Direct Preference Optimization: Your Language Model is Secretly a Reward Model [paper]

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn

Stanford University, Chan Zuckerberg Biohub

Problem

1. RLHF is a **complex** and often **unstable** procedure, first fitting a reward model that reflects the human preferences, and then fine-tuning the large unsupervised LM using reinforcement learning to maximize this estimated reward without drifting too far from the original model.

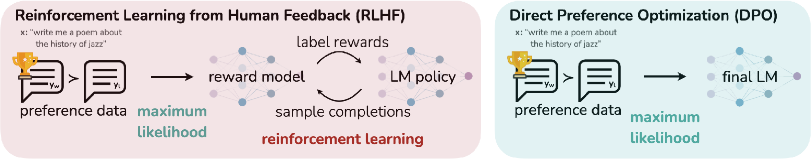

Figure 1: From RLHF to DPO

Method

- Deriving the DPO objective

  RL Fine-Tuning Phase:

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

  The optimal solution of KL-constrained reward maximization objective: :paperclip:

  $$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\mathrm{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right)$$

  $Z(x)$ is a partition function, which is hard to estimate.

  This makes it hard to utilize in practice. Take the logarithm of both sides, we have:

  $$r(x, y) = \beta \log \frac{\pi_r(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta \log Z(x).$$

Method

- Deriving the DPO objective

  Recall Bradley-Terry model:

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}$$

  Substituting the reparameterized $r(x, y)$ into BT model, the optimal RLHF policy π^* satisfy the preference model:

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

Method

- Deriving the DPO objective

  Now we have the probability of human preference data interms of the optimal policy rather than the reward model. To solve it, formualting a maximum likelihood objective for π_θ:

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

Method

- Deriving the DPO objective

  The gradient with respect to θ:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \bigg] \bigg]$$

  where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$,

  is the reward implicitly defined by the language model π_θ and reference model π_{ref}.

Method

- Utilization 1. Sample $y_1, y_2 \sim \pi_{\text{ref}}(.|x)$ for every prompt x, label them with human preferences. 2. Optimize π_θ to minimize $\mathcal{L}_{\text{DPO}}$.