大模型（LLMs）强化学习面

来自： AiGC面试宝典

宁静致远

2024年01月27日 20:47

1 简单介绍强化学习?

强化学习：（Reinforcement Learning）一种机器学习的方法，**通过从外部获得激励来校正学习方向从而获得一种自适应的学习能力。**

2 简单介绍一下 RLHF?

基于人工反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）：**构建人类反馈数据集，训练一个激励模型，模仿人类偏好对结果打分**，这是GPT-3后时代大语言模型越来越像人类对话核心技术。

3. 奖励模型需要和基础模型一致吗?

不同实现方式似乎限制不同。（待实践确认）colossal-ai的coati中需要模型有相同的tokenizer，所以选模型只能从同系列中找。在ppo算法实现方式上据说trlx是最符合论文的。

4. RLHF 在实践过程中存在哪些不足?

1. 不足点1：人工产生的偏好数据集成本较高，很难量产；
2. 不足点2：三个阶段的训练（SFT->RM->PPO）过程较长，更新迭代较慢；
3. 不足点3：PPO 的训练过程同时存在4个模型（2训练，2推理），对计算资源的要求较高。

5. 如何解决 人工产生的偏好数据集成本较高，很难量产问题?

- 解决方法：AI 专家替代派
- 代表方法：

1. RLAIF

该方法的**核心在于通过AI 模型监督其他 AI 模型，即在SFT阶段，从初始模型中采样，然后生成自我批评和修正，然后根据修正后的反应微调原始模型。** 在 RL 阶段，从微调模型中采样，使用一个模型来评估生成的样本，

并从这个 AI 偏好数据集训练一个偏好模型。 然后使用偏好模型作为奖励信号对 RL 进行训练，即 RL from AI Feedback（RLAIF）。

1. RRHF

RRHF(Rank Response from Human Feedback) 不需要强化学习，可以利用不同语言模型生成的回复，包括 ChatGPT、GPT-4 或当前的训练模型。RRHF通过对回复进行评分，并通过排名损失来使回复与人类偏好对齐。RRHF 通过通过排名损失使评分与人类的偏好（或者代理的奖励模型）对齐。RRHF 训练好的模型可以同时作为生成语言模型和奖励模型使用

6. 如何解决三个阶段的训练（SFT->RM->PPO）过程较长，更新迭代较慢问题?

- 解决方法：微调数据优化派
    - 方法介绍：该类方法的核心在于仅仅通过优质数据集的获取和产生，以训练得到一个效果较好的 SFT 模型，而无需进行 RM 和 PPO 的训练。
- 代表方法：

1. LIMA

LIMA(Less Is More for Alignment) 即浅层对齐假说，即一个模型的知识和能力几乎完全是在预训练中学习的，而对齐则是教会它与用户交互时如何选择子分布。如果假说正确，对齐主要有关于学习方式，那么该假说的一个推论是，人们可以用相当少的样本充分调整预训练的语言模型。因此，该工作假设，对齐可以是一个简单的过程，模型学习与用户互动的风格或格式，以揭示在预训练中已经获得的知识和能力。

1. MAYBE ONLY 0.5% DATA IS NEEDED

本文主要从数据角度来探讨如何降低 LLM 训练阶段的成本，提高数据效率。为了实现该目的，作者通过从现有数据中识别出最有价值的核心样本来帮助模型获取下游任务的知识，并仅用少量数据来实现可比甚至更好的性能。

7. 如何解决 PPO 的训练过程同时存在4个模型（2训练，2推理），对计算资源的要求较高 问题?

- 解决方法：训练过程改造派
    - 方法介绍：该类方法通常通过改造模型的训练方式（如只保留SFT和RM），以提高训练效率并减少训练成本。
- 代表方法：

1. RAFT

RAFT（Reward rAnked FineTuning），它基于关于通过奖励和监督微调对样本进行排序的组合的形式。

1. DPO

DPO(Direct Preference Optimization) 提出了一种使用二进制交叉熵目标来精确优化LLM的方法，以替代基于 RLHF 的优化目标，从而大大简化偏好学习 pipeline。