

dpo缺点及DPOP解决方案



AI星际漫游

vx公众号[AI星际漫游]

来自专栏 · 每天学点深度学习 >

11 人赞同了该文章 >

一、存在的问题：标准直接偏好优化⁺ (DPO) 的失败模式

在大型语言模型⁺ (LLM) 的对齐中，目标是使模型的行为符合人类的偏好。直接偏好优化 (DPO) 是一种很有前景的技术，它试图直接通过偏好数据（即给定提示 x ，人类更喜欢回答 y_w 而不是 y_l ）来优化模型，而无需像传统的基于人类反馈的强化学习 (RLHF) 那样显式地训练一个奖励模型。

1. 标准 DPO 的目标和机制

DPO 的核心思想是，存在一个隐式的奖励函数 $r^*(x, y)$ ，人类的偏好是根据这个奖励函数生成的。一个理想的模型 π^* 会根据这个奖励函数来分配概率：

$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x) \exp(\beta r^*(x, y))$$

其中：

- $\pi_{\text{ref}}(y|x)$ 是参考模型的概率（通常是经过监督微调⁺ (SFT) 的模型）。
- β 是一个超参数，控制奖励函数对概率分布的影响程度。

DPO 的目标是找到一个策略 π_θ (模型参数为 θ)，使其能够最好地解释观察到的偏好数据。它通过最大化以下目标函数（或最小化其负数形式的损失函数）来实现这一点：

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

这里：

- D 是人类偏好数据集，包含 (x, y_w, y_l) 的三元组。
- y_w 是偏好的 (winning) 回答， y_l 是不偏好的 (losing) 回答。
- $\sigma(z) = 1/(1 + e^{-z})$ 是 sigmoid 函数。
- $\beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ 可以看作是模型 π_θ 相对于参考模型 π_{ref} 对回答 y 给出的隐式奖励⁺ $\hat{r}_\theta(x, y)$ 。

所以，DPO 损失可以更简洁地写为：

$$L_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l))]$$

DPO 试图最大化偏好回答 y_w 的隐式奖励与不偏好回答 y_l 的隐式奖励之间的差异。

2. DPO 的失败模式：降低偏好样本的似然性

尽管标准的 DPO 损失函数旨在提高偏好回答 y_w 相对于不偏好回答 y_l 的“相对概率”（即 $\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}$ 增大），但它可能无意中导致模型降低对偏好回答 y_w 本身的绝对似然性 $\log \pi_\theta(y_w|x)$ 。

理论原因和机制：

- 关注相对差异而非绝对值：

$\hat{r}_{\theta}(x, y_l)$ 的

赞同 11

添加评论

分享

喜欢

收藏

申请转载

...

关于作者



AI星际漫游

vx公众号[AI星际漫游]

回答

37

文章

76

关注者

141

关注他

发私信



证 $\log \pi_{\theta}(y_w|x)$ 会增加。

- **低编辑距离问题**：这种失败模式在偏好对 (y_w, y_l) 之间文本差异很小（即编辑距离低）时尤为突出。
 - 假设 $y_w = S_1 S_2 \dots S_k W_1 \dots W_m$ 和 $y_l = S_1 S_2 \dots S_k L_1 \dots L_p$ ，其中 S_i 是共享的前缀词元， W_j 和 L_j 是不同的后续词元。
 - 为了最大化 y_w 和 y_l 之间的奖励差异，DPO 的梯度可能会集中在使模型增加区分性词元 (W_j) 的概率，并降低 y_l 中对应词元 (L_j) 的概率。
 - 然而，DPO 在增加区分词元概率的同时，**可能会减少共享前缀 S_i 或 y_w 中后续词元 W_j 的概率**。这可以被理解成一种“拆东墙补西墙”的效应。如果对共享部分的概率惩罚过大，即使区分部分的相对概率增加了，偏好序列 y_w 的整体（绝对）对数似然性也可能下降。
- **梯度分析**：考虑 DPO 损失对模型输出 logits 的梯度。对于偏好序列 y_w 中的一个词元 $y_{w,t}$ ，其梯度的贡献大致为：

$$\frac{\partial L_{\text{DPO}}}{\partial \text{logit}(y_{w,t})} \propto \left(1 - \sigma(\Delta r)\right) \left(\frac{1}{\pi_{\theta}(y_{w,t}|x, y_{w,1:t-1})} - \frac{1}{\pi_{\text{ref}}(y_{w,t}|x, y_{w,1:t-1})} \right)$$

