第 6 卷◆第 6 期◆版本 1.0◆2023 年
文章类型：论文|刊号（ISSN）：2630-4759 /（中图刊号）：270GL018

# 大语言模型中的上下文面板调节与奖励模型

文木源
GPT DESK PTE LTD
DOI:10.12238/ej.v6i6.1239

[摘　要] 直接偏好优化(DPO)旨在符合人类偏好,同时减少强化学习的复杂性。传统方法如人类反馈强化学习(RLHF)首先匹配奖励模型与提示和偏好,然后使用强化学习(RL)来找到最大化奖励的策略。相比之下,DPO通过直接优化策略来满足偏好,无需显式奖励函数或强化学习,简化了过程。DPO是微调语言模型以保持与人类反馈一致的更直接、更有效的方法。此外,OpenAI提到他们通过模仿人类评分来训练模型,以帮助改善RLHF。下一步是将模型拟合到含有丰富“条件”的数据集上,例如训练模型生成包含记忆、条件、目标、计划、未来任务的面板,并使用这个面板进行训练。这些条件将“创意写作任务”转变为“分配材料”的任务,减少了创意写作中的熵。条件强化学习微调(C-RLFT)使得大语言模型能够理解和生成类人文本、适应新信息和个性化响应,同时保持相关性和连贯性。未来的改进工作包括使用RLHF或RLAIF改善条件面板、数据集和模型之间的迭代、使模型与现实世界需求保持一致,以及基于0阶优化构建新的基础模型。这些方向旨在使大语言模型更高效、符合人类偏好,并能在各种环境中运行,包括边缘计算设备。
[关键词] 直接偏好优化；人类反馈强化学习；条件面板；创意写作熵降低；C-RLFT训练；边缘计算
中图分类号：TV149.2　文献标识码：A

## Contextual Panel Conditioning and Reward Models in Large Language Models

Muyuan Wen
GPT DESK PTE LTD

[Abstract] Direct preference optimization (DPO) aims to match human preferences while reducing the complexity of reinforcement learning. Traditional methods such as reinforcement learning with human feedback (RLHF) first match reward models with cues and preferences, and then use reinforcement learning (RL) to find policies that maximize rewards. In contrast, DPO simplifies the process by directly optimizing the policy to satisfy preferences without explicit reward functions or RL processes. DPO is a more direct and potentially more efficient way to fine-tune a language model to remain consistent with human feedback. Additionally, OpenAI mentioned that they trained the model by imitating human ratings to help improve RLHF. The next step is to fit the model to a data set containing rich ”conditions” .For example, the training model generates a panel containing memories, conditions, goals, plans, and future tasks, and uses this panel for training. These conditions transform the ”creative writing task” into the task of ”distributing materials”, reducing entropy in creative writing. Conditional reinforcement learning fine-tuning (C-RLFT) enables large language models to understand and generate human-like text, adapt to new information, and personalize responses while maintaining relevance and coherence. Future improvements include improving conditional panels using RLHF or RLAIF, iteration between datasets and models, aligning models with real-world needs, and building new base models based on 0-order optimization. These directions aim to make large language models more efficient, consistent with human preferences, and able to run in a variety of environments, including edge computing devices.
[Key words] Direct Preference Optimization; Human Feedback Reinforcement Learning; Conditional Panel; Creative Writing Entropy Reduction; C-RLFT Training; Edge Computing

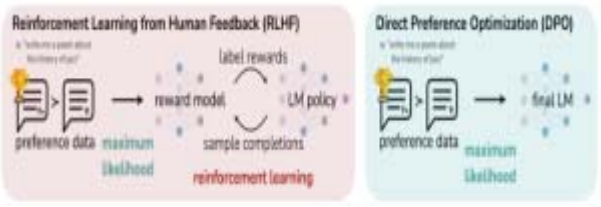
在创意写作等复杂任务中,使用详细的上下文面板调节大语言模型(LLM)能显著地降低熵。通过使用详细的上下文(例如人物和地点描述以及作者的计划和记忆)将高熵创意写作任务转换为更加结构化的“指定材料的写作”任务,模型可以以高精度和接近零损失的方式执行。

OpenAI使用了类似的技术,采用人工评分来根据人类反馈进行强化学习(RLHF)。该研究还结合了RecurrentGPT的方法,使用LLM从原始文本输入生成初始条件面板。训练包括两个部分：具有全局信息条件面板的任务以及学习从单个任务生成这些面板。

实验表明,在没有进一步强化学习的情况下,模型在创意写作任务的评估数据集上表现良好,实现了低损失和高准确率。这与在纯文本上训练的模型形成鲜明对比,后者的表现并不那么合格。

如果需要随机性或创造性,可以通过人类偏好或其他受控过程有意引入熵。这种平衡的方法可以维持大语言模型的生成能力,同时显著降低其产出的不可预测性。

## 1 大语言模型包含价值函数/奖励模型




图片翻译自：Rafailov,Rafael,etal.”Direct preference optimization: Your language model is secretly a reward model.” arXiv preprint arXiv:2305.18290(2023).

