Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Author: Rafailov et al.

created by paper2slides

2023-05-29

https://arxiv.org/abs/2305.18290

Executive Summary

- Large-scale LMs show exceptional capabilities, but steering their behavior effectively is challenging due to their unsupervised training.
- Current methods use RLHF, which trains a reward model from human feedback and optimizes the LM using RL to match human preferences.
- RLHF is complex and unstable, often involving reward model training and RL fine-tuning with risks of the model drifting.
- Direct Preference Optimization(DPO) directly optimizes the policy based on human preferences using a simple classification loss, avoiding RL and separate reward model training.

Introduction

Challenges with Large-Scale LMs:

- Unsupervised training leads to unpredictable behavior.
- Steering models to align with human preferences is complex.

Current Solutions (RLHF):

- RLHF fine-tunes LMs to match human preferences.
- Involves two-stage process: training a reward model and RL optimization.
- Highly complex, unstable, and computationally intensive.

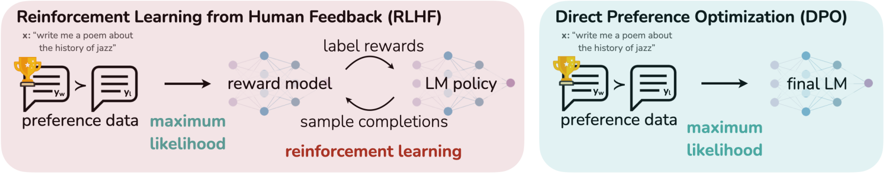

Figure: DPO optimizes for human preferences directly without RL.

Proposed Method: Direct Preference Optimization

Key Insight:

- Leverage mapping between reward functions and optimal policies.
- Optimize constrained reward maximization problem as classification.

Simplification and Stability:

- Single-stage policy training with classification loss.
- Avoid explicit reward models and RL, reducing complexity.
- No in-loop sampling or intensive hyperparameter tuning.

Formal Definition and Objective

KL-Constrained Reward Maximization:

$$\max_{\pi_\theta} \mathbb{E}_{x,y\sim\pi_\theta}[r(x,y)] - \beta \mathbb{D}_{\mathrm{KL}}[\pi_\theta \| \pi_{\text{ref}}] \tag{1}$$

Optimal Solution:

$$\pi_r(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y)\right) \tag{2}$$

- β: Controls deviation from the reference policy π_{ref}.
- $Z(x)$: Partition function ensuring normalization.

Reparameterization

Reparameterized Reward Function:

$$r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{3}$$

Objective Without Explicit Reward Model:

$$p^*(y_1 \succ y_2|x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)} \tag{4}$$

DPO Loss Function

Derive DPO Loss:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \tag{5}$$

Intuition:

- Increases log-probability of preferred responses relative to dispreferred.
- Controls deviation with dynamic importance weight.

Experimental Setup

Controlled Sentiment Generation:

- Dataset: IMDb movie reviews.
- Model: GPT-2-large SFT on IMDb reviews.
- Reward: Pre-trained sentiment classifier.

Summarization:

- Dataset: Reddit TL;DR, human preference data.
- Model: GPT-J SFT on human-written summaries.

Single-Turn Dialogue:

- Dataset: Anthropic Helpful and Harmless (HH).
- Preference data: 170k dialogues with labeled preferences.

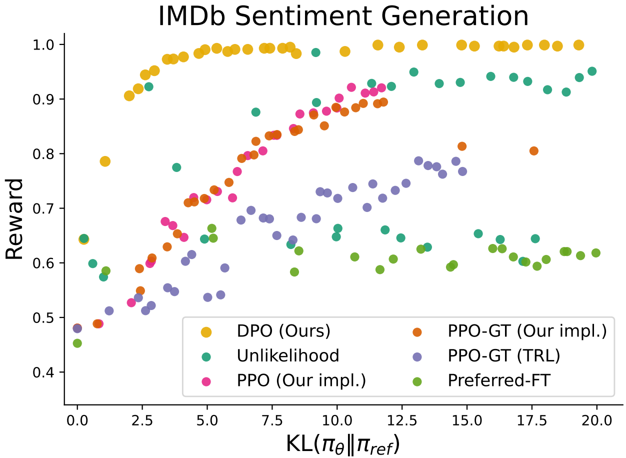

Figure: Expected reward vs. KL to reference policy.

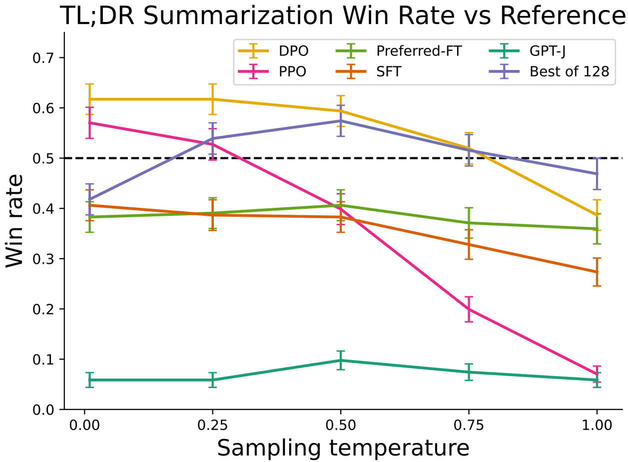

Figure: Win rates for TL;DR Summarization task.

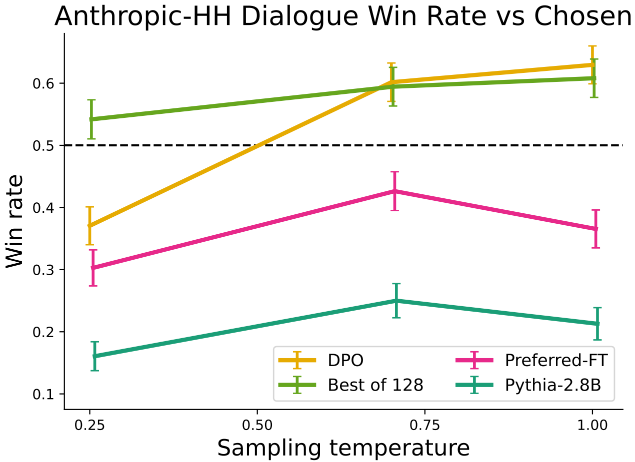

Figure: Win rates for Anthropic-HH one-step dialogue.

Generalization to New Input Distribution

Experiment on CNN/DailyMail Dataset:

- Evaluate DPO and PPO policies on new distribution.

	Win Rate (Temp 0)	Win Rate (Temp 0.25)
DPO	0.36	0.31
PPO	0.26	0.23

Table: GPT-4 win rates for CNN/DailyMail summarization.

- DPO maintains higher win rates even on unexpected inputs.
- Demonstrates robustness and generalizability of DPO policies.

Conclusion and Future Work

Summary:

- Direct Preference Optimization(DPO) optimizes LMs from preferences efficiently without RL.
- Achieves superior performance in diverse, real-world tasks.

Future Directions:

- Explore broader applications beyond language modeling.
- Investigate generalization and robustness across varied domains deeply.
- Scale experiments to state-of-the-art models with larger parameters.