其中

$$\Delta r = \hat{\theta}(x, y_w) - \hat{\theta}(x, y_l)$$

对于不偏好序列 y_l 中的一个词元 $y_{l,t}$ ，其梯度的贡献大致为：

$$\frac{\partial L_{\text{DPO}}}{\partial \text{logit}(y_{l,t})} \propto -\left(1 - \sigma(\Delta r)\right) \left(\frac{1}{\pi_{\theta}(y_{l,t}|x, y_{l,1:t-1})} - \frac{1}{\pi_{\text{ref}}(y_{l,t}|x, y_{l,1:t-1})} \right)$$

虽然这些梯度试图增加 y_w 的相对奖励并减少 y_l 的相对奖励，但当 y_w 和 y_l 高度相似时，这些更新可能以复杂的方式相互作用，导致 y_w 的整体概率下降。DPO 可能只关注少数区分性词元的 logits。

二、解决方案：DPO-Positive⁺ (DPOP)

为了解决标准 DPO 方法中存在的降低偏好样本似然性的问题，作者设计了一种新的损失函数和训练程序，称为 DPO-Positive (DPOP)。

1. DPOP 的目标

DPOP 的核心目标是在保持 DPO 区分偏好和不偏好回答能力的同时，**避免或减轻对偏好回答 y_w 绝对似然性的不必要惩罚，甚至促进其提升**。它通过在损失函数中引入一个新的惩罚项来实现这一目标，该惩罚项专门用于激励模型维持偏好样本的高对数似然性。

2. DPOP 损失函数

DPOP 完整损失函数如下：

$$L_{\text{DPOP}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \left(\log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) - \lambda \cdot \max \left(0, \log \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\theta}(y_w|x)} \right) \right) \right]$$

这里：

- π_{θ} 是当前正在优化的模型。
- π_{ref} 是参考模型（通常是经过 SFT 的模型）。
- (x, y_w, y_l) 分别是提示、偏好的回答和不偏好的回答。
- β 是控制 KL 正则化强度的超参数，与 DPO 中的定义一致。
- $\lambda > 0$ 是一个新的超参数，用于控制新增惩罚项的权重。
- σ 是 sigmoid 函数。
- 核心的新增惩罚项是 $-\lambda \log \pi_{\theta}(y_w|x)$

3. DPOP 如何解决失败模式

DPOP 通过其新增的惩罚项来解决 DPO 的失败模式：

- **激励维持偏好样本的似然性：**惩罚项 $\text{max}\left(0, \log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\theta}(y_w|x)}\right)$ ($\pi_{\theta}(y_w|x)$) 的作用机制如下：
 - 当 $\pi_{\theta}(y_w|x) \geq \pi_{\text{ref}}(y_w|x)$ 时（即当前模型对偏好回答 y_w 的似然性不低于参考模型时）， $\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\theta}(y_w|x)} \leq 0$ ，此时 $\text{max}(0, \dots)$ 项为 0，惩罚项不起作用。
 - 当 $\pi_{\theta}(y_w|x) < \pi_{\text{ref}}(y_w|x)$ 时（即当前模型降低了偏好回答 y_w 的似然性）， $\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\theta}(y_w|x)} > 0$ ，此时惩罚项变为一个正值，通过损失函数的最小化过程，模型将被激励去提高 $\pi_{\theta}(y_w|x)$ ，使其至少不低于（或努力恢复到） $\pi_{\text{ref}}(y_w|x)$ 的水平。
- **修正梯度方向：**DPO 在处理低编辑距离样本时，对于差异词元之后的词元可能会产生“错误方向”的梯度更新，从而降低正确词元的概率。DPOP 的新惩罚项修正了这一点。在 $\pi_{\text{ref}}(y_w|x) < 1$ 的情况下（即 $\frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} < 1$ ），DPOP 损失函数对于偏好回答 y_w 中词元的梯度变为：

$$\nabla_{\theta_j} [\log \pi_{\theta}(t_k|y_w^{1:k-1}, x) - \log \pi_{\theta}(t_k|y_l^{1:k-1}, x) + \lambda \log \pi_{\theta}(t_k|y_w^{1:k-1}, x)]$$

