DPO

research paper explained

SECTIONS

Here are the sections we are going to see about

Abstract

Introduction

Preliminaries

DPO Method

Experiments

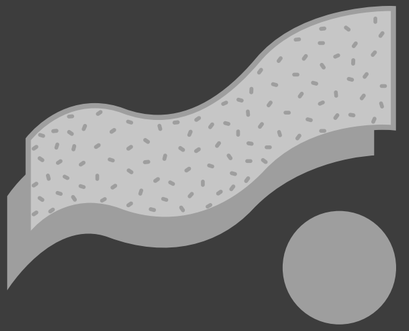
Conclusion

ESSENCE

“

DPO is a computationally efficient method that calculates the log probabilities of preferred and dispreferred completions under a model and optimizes its params in a way to increase the likelihood of preferred responses and decrease those dispreferred to align the model with human preferences

ABSTRACT



LLMs have a broad knowledge but since it is difficult to control its generation behavior we have control mechanisms like RLHF to steer the generation as per preference but it is complex and unstable in nature

The aim of the paper is to map between the reward function and optimal reward policies for optimizing model to align with preferences by a single stage policy training approach making it computationally efficient

DPO - Direct Preference Optimization is a lightweight, highly performant method to train LLMs on preference datasets without a reward model. Experiments have shown that DPO aligns the model with the preferences better than RLHF

INTRODUCTION

- **Challenge Existing:** LLMs possess great capabilities due to it is training on a large variety of datasets with different goals, skillsets, and priorities. But some of them are not desirable. For eg: We need the model to know about common misconceptions among people but should be aware that it is a misconception
- **Existing Solution:** In simple terms, it is important to select model responses and have steering control over model generation capabilities to match our own preferences for which mechanisms like RLHF are used
- **RLHF:** Reinforcement Learning with Human Feedback is a method based on RL which involves creating a reward model based on preference datasets and utilizing it to optimize the SFT model performance
- **Problems in RLHF:** RLHF is very expensive and complex due to its involvement of multiple training loops

INTRODUCTION

DPO

01 DPO has the same objective of reward maximization using KL-divergence constraint like RLHF but simpler to train and implement

02 DPO update increases the relative log probability of preferred over dispreferred responses

03 Relies on a theoretical preference model called Bradley-Terry model to measure how well the model aligns with preference datasers

04 Uses a loss function called DPO loss eliminating the need for reward model

PRELIMINARIES - RLHF

RLHF occurs in three different phases. They are:

SFT phase: By supervised finetuning, on a high-quality dataset like instruction tuning it will result in a model called **πSFT**

Reward modeling phase: SFT is prompted with a question and a pair of answers (preferred, dispreferred) -> **(y1, y2) ~ πSFT(y | x)**

The preferences are assumed to be generated by a latent reward model **r∗(y, x)**

For preference distribution, the reward model follows a Bradley-Terry model as follows

p ∗ (y1 ≻ y2 | x) = exp (r ∗ (x, y1)) / exp (r ∗(x, y1)) + exp (r ∗(x, y2)) -> eq.1

Here **y1 = yw -> preferred response** and **y2 = yl -> dispreferred response -> yw > yl**

PRELIMINARIES - RLHF

With a Dataset,

$$\mathcal{D} = \left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$$

a reward model **rφ(x, y)** can be parameterized under φ and estimate the params via maximum likelihood

Framing the above as a binary classification problem the negative likelihood loss can be formulated as

LR(rφ, D) = −E(x,yw,yl)∼D log σ(rφ(x, yw) − rφ(x, yl)) -> eq.2

where σ is logistic function

PRELIMINARIES - RLHF

In LMs, **$r_\varphi(x, y)$** is usually **$\pi^{SFT}(y \mid x)$** with a **final layer added** to provide the reward value.

To ensure low variance in reward function, prior works normalize the rewards such that,

$$E_{x,y\sim D}\,[r_\varphi(x, y)] = 0 \text{ for all } x$$

RL Finetuning phase:

The reward function is used to optimize the model as follows

$$\max_{\pi_\theta} E_{x\sim D, y\sim\pi_\theta(y|x)}[r_\varphi(x, y)] - \beta D_{KL}[\pi_\theta(y \mid x) \,||\, \pi_{ref}(y \mid x)] \rightarrow \text{eq. 3}$$

where **π_θ -> LLM, π_{ref} -> reference policy or model**

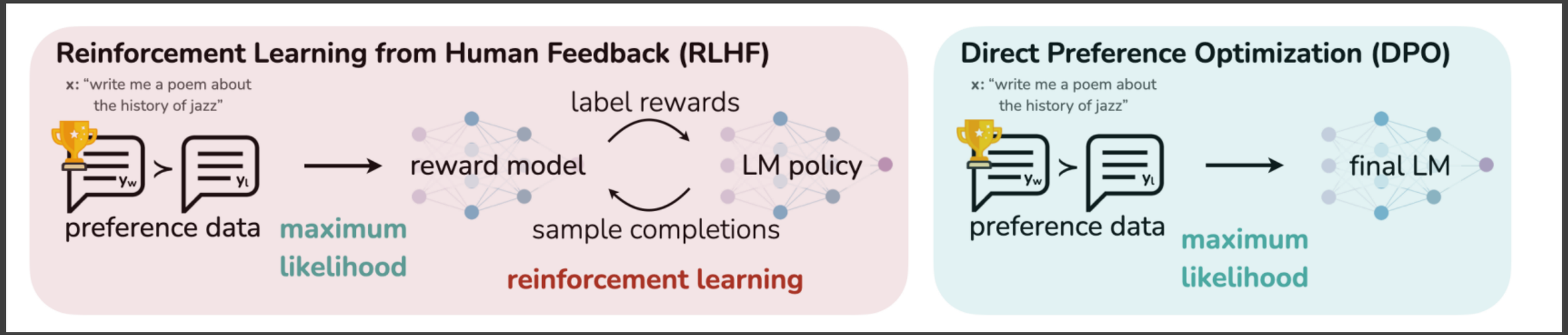
The final reward function r(x,y) can be calculated with the following equation,