图1DPO直接优化人类偏好,而避免了强化学习。现有的方法是先用一组提示和人类对一对回应的偏好来拟合一个奖励模型,然后用强化学习来寻找一个最大化学习到的奖励的策略,以便对人类反馈进行微调语言模型。相比之下,DPO直接用一个简单的分类目标来优化最能满足偏好的策略,拟合一个隐式的奖励模型,其相应的最优策略可以以闭合形式提取出来。

该图描述了两种基于人类反馈优化语言模型的方法：人类反馈强化学习(RLHF)和直接偏好优化(DPO)。

1.1基于人类反馈的强化学习(RLHF)：

流程：涉及收集偏好数据(例如不同模型输出的人类比较)并使用它来标记奖励。

奖励模型:使用最大似然来训练奖励模型,以根据偏好数据预测这些奖励。

LM策略:在奖励模型的奖励的指导下,使用强化学习更新语言模型(LM)策略。LM生成样本补全,并对其进行评估,并且该反馈循环继续完善LM的输出。

1.2直接偏好优化(DPO)：

流程：DPO也从收集偏好数据开始,但通过直接优化语言模型以满足这些偏好而有所不同。

最终LM：DPO没有创建单独的奖励模型并使用强化学习,而是使用偏好数据的最大似然来直接训练最终的LM。没有明确的奖励函数或强化学习步骤。

目标：DPO的主要目标是满足从偏好数据确定的人类偏好,有效地将优化转化为简单的分类问题。

DPO的设计目的是为了符合人类的偏好,同时避免强化学习的复杂性。人类反馈强化学习(RLHF)等传统方法首先将奖励模型与一组提示和偏好相匹配,然后使用强化学习(RL)来找到最大化学习奖励的策略。另一方面,DPO在直接满足偏好的基础上优化策略,通过不需要显式奖励函数或强化学习来简化过程。

DPO是一种更直接、可能更有效的微调语言模型以与人类反馈保持一致的方法,因为它省去了RLHF中涉及的中间步骤。

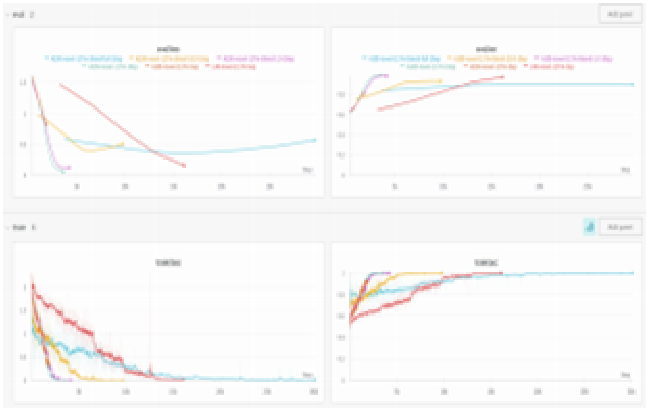
还有其他研究支持我们的发现。此外,OpenAI也提到他们通过模仿人类评分来训练模型,以期帮助改善人类反馈强化学习。

值得考虑的下一步是将模型拟合到具有大量“条件”的数据集上。我们训练模型如何生成一个包含记忆、条件、目标、计划、未来任务等信息的面板,并使用这个面板来训练模型。

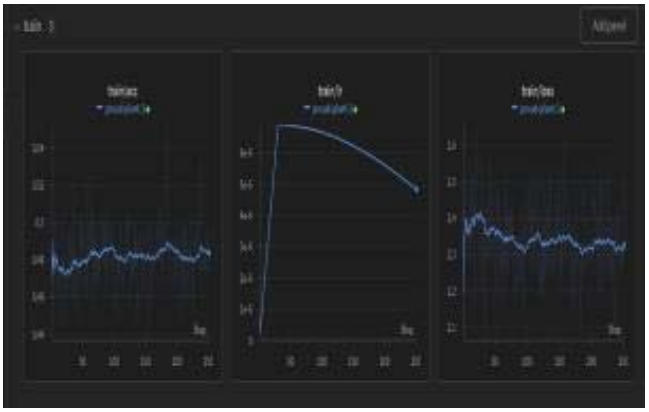
目前的实验表明,即使是具有最高熵的任务——创意写作,一旦条件良好,模型也几乎不会包含熵。我们使用的面板包括最新段落、人物列表和描述、地点列表和描述、作者的计划和目标以及记忆。这些条件使得“创意写作任务”变成了“分配材料”的任务。

条件面板由大语言模型生成,我们将原始段落放入大语言模型并通过提示获取初始条件面板,如RecurrentGPT训练数据包含两部分,一部分是来自全局信息的带有条件面板的任务,另一部分是如何从单个任务生成条件面板。