展开后，对于词元 t_k （假设其在词汇表中的索引为 i ），其梯度关于第 j 个 logit θ_j 的分量为：

$$= \begin{cases} \lambda(1 - s_j^{y_w^{<k}, x}) + s_j^{y_l^{<k}, x} - s_j^{y_w^{<k}, x} & \text{if } i=j \\ -(\lambda+1)s_j^{y_w^{<k}, x} + s_j^{y_l^{<k}, x} & \text{if } i \neq j \end{cases}$$
 其中 $s_j^{(context, x)}$ 是在给定上下文和输入 x 的条件下，模型词汇表中第 j 个词元的 softmax 概率。对于足够大的 λ 值，当 $i=j$ （即词元 t_k 对应正确的 logit）时，梯度保证为正；当 $i \neq j$ （对应错误的 logits）时，梯度保证为负（只要 $s_j^{y_w^{1:k-1}, x} > 0$ ）。这就修正了 DPO 中可能出现的降低正确词元概率的问题。

- **确保偏好样本的高似然性：**通过这种优化压力，模型不能再仅仅通过大幅降低不偏好样本的似然性（相比于降低偏好样本似然性的程度）来最小化损失。它必须同时确保偏好样本的似然性相对于参考模型保持在较高水平。

4. DPOP 中的隐式奖励参数化

DPOP 损失函数保留了在 Bradley-Terry 模型下拟合偏好数据的特性。其隐式奖励参数化如下：

- 对于不偏好的回答 y_l ：

$$\hat{r}_{\text{DPOP}}(x, y_l) = \beta \log\frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$$
 （这与标准 DPO 对 y_l 的奖励形式相同）
- 对于偏好的回答 y_w ：

$$\hat{r}_{\text{DPOP}}(x, y_w) = \beta \left[\log\frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda \cdot \text{max}\left(0, \log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\theta}(y_w|x)}\right) \right]$$
 可以看到，对偏好回答 y_w 的奖励在其原始 DPO 形式的基础上，增加了一个与新惩罚项相关的调整。这个调整确保了如果 $\pi_{\theta}(y_w|x)$ 相对于 $\pi_{\text{ref}}(y_w|x)$ 过低，那么其有效奖励会受到相应的“惩罚”或“修正”，从而激励模型提升 $\pi_{\theta}(y_w|x)$ 。

5. 与对比损失⁺的联系 (Connection to Contrastive Loss)

尽管 DPOP 的主要动机是解决 DPO 的失败模式，但它也与对比学习中的对比损失 (contrastive loss) 有一定的联系。DPOP 的损失函数形式可以被看作是带有特定间隔 (margin) 的对比损失，其中间隔 $m = \log\frac{1}{\pi_{\text{ref}}(y_w|x)}$ 。标准 DPO 类似于缺少了对比损失中“相似点对”项和间隔项的公式，而 DPOP 通过引入新的惩罚项（可以类比为相似点对的项和间隔）弥补了这一不足。

通过这些机制，DPOP 旨在更稳健地从偏好数据中学习，避免标准 DPO 可能出现的性能退化，并在多种情况下展现出更优越的性能。

编辑于 2025-05-14 00:26 · 浙

写下你的评论...



还没有评论，发表第一个评论吧

推荐阅读



**DPO正例概率不应该下降？
DPO在实践中的反思与改进**

leshare

DPO和实现代码

1. 提出 DPO 的背景RLHF 的缺陷：基于人类反馈的强化学习（RLHF）过程复杂且不稳定，需要拟合奖励模型并使用强化学习微调大型无监督语言模型（LM）。DPO 的目的：通过利用奖励函数与最优策...

Evan Lee



从原理到实战技巧，详解DPO的优缺点

黑木

**（详解）
言模型！**

引言在上
了DPO这
作为PPC
于大模型
齐。其中
一个问题
黑木