Report on Comparison of RLHF and DPO Based on Sample Efficiency, Response Quality, and Computation Cost

Paramananda Bhaskar

21CS30035

Disclaimer: Due to Kaggle's resource constraints (maximum session duration of 10 hours), we were only able to train for **1 epoch** in **RLHF PPO** and **DPO**. The performance **would improve significantly with more training epochs**.

Introduction

Reinforcement Learning from Human Feedback (RLHF) using Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) are two prominent approaches for fine-tuning language models using human feedback. This report contrasts these methods based on three key factors: sample efficiency, response quality, and computation cost.

Sample Efficiency

Sample efficiency refