (GRPO)** 作为强化学习框架。GRPO 是一种高效的RL算法, 旨在减少训练成本, 提升性能。GRPO 的核心思想是通过**组内相对奖励**来优化策略模型, 而不是依赖传统的批评模型 (critic model)。

• GRPO 的优化目标:

GRPO 通过从旧策略中采样一组输出, 计算每个输出的相对优势 (advantage), 然后优化策略模型以最大化这些优势。具体公式如下:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 \right) \right)$$

其中， A_i 是优势函数，通过组内奖励的均值和标准差计算得出。

• 优势：

GRPO 避免了传统RL中需要训练批评模型的复杂性，显著降低了计算成本，同时保持了RL的有效性。

关于GRPO算法更详细的理解，可以移步[DeepSeek-V3-GRPO理解-CSDN博客](#)

3. 奖励模型设计

奖励模型是RL训练中的关键部分，它决定了模型优化的方向。DeepSeek-R1-Zero 的奖励模型主要由两部分组成：

1. 准确性奖励 (Accuracy Reward)：

- 用于评估模型生成的答案是否正确。
- 对于数学问题等有确定性答案的任务，模型需要以特定格式（如框内答案）提供最终答案，以便通过规则进行验证。
- 对于编程问题（如LeetCode），使用编译器根据预定义的测试用例生成反馈。

2. 格式奖励 (Format Reward)：

- 强制模型在生成答案时遵循指定的格式，例如将推理过程放在 `<think>` 和 `</think>` 标签之间，答案放在 `<answer>` 和 `</answer>` 标签之间。
- 这种格式奖励确保了模型输出的结构化和可读性。

• 不使用神经奖励模型：

文中提到，作者没有使用神经奖励模型的原因主要有以下几点：

• 奖励攻击 (Reward Hacking)：

- 在大规模RL训练中，神经奖励模型容易受到奖励攻击的影响，即模型可能会通过生成符合奖励模型偏好但不正确的答案来“欺骗”奖励系

• 训练复杂性：

- 神经奖励模型需要额外的训练资源，并且重新训练奖励模型会增加整个训练流程的复杂性。

• 简化流程：

- 基于规则的奖励系统更加简单直接，能够有效引导模型生成正确的答案和符合格式的输出生，同时减少训练中的不确定性。

4. 训练模板

为了引导模型在RL过程中生成符合要求的输出，作者设计了一个简单的训练模板。模板要求模型首先生成推理过程，然后生成最终答案。模板如下：

```
1 | A conversation between User and Assistant. The user asks a question, and the Assistant solves it.
2 | The assistant first thinks about the reasoning process in the mind and then provides the user with the answer.
3 | The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively,
4 | i.e., <think> reasoning process here </think> <answer> answer here </answer>.
5 | User: prompt. Assistant:
```

这个模板确保了模型在生成答案时遵循指定的格式，同时避免了内容上的偏见，使得模型能够在RL过程中自然地进化。

5. 自我进化过程

DeepSeek-R1-Zero 的训练过程展示了模型如何通过RL自我进化，逐步提升推理能力。以下是其自我进化过程的关键点：

1. 初始性能：

- 在RL训练开始时，DeepSeek-R1-Zero 的表现相对较弱。例如，在AIME 2024基准测试中，初始的pass@1得分仅为15.6%。

2. 逐步提升：

- 随着RL训练的进行，模型的推理能力逐步提升。经过数千个RL步骤后，DeepSeek-R1-Zero 在AIME 2024上的pass@1得分提升至71.0% OpenAI的o1-0912模型的表现。
- 通过多数投票 (majority voting) ，



just do it now

关注

👍 29



3. 复杂行为的涌现：

1. 在RL过程中，模型自发地发展出一些复杂的推理行为，如**自我验证**、**反思**和**生成长链推理过程**。
2. 这些行为并非通过显式编程实现，而是模型在RL环境中自我探索和优化的结果。

4. “Aha Moment”：

1. 在训练过程中，模型出现了“Aha Moment”，即模型在解决某些问题时，突然意识到需要重新评估其初始推理步骤。这种行为展示了RL在自主学习和改进方面的潜力。