$$r(x, y) = r_\varphi(x, y) - \beta(\log \pi_\theta(y \mid x) - \log \pi_{ref}(y \mid x))$$

and the aim is to maximize this reward using PPO (Proximal Policy Optimization)

DPO Method

DPO


Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
label rewards
reward model
LM policy
sample completions
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
y_l
preference data
maximum likelihood
final LM

DPO METHOD

To solve the challenges of RLHF on large scale problem DPO approach was introduced which will bypass the reward modeling step and directly optimizes a language model using a preference data

Here loss function is transformed into a loss function over policies which allows skipping the reward modeling step but still preference model - Bradly Terry is used for the optimization

DERIVING DPO OBJECTIVE

eq.3 (reward model to optimize LLM) under a general reward function R with KL-constrained reward maximization objective becomes

$$\pi r(y \mid x) = 1/Z(x)\ \pi ref(y \mid x)\ \exp(1/\beta\ r(x, y)) \rightarrow \text{eq.4}$$

where

$$Z(x) = \sum y\ \pi ref(y \mid x)\ \exp(1/\beta\ r(x, y))$$

Z(x) is the partition function

Rearranging eq.4 with some algebra in terms of optimal policy πr, reference policy πref, and Z(.)

$$r(x, y) = \beta \log(\pi r(y \mid x)/\pi ref(y \mid x)) + \beta \log Z(x) \rightarrow \text{eq.5}$$

DERIVING DPO OBJECTIVE

eq.3 (reward model to optimize LLM) under a general reward function R with KL-constrained reward maximization objective becomes

$$\pi r(y \mid x) = 1/Z(x)\ \pi ref(y \mid x)\ \exp(1/\beta\ r(x, y)) \rightarrow \text{eq.4}$$

where

$$Z(x) = \sum y\ \pi ref(y \mid x)\ \exp(1/\beta\ r(x, y))$$

Z(x) is the partition function. Complete derivation in Appendix 1

Rearranging eq.4 with some algebra in terms of optimal policy πr, reference policy πref, and Z(.)

$$r(x, y) = \beta \log(\pi r(y \mid x)/\pi ref(y \mid x)) + \beta \log Z(x) \rightarrow \text{eq.5}$$

DERIVING DPO OBJECTIVE

Reparameterizing to ground truth reward r* optimal model becomes π* as follows

$$r(x, y) = \beta \log(\pi^*(y \mid x)\, \pi_{ref}(y \mid x)) + \beta \log Z(x)$$

Since the Bradley-Terry model depends only on rewards' difference

$$p^*(y1 \succ y2 \mid x) = \sigma(r^*(x, y1) - r^*(x, y2))$$

Optimal RLHF policy π* under the BT model becomes the preference model after some operations becomes, -> refer Appendix 6

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

DERIVING DPO OBJECTIVE

Now we have an optimal policy from which we can formulate a maximum likelihood objective for πθ based on eq.2

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

The above equation shows the formulation of DPO loss which bypasses the reward modeling step

DPO UPDATE

The gradient of the loss function LDPO increases the likelihood of the preferred completions yw and decreases the likelihood of dispreferred completions yl

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

where,

$$\hat{r}_\theta(x, y) = \beta \log(\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x))$$

is the final DPO reward function

EXPERIMENTS

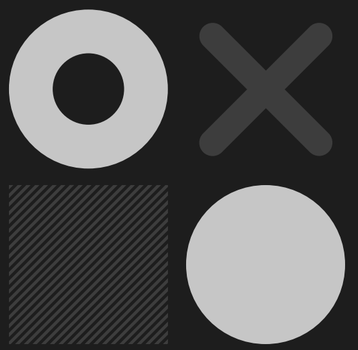
Tasks: There were three different tasks for evaluation. They are controlled sentiment generation, summarization, and single-turn dialog generation

Evaluation: For evaluation KL divergence and GPT-4 are used for evaluation

Methods: Zero-shot prompting with GPT-J in summarization and 2-shot prompting with Pythia 2.8B for dialog task

HOW WELL DID DPO OPTIMIZE

After conducting a series of experiments with **KL∈{3, 6, 9, 12} for PPO**, **β∈{0.05, 0.1, 1.5}, α∈{0.05, 0.1, 0.5, 1} for DPO**, it showed that **DPO** performs **better** than **PPO** at achieving the **highest reward with low KL divergence.**



CONCLUSIONS

DPO is a simple training paradigm for training LM from preferences without RL since it identifies a mapping between LM policies and reward function functions. With a simple cross-entropy loss, LMs are trained to align with preferences

Some of the advantages of DPO are DPO performs similar or better than existing RLHF algorithms including RLHF based on PPO. DPO also reduces the barrier for training more LMs on human preferences datasets