0.1捕鱼送6元

EMCm7DuGMf9lBRLV

0.1捕鱼送6元扩散LLM推理用类GRPO学习！优于单独SFT，UCLA、Meta新框架d1开源

机器之心报道

编辑：陈陈、杜伟

当前，强化学习（RL）方法在最近模型的推理任务上取得了显著的改进，比如 DeepSeek-R1、Kimi K1.5，显示了将 RL 直接用于基础模型可以取得媲美 OpenAI o1 的性能。

不过，基于 RL 的后训练进展主要受限于自回归的大语言模型（LLM），它们通过从左到右的序列推理来运行。

与此同时，离散扩散大语言模型（dLLM）成为有潜力的语言建模的非自回归替代。不像以因果方式逐 token 生成文本的自回归模型那样，dLLM 通过迭代去噪过程生成文本，在多步骤操作中优化序列的同时并通过双向注意力利用过去和未来的上下文。其中，LLaDA 等开放的掩码 dLLM 实现了媲美同尺寸自回归模型的性能，而 Mercury 等闭源 dLLM 进一步展现了出色的推理延迟。

然而，顶级的开源 dLLM 并没有使用 RL 后训练，使得这一有潜力的研究方向还有很大的挖掘空间。这一范式转变引出了重要的问题：RL 后训练如何在非自回归上下文中高效地实现？

RL 算法适应掩码 dLLM 面临一些独特的挑战，原因在于自回归模型采用的已有方法（如 PPO、GRPO）通过计算生成序列的对数概率来估计和优化策略分布，导致无法直接应用于 dLLM。虽然这种计算在自回归模型中通过序列因式分解很容易实现，但 dLLM 由于它们的迭代、非序列生成过程而缺乏这种自然分解。

为了解决这些问题，来自 UCLA 和 Meta AI 的研究者提出了一个两阶段后训练框架 d1，从而可以在掩码 dLLM 中进行推理。在第一阶段，模型在高质量推理轨迹中进行监督微调；在第二即 RL 阶段，研究者引入了用于掩码 dLLM 的新颖策略梯度方法 diffu-GRPO，它利用提出的高效一步（one-step）对数概率估计在 GRPO 的基础上创建。

研究者表示，他们的估计器利用了随机提示词掩码，作为策略优化的一种正则化，使得可以扩展 per batch 的梯度更新数量并减少 RL 训练所需的在线生成数量。这将极大地降低计算时间。

在实验部分，研究者使用 LLaDA-8B-Instruct 作为基础模型实例化 d1。他们将 d1-LLaDA 的性能与基础 LLaDA 模型以及仅使用 SFT 和仅使用 diffu-GRPO 训练的 LLaDA 模型进行比较。结果表明，d1 在四个数学和逻辑推理基准测试中始终优于基础模型，如下图 1 所示。d1-LLaDA 同样优于仅使用 SFT 方法和仅使用 diffu-GRPO 方法的模型。

方法概览

d1 是一个两阶段框架，通过依次结合监督微调（SFT）和在线强化学习（RL）来增强预训练掩码 dLLMs 的推理性能。

其中，在线强化学习（特别是 GRPO 算法）已被证明能有效提升离线训练语言模型的性能。然而，GRPO 的学习策略并不能直接泛化到 dLLMs。

GRPO 的目标函数（如公式 3 所示）需要同时计算当前策略 π_θ 和旧策略 π_θold 在以下两个层面的（对数）似然比：

核心问题在于：研究者需要高效计算 dLLMs 生成内容的逐 token 对数概率和序列对数概率。

自回归（AR）模型，如 Transformer，直接对每个 token 的对数概率进行建模，并且可以通过链式法则使用一次前向传递轻松计算出序列级别的对数概率

同样，KL 项可以分解为。

与 AR 模型不同，dLLMs 不遵循序列对数概率的顺序分解。同时，每个 token 的对数概率计算成本也很高，因为解码过程中需要多次调用掩码预测器 f_θ。基于此，该研究提出了一个高效的对数概率估计器。

对于序列对数概率，该研究使用均场近似方法，将其分解为独立的每个 token 对数概率的乘积。

对于每个 token 的对数概率，该研究引入了一种估计方法，该方法仅调用一次 f_θ。

基于新引入的对数概率估计器，该研究将 GRPO 扩展到掩码 dLLMs，推导出 diffu-GRPO 的损失函数。

算法如下图所示。

实验结果

表 1 报告了基线模型 LLaDA-8B-Instruct 与采用不同后训练优化方案的模型，在四项任务上的零样本性能对比。

图 3 绘制了有效 token 的平均数量：

基于实验，该研究得出以下主要发现：

diffu-GRPO 在所有 12 种设置中都一致优于基础的 LLaDA 和 SFT（监督式微调）。diffu-GRPO 和 SFT 都相较于 LLaDA-8B-Instruct 基线有所提升，但 diffu-GRPO 显示出更持续且幅度更大的增益。具体来说，diffu-GRPO 在所有 12 种设置中都优于 LLaDA-8B-Instruct 和 SFT，而 SFT 仅在其中的 7 种设置中优于 LLaDA-8B-Instruct，这表明diffu-GRPO 相比于单独的 SFT 实现了更强的整体性能提升。

LLaDA+diffu-GRPO 在所有设置中都优于基础的 LLaDA-8B-Instruct 模型，而 d1-LLaDA 在每种情况下都超过了 LLaDA+SFT。这表明，无论初始化是来自预训练模型还是经过 SFT 调整的检查点，diffu-GRPO 都能提供可靠的性能提升。

d1 训练方案实现了最显著的性能提升。通过先进行监督微调（SFT）、再结合 diffu-GRPO 训练所形成的 d1-LLaDA 模型，产生了超越单一方法的叠加增益。这种组合式方法在 12 个实验设置中有 11 项优于纯 diffu-GRPO 方案，表明两个训练阶段存在协同效应。

定性结果表明，在 SFT 和 d1-LLaDA 生成中出现了顿悟时刻。尽管与 LLaDA-8B-Instruct 相比，生成序列长度为 128 和 256 的性能随着 SFT、diffu-GRPO 和 d1 有所提高，但从质的方面看，在生成的推理轨

迹中并未观察到显著差异。然而当序列长度达到 512 时，该研究开始观察到 SFT 和 d1-LLaDA 模型展现出两种关键能力：自我修正机制和回溯行为。

澳洲幸运8预测网站

一分快三最简单的赚钱方法

澳洲幸运8番摊稳赢法

168飞艇免费计划入口

幸运5精准100%免费计划

澳洲幸运5官方开奖直播

大小单双快三软件

极速赛车10个数字规律

澳洲五开奖记录表

澳洲幸运10百度百科

澳洲幸运10漏洞公式

澳洲幸运10六码永久公式

飞艇一天赢3万公式

澳洲幸运10冠军计划规律图片

正规分分彩app下载

澳洲幸运10十句口诀

2025澳洲幸运5官网开奖直播

幸运澳洲10开奖结果

澳洲幸运10官网开奖号码