6. 挑战与局限性

尽管DeepSeek-R1-Zero 展示了强大的推理能力，但它也面临一些挑战：

1. 可读性差：

1. 由于模型是通过纯RL训练的，生成的推理过程往往难以阅读，缺乏清晰的结构和格式。

2. 语言混合：

1. 在生成推理过程时，模型有时会混合使用多种语言（如中英文混合），导致输出不够一致。

3. 输出格式不一致：

1. 尽管有格式奖励，模型在某些情况下仍然会生成不符合要求的输出格式。

7. 总结

DeepSeek-R1-Zero 的构建过程展示了如何通过**纯强化学习**激励大型语言模型自我进化，从而提升推理能力。尽管它面临可读性和语言混合等挑战，推理任务上的表现证明了RL在提升模型推理能力方面的巨大潜力。DeepSeek-R1-Zero 的成功为未来的研究提供了新的方向，尤其是在无需监督数据如何通过RL进一步优化模型的推理能力。

通过DeepSeek-R1-Zero，作者验证了LLMs可以通过纯RL自我进化，生成复杂的推理行为，这为后续的DeepSeek-R1模型奠定了基础。

1.3 DeepSeek-R1：带冷启动的强化学习

DeepSeek-R1在RL之前引入了冷启动数据和多阶段训练管道。首先，作者收集了数千个冷启动数据来微调DeepSeek-V3-Base模型，然后进行推理与训练。在RL接近收敛时，通过拒绝采样生成新的SFT数据，并结合来自DeepSeek-V3的监督数据进行微调。最后，模型在所有场景的提示下进行额外的训练，得到DeepSeek-R1。



DeepSeek-R1 是在 DeepSeek-R1-Zero 的基础上进一步优化的模型，旨在解决 DeepSeek-R1-Zero 面临的**可读性差**和**语言混合**等问题，同时进一步提升推理能力。DeepSeek-R1 通过引入**冷启动数据**和**多阶段训练管道**，显著提升了模型的表现。以下是 DeepSeek-R1 的详细介绍：

1. 背景与动机

DeepSeek-R1-Zero 展示了通过纯强化学习（RL）激励模型自我进化的潜力，但其生成的推理过程可读性较差，且存在语言混合等问题。为了解决这进一步提升模型的推理能力，作者设计了 DeepSeek-R1。DeepSeek-R1 的核心思想是通过**冷启动数据**和多阶段训练，引导模型生成更清晰、更一致的推理过程，同时保持强大的推理能力。

2. DeepSeek-R1 的训练管道

DeepSeek-R1 的训练管道分为四个主要阶段：

1. 冷启动（Cold Start）：

- 在 RL 训练之前，DeepSeek-R1 首先通过**冷启动数据**对基础模型进行微调。冷启动数据由数千个长链推理（Chain-of-Thought, CoT）示例组成，通过多种方式生成：
 - 使用少样本提示（few-shot prompting）生成详细的推理过程。
 - 直接提示模型生成带有反思和验证的详细答案。
 - 从 DeepSeek-R1-Zero 的输出中筛选高质量的推理过程。



just do it now

关注

觉得

👍 29

- 冷启动数据的引入显著提升了模型的可读性和推理能力，避免了 RL 训练初期的**不稳定冷启动阶段**。

2. 推理导向的强化学习 (Reasoning-oriented RL) :

- 在冷启动微调之后，DeepSeek-R1 进入推理导向的 RL 训练阶段。这一阶段的训练过程与 DeepSeek-R1-Zero 类似，但引入了**语言一致性奖励**，言混合问题。
- **语言一致性奖励**：通过计算推理过程中目标语言单词的比例，确保模型生成的推理过程语言一致。
- 尽管语言一致性奖励略微降低了模型的性能，但它显著提升了输出的可读性，符合人类偏好。

3. 拒绝采样与监督微调 (Rejection Sampling and Supervised Fine-Tuning, SFT) :

- 在推理导向的 RL 训练接近收敛时，作者通过**拒绝采样**从 RL 检查点生成新的 SFT 数据。
- **推理数据**：从 RL 检查点生成推理轨迹，并通过规则奖励和生成奖励模型（使用 DeepSeek-V3 进行判断）筛选出正确的推理过程。
- **非推理数据**：包括写作、事实问答、自我认知和翻译等任务的数据，这些数据来自 DeepSeek-V3 的 SFT 数据集。
- 最终，作者收集了约 80 万条训练样本（60 万条推理数据和 20 万条非推理数据），并对 DeepSeek-V3-Base 模型进行了两轮微调。