这个过程可以通过强化学习或其他优化方法来优化。幸运的是,大语言模型包含奖励模型。



如上图所示,在多种条件下,损失可以约为零,并且在评估数据集上表现良好。可以不再应用进一步的RL。我们的任务是创意写作,名称中带有“混合”的实验中使用的数据集可能会与消融研究的其他任务混合。如果我们直接在创意写作的纯文本上训练大语言模型,训练和评估损失都不会低于2.3,并且准确率将始终低于0.5。创意写作纯文本训练时的训练损失图如下：



综上所述，“条件”在大语言模型中非常有用，它可以帮助大语言模型降低熵。我们可以通过大语言模型优化原始数据集的条件面板，从而提供更好的数据质量。

如果我们需要熵，我们可以通过人类偏好或其他外部控制过程（如马尔可夫随机场）引入熵，也可以通过这些方法修改面板。

## 2 应用与未来展望

基于条件的大语言模型在下面这些事情上都很擅长：

•基于大语言模型的实时聊天系统

•创意写作

•具有上下文感知的即时消息和语音翻译

用于训练大型语言模型（LLM）的条件强化学习微调（C-RLFT）方法具有多项优势，使其对于实时聊天系统、创意写作以及上下文感知即时消息和语音等任务特别有效。

2.1基于LLM的实时聊天系统：

C-RLFT可以训练LLM高度适应用户输入，这对于实时交互至关重要。该模型可以学习理解和响应各种主题和对话风格，从而实现更加自然和引人入胜的对话。

通过强化学习，可以对模型进行微调，以满足个人用户的偏好，使交互更加个性化和上下文相关。

2.2创意写作：

C-RLFT使大语言模型对文学手法、叙事结构和角色发展有细致入微的理解。“条件”确保模型保持一致性和创造力，保持给定作品的风格和主题元素。

强化学习组件则有助于保持创意输出的连续性和连贯性，这是产生高质量创意写作的关键要素。

2.3上下文感知即时消息和语音翻译：

在即时消息和语音翻译中，维护对话的上下文至关重要。经过C-RLFT训练的模型能够更好地跟踪对话线索并适应上下文的细微差别，这对于准确且有意义的翻译至关重要。

微调过程使大语言模型能够从实时反馈中学习，从而持续改进翻译质量和上下文感知响应。

总之，C-RLFT使大语言模型能够理解和生成类人文本、适应新信息和个性化响应，同时保持需要深入理解语言和上下文的任务所需的相关性和连贯性。这使得C-RLFT训练的模型对于需要实时交互、创造力和情境感知的应用程序非常有效。

在此基础上，未来还有下列非常有意义的改进工作：

2.3.1使用RLHF或RLAIF改善条件面板：

•RLHF（人类反馈强化学习）和RLAIF（增强模仿反馈强化学习）可用于细化指导语言模型训练的条件面板.通过基于反馈机制迭代更新这些面板，可以改进数据集，使其更符合人类偏好，并更有效地指导模型的响应。

•更好的数据集：这些技术的应用将带来更高质量的数据集，这对于训练模型至关重要，该模型可以生成与人类判断和效用密切相关的回复。

2.3.2数据集和模型之间的迭代：

•该策略涉及一个循环过程，改进的数据集会产生更好的模型，而更好的模型又会生成改进的数据集。这是一种共生关系，数据集和模型通过连续迭代不断完善。

•这个迭代过程确保持续学习和适应，使模型保持最新的数据和趋势，这对于保持模型的相关性和性能至关重要。

2.3.3使模型与现实世界的需求保持一致：

•这意味着确保模型能够理解并响应现实世界任务和问题的复杂性和细微差别。

•实际应用：这不仅涉及文本训练，还涉及理解上下文、用户意图和人类交流的微妙之处，并能够将这种理解应用于医疗保健、教育、客户服务等实际应用中。

2.3.4基于0阶优化构建新的基础模型：

•使用0阶优化创建基本模型是非常具有吸引力的，这是一种不需要梯度计算的优化方法。这种方法可以显著提高内存效率并且可能更快。

•二进制模型：进行位运算的二进制模型可以显著减少所需的计算资源，使其更适合部署在内存和处理能力有限的边缘设备上。

•边缘计算：通过创建轻量级且快速的模型，可以直接在用户设备上部署大语言模型（边缘计算），从而增强隐私、减少延迟并启用离线功能。

总体而言，以上列出的这些未来方向旨在完善 创建和使用大语言模型的过程，使其更加高效、符合人类偏好，并能够在包括边缘计算设备在内的各 种环境中运行。这些进步可以显著增加大语言模型在各个领域的应用范围和影响。

## [参考文献]

[1]Zhou,Wangchunshu,etal.”RecurrentGPT:InteractiveGeneration of(Arbitrarily)Long Text.” arXiv preprint arXiv:2305.13304(2023).