可学习Beta值的DPO算法研究：自适应探索-利用平衡的新方法

作者名[1]

[1]机构, 邮箱

March 9, 2025

Abstract

本文提出了一种创新的Direct Preference Optimization (DPO)算法变体——Learnable Beta DPO，通过引入动态可学习的β参数来实现对探索-利用平衡的自适应控制。传统DPO算法使用固定的β超参数来平衡参考策略和偏好学习，这限制了其在复杂多变场景下的优化潜力。我们设计了一个与策略模型紧密耦合的BetaHead网络，能够根据输入上下文动态调整β值，从而在模型熟悉的领域保持保守学习策略，在不熟悉的领域加大探索力度。实验结果表明，Learnable Beta DPO相比固定β的标准DPO，在性能、泛化能力和样本效率方面均有显著提升。

关键词

Direct Preference Optimization, 大语言模型, 人类偏好对齐, 自适应学习, 信息融合

1 引言

近年来，大型语言模型（LLMs）的发展日新月异，而如何使这些模型更好地对齐人类偏好已成为重要研究方向。Direct Preference Optimization (DPO) [2]作为一种直接优化语言模型的算法，因其简洁高效而广受关注。相比传统的基于人类反馈的强化学习方法（RLHF）[1]，DPO避免了复杂的奖励模型训练和策略迭代过程，大大简化了人类偏好对齐的技术路线。

然而，标准DPO算法使用固定的β超参数来平衡参考策略和偏好学习，这导致了两个主要局限：一是上下文不敏感，无法根据不同输入内容自适应调整学习策略；二是优化效率受限，在不同难度任务上无法精细控制探索与利用的平衡。本研究旨在克服这些局限，提出一种更灵活、更有效的DPO变体算法。

本文的主要贡献包括：(1) 提出了Learnable Beta DPO算法，通过引入动态可学习的β参数实现自适应探索-利用平衡；(2) 设计了与策略模型紧密耦合的BetaHead网络，能够根据上下文动态调整β值；(3) 从信息融合的角度重新诠释了DPO算法，为理解β参数作用提供了新视角；(4) 通过大量实验验证了所提方法的有效性，展示了其在多个领域的性能优势。

2 理论基础

这里是理论基础内容。将详细介绍标准DPO算法、固定β的局限性以及信息融合视角下的DPO解析。

3 Learnable Beta DPO方法

这里是方法章节内容。将详细介绍BetaHead网络设计、动态β计算公式以及训练算法。

4 实验设置

这里是实验设置章节内容。将详细介绍模型与数据集、评估指标、对比基线以及具体的实验参数设置。

5 实验结果与分析

这里是实验结果章节内容。将详细分析整体性能比较、不同领域的学习效果分析以及消融实验结果。

6 结论与未来工作

这里是结论章节内容。将总结本文的主要贡献、局限性以及未来工作方向。

References

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.