4. 全场景强化学习 (Reinforcement Learning for All Scenarios) :

- 为了进一步对齐人类偏好，DeepSeek-R1 进行了第二阶段的 RL 训练，旨在提升模型的**有用性和无害性**，同时保持其推理能力。
- 对于推理数据，继续使用规则奖励来指导数学、代码和逻辑推理任务的学习。
- 对于通用数据，使用奖励模型来捕捉人类偏好，确保生成的响应既有用又无害。

3. DeepSeek-R1 的性能

DeepSeek-R1 在多个基准测试中表现出色，尤其在推理任务上表现优异。以下是其主要性能亮点：

1. 推理任务：

- 在 AIME 2024 上，DeepSeek-R1 的 pass@1 得分为 79.8%，略高于 OpenAI 的 o1-1217。
- 在 MATH-500 上，DeepSeek-R1 的 pass@1 得分为 97.3%，与 OpenAI-o1-1217 相当，显著优于其他模型。
- 在 Codeforces 上，DeepSeek-R1 的 Elo 评分为 2029，超过了 96.3% 的人类参赛者。

1. 知识任务：

在 MMLU、MMLU-Pro 和 GPQA Diamond 等知识基准测试中，DeepSeek-R1 的表现显著优于 DeepSeek-V3，尽管略低于 OpenAI-o1-1217，他闭源模型。

2. 其他任务：

DeepSeek-R1 在创意写作、问答、编辑和摘要等任务上也表现出色。例如，在 AlpacaEval 2.0 上，DeepSeek-R1 的长度控制胜率为 87.1%；在 ArenaHard 上的胜率为 92.3%。

4. DeepSeek-R1 的挑战与未来工作

尽管 DeepSeek-R1 在多个任务上表现出色，但仍面临一些挑战：

1. 通用能力不足：

在函数调用、多轮对话、复杂角色扮演和 JSON 输出等任务上，DeepSeek-R1 的表现不如 DeepSeek-V3。

2. 语言混合问题：

DeepSeek-R1 主要针对中文和英文优化，处理其他语言时可能出现语言混合问题。

3. 提示敏感性：

DeepSeek-R1 对提示 (prompt) 较为敏感，少样本提示 (few-shot prompting) 会降低其性能，建议使用零样本提示 (zero-shot prompting) 。

未来的工作将集中在提升通用能力、解决语言混合问题、优化提示工程以及扩展软件工程任务的 RL 训练。

6. 总结

DeepSeek-R1 通过引入冷启动数据和多阶段训练策略，显著提升了模型的推理能力和可读性。它在多个推理任务上的表现与 OpenAI 的 o1-1217 相当，在某些任务上超越了它。通过蒸馏技术，DeepSeek-R1-Distill-Lite 模型在保持推理能力的同时，大幅降低了推理成本。这些成果为构建更强大、更高效的推理模型提供了新的方向，尤其是在如何通过 RL 训练提升模型性能方面。  [关注](#)  29 

1.4 蒸馏：赋予小模型推理能力

作者探索了将DeepSeek-R1的推理能力蒸馏到小型密集模型中。使用Qwen2.5-32B作为基础模型，直接蒸馏DeepSeek-R1的表现优于在其上应用RL，14B模型在多个基准测试中显著优于现有的开源模型。

在论文中，**蒸馏（Distillation）**是将 DeepSeek-R1 的推理能力迁移到更小的模型上的关键步骤。通过蒸馏，作者成功地将 DeepSeek-R1 的强大推理能力传递给了多个小型密集模型（如 1.5B、7B、14B、32B 和 70B 模型），这些模型在多个基准测试中表现优异，甚至超越了现有的开源模型。以下是蒸馏的详细介绍：

1. 蒸馏的背景与动机

大型语言模型（如 DeepSeek-R1）在推理任务上表现出色，但其庞大的参数量和计算需求限制了其在资源受限环境中的应用。为了在保持高性能的同时降低推理成本，作者探索了将 DeepSeek-R1 的推理能力蒸馏到更小的模型上。蒸馏的核心思想是通过**知识迁移**，将大模型的复杂推理模式传递给小模型，提升小模型的推理能力。

