DPO

1.简介

基于 **人类反馈的强化学习（RLHF）** 是一个复杂且不稳定的过程，拟合一个反映人类偏好的奖励模型，然后使用强化学习对大语言模型进行微调，以最大限度地提高估计奖励，同时又不能偏离原始模型太远。这涉及训练多个 LM，并在训练循环中从 LM 采样，从而产生大量的计算成本。

本文作者提出了 **直接偏好优化（DPO）** 算法，它稳定、高效且计算量轻，**无需拟合奖励模型，也无需在微调期间从 LM 采样或执行显著的超参数调整**。

实验表明，DPO 可以微调 LMs，使其与人类偏好保持一致，与现有方法一样或更好。值得注意的是，DPO 在情绪控制的能力上超越了 RLHF，提高了总结和单轮对话的响应质量，同时大大简化了实现和训练。

2.RLHF pipeline

RLHF 通常由 3 个阶段组成：

1. **监督微调 (SFT)**：高质量数据集上通过监督学习
2. **偏好采样和奖励学习 (RM)**：标注排序的判别式标注成本远远低于生成答案的生成式标注。
3. **强化学习微调 (PPO)**：在对 SFT 模型进行微调时生成的答案分布也会发生变化，会导致 RM 模型的评分会有偏差，需要用到强化学习.

2.1 SFT 阶段

RLHF 通常从一个通用的预训练 LM 开始，该 LM 在高质量数据集上通过监督学习（最大似然）对感兴趣的下游任务（如对话、指令跟随、总结等）进行微调，以获得模型 π^{SFT} 。

2.2 Reward 建模阶段

在第二阶段，用 x 提示 π^{SFT} 产生一对答案 $(y_1, y_2) \sim \pi^{SFT}$ 。通过人类标注，得到偏好标签 $y_w \succ y_l$ ，其中 y_w 表示首选 prompt， y_l 表示非首选 prompt。

通过静态数据集 $D = \left\{x^i, y_w^i, y_l^i\right\}_{i=1}^N$，可以将奖励模型 $r_\phi(x, y)$ 参数化，并通过极大似然估计参数。将问题定义为二元分类，有负对数似然损失：

$$\mathcal{L}_R\left(r_\phi, \mathcal{D}\right) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(r_\phi\left(x, y_w\right) - r_\phi\left(x, y_l\right)\right)\right]$$

其中 σ 是 `sigmoid` 函数。奖励模型 $r_\phi(x, y)$ 通常由 π^{SFT} 进行初始化，并在最后一个 Transformer 层之后添加线性层，该层为奖励值生成单个标量预测。

2.3 RL 微调阶段

在 RL 阶段，使用学习到的奖励函数来对语言模型进行打分。特别是，制定了以下优化问题：

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)}\left[r_\phi(x, y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \| \pi_{\mathrm{ref}}(y \mid x)\right]$$

其中 β 是控制 π_θ 偏离基本参考策略 π_{ref} 的参数。在实践中，语言模型策略 π_θ 也被初始化为 π_{ref} 。**添加的 β 约束很重要，因为它可以防止模型偏离奖励模型准确的分布太远**，以及保持生成多样性和防止模式崩溃为单个高奖励答案。

由于语言生成的离散性，这个目标是不可微的，并且通常使用强化学习进行优化。标准方法是构造奖励函数 $r(x, y) = r_\phi(x, y) - \beta\left(\log \pi_\theta(y \mid x) - \log \pi_{ref}(y \mid x)\right)$，并利用 PPO 最大化。

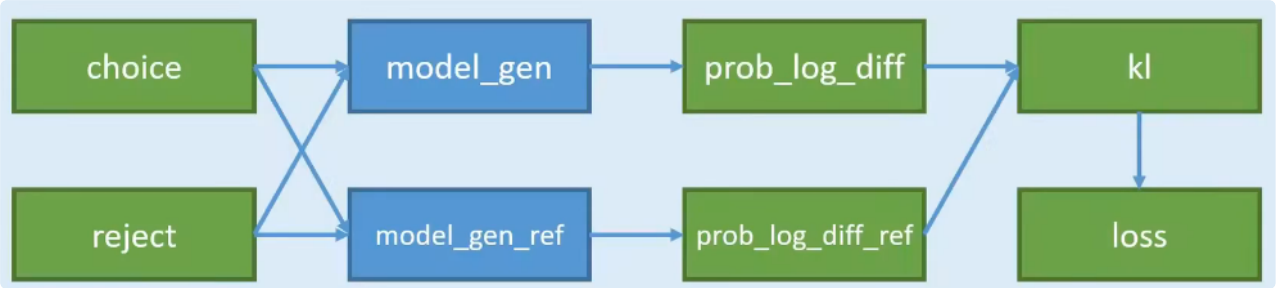
3.直接偏好优化（DPO）

与之前的 RLHF 方法不同，**DPO 绕过了奖励建模步骤，并使用偏好数据直接优化语言模型**。

3.1 PPO 算法总览

1. 对一个问题，有两个回答 choice 和 reject，不是一个一定正确，一个一定不正确；而是训练出的语言模型，更加 prefer 哪一种，即希望语言模型以哪一种方式来回答。
2. 准备两个模型 model_gen 和 model_gen_ref，其实是一摸一样的模型，只不过在训练过程中，只会训练其中一个，另外一个是不训练的。
3. 把两两份数据，分别输入到两个模型中计算，可以得到 4 份概率；