2. 蒸馏的数据来源

蒸馏过程使用了 DeepSeek-R1 生成的 **80 万条训练数据**，这些数据包括：

- **60 万条推理数据**：通过拒绝采样从 DeepSeek-R1 的 RL 检查点生成，涵盖了数学、代码、科学和逻辑推理等任务。
- **20 万条非推理数据**：包括写作、事实问答、自我认知和翻译等任务的数据，来自 DeepSeek-V3 的 SFT 数据集。

这些数据确保了蒸馏后的模型不仅具备强大的推理能力，还能处理多种通用任务。

3. 蒸馏的目标模型

作者选择了多个开源模型作为蒸馏的目标模型，包括：

- **Qwen 系列**：Qwen2.5-1.5B、Qwen2.5-7B、Qwen2.5-14B、Qwen2.5-32B。
- **Llama 系列**：Llama-3.1-8B、Llama-3.3-70B-Instruct。

这些模型在蒸馏前已经具备一定的语言理解和生成能力，但推理能力相对较弱。通过蒸馏 DeepSeek-R1 的输出，这些模型的推理能力得到了显著提升。

4. 蒸馏的训练过程

蒸馏过程主要包括以下步骤：



1. 数据生成：

- 使用 DeepSeek-R1 生成 80 万条训练数据，涵盖推理和非推理任务。
- 对于推理任务，通过拒绝采样筛选出正确的推理过程。
- 对于非推理任务，使用 DeepSeek-V3 的 SFT 数据集。

1. 监督微调（SFT）：

- 使用生成的 80 万条数据对目标模型进行监督微调。
- 微调过程中，模型学习 DeepSeek-R1 的推理模式和生成风格。
- 作者对每个目标模型进行了两轮微调，以确保知识迁移的效果。

1. 蒸馏后的模型评估：

- 蒸馏后的模型在多个基准测试上进行了评估，包括 AIME 2024、MATH-500、GPQA Diamond、LiveCodeBench 和 Codeforces 等。
- 评估结果表明，蒸馏后的模型在推理任务上的表现显著优于蒸馏前的模型，甚至超越了现有的开源模型。

5. 蒸馏的效果

蒸馏后的模型在多个基准测试中表现出色，以



just do it now

关注

29



Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	Code
	pass@1	cons@64	pass@1	pass@1	pass@1	rat
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	7
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	7
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	18
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	13
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	9
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	11
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	14
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	16
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	12
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	16

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

从表中可以看出，蒸馏后的模型在多个基准测试上的表现显著优于蒸馏前的模型，尤其是 DeepSeek-R1-Distill-Qwen-32B 和 DeepSeek-R1-Distill-Llama-70B 其表现甚至超越了 OpenAI 的 o1-mini 模型。

6. 蒸馏的优势

蒸馏技术具有以下优势：

1. 计算效率高：

蒸馏后的模型参数量显著减少，计算成本大幅降低，适合在资源受限的环境中部署。

2. 性能优异：

蒸馏后的模型在推理任务上的表现接近甚至超越大模型，展示了知识迁移的有效性。

3. 易于扩展：

蒸馏技术可以应用于多种模型架构和任务，具有广泛的适用性。

7. 蒸馏的局限性

尽管蒸馏技术具有显著优势，但也存在一些局限性：

1. 依赖大模型：

蒸馏的效果依赖于大模型的质量和生成的数据，如果大模型本身表现不佳，蒸馏后的模型性能也会受限。

2. 知识损失：

在蒸馏过程中，部分复杂的推理模式可能无法完全传递给小模型，导致性能略有下降。

8. 未来工作

作者计划在未来的研究中进一步探索以下方向：

1. 结合RL训练：

在蒸馏的基础上加入RL训练，进一步提升小模型的推理能力。

2. 多任务蒸馏：

探索如何将大模型在多个任务上的知识同时蒸馏给小模型，提升其通用能力。

3. 自动化蒸馏：

开发自动化蒸馏框架，减少人工干预，提升效率。

觉得



just do it now

关注

29