4. 4 份数据中，其中有 2 份是想要的，2 份是不想要的；2 份想要的做差，得到 `pro_log_diff`，2 份不想要的做差 `pro_log_diff_ref`
5. 拿 2 份做差的数据，计算 KL 散度；惩罚 policy 模型对正样本概率的下降和负样本概率的上升
6. 以 KL 散度计算 Loss



3.1 DPO 目标函数

类似于奖励建模方法，策略目标变为：（推导过程详见原论文）

$$\mathcal{L}_{\mathrm{DPO}}\left(\pi_\theta ; \pi_{\mathrm{ref}}\right)=-\mathbb{E}_{\left(x, y_w, y_l\right) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)}-\beta \log \frac{\pi_\theta\left(y_l \mid \right.\right.\right.$$

通过这种方式，绕过了显式奖励建模步骤，同时也避免了执行强化学习优化的需要。

逐步分析这个优化目标：首先，σ 函数里面的值越大，L_DPO 越小。即最大化 y_w 和 y_l 的奖励函数：

$$r_w=\log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)}$$

$$r_l=\log \frac{\pi_{\mathrm{ref}}\left(y_l \mid x\right)}{\pi_\theta\left(y_l \mid x\right)}$$

- 对于人类偏好结果 y_w ，我们期望 $\pi_\theta(y_w \mid x)$ 越大越好；
- 对于人类非偏好结果 y_l ，我们期望 $\pi_\theta(y_l \mid x)$ 越小越好。
- 如果 $\pi_{\mathrm{ref}}\left(y_w \mid x\right)$ 比较小，说明参考模型 π^{ref} 没有正确分类该偏好响应 y_w ，此时 r_w 的奖励系数很大。
- 如果 $\pi_{\mathrm{ref}}\left(y_l \mid x\right)$ 比较大，说明参考模型 π^{ref} 没有正确分类该非偏好响应 y_l ，此时 r_l 的奖励系数很大

3.2 DPO outline

1. 对于每个 prompt x，从参考策略中采样补全 $(y_1, y_2) \sim \pi_{\text{ref}}(\cdot \mid x)$，用人类偏好进行标记以构建离线偏好数据集 $D = \left\{x^i, y_w^i, y_l^i\right\}_{i=1}^N$ 。
2. 对于给定的 π_{ref} 、 D 和 β ，优化语言模型 π_θ 以最小化 L_{DPO} 。

由于偏好数据集使用 π^{SFT} 进行采样，因此只要可用，就会初始化 $\pi_{\text{ref}} = \pi^{SFT}$ 。在实践中，人们更愿意重用公开的偏好数据集，而不是生成样本并收集人类偏好。这时我们通过最大化首选 prompt (x, y_w) 的似然来初始化 π_{ref} ，即

$$\pi_{\text{ref}} = \arg\max_{\pi} \mathbb{E}_{x, y_w \sim \mathcal{D}} \left[\log \pi\left(y_w \mid x\right)\right]$$

该过程有助于缓解真实 π_{ref} 与 DPO 使用的 π_{ref} 之间的分布偏移。

4.实验

- **最大化奖励的同时最小化 KL 散度**。可以看到 DPO 在保持较小 KL 散度时，也能够达到最大奖励。而 PPO 随着奖励的增大，KL 散度也在增大。
- **对不同采样温度的鲁棒性**。DPO 在不同的采样温度下全面优于 PPO，同时在 Best of N 基线的最佳温度下也更胜一筹。

5.结论

基于人类反馈的强化学习（RLHF）是一个复杂且不稳定的过程，首先拟合一个反映人类偏好的奖励模型，然后使用强化学习对大语言模型进行微调，以最大限度地提高估计奖励，同时又不能偏离原始模型太远。这涉及训练多个 LM，并在训练循环中从 LM 采样，从而产生大量的计算成本。

本文作者提出了直接偏好优化（DPO）算法，它稳定、高效且计算量轻，无需拟合奖励模型，也无需在微调期间从 LM 采样或执行显著的超参数调整。

实验表明，DPO 可以微调 LMs，使其与人类偏好保持一致，与现有方法一样或更好。值得注意的是，DPO 在情绪控制的能力上超越了 RLHF，提高了总结和单轮对话的响应质量，同时大大简化了实现和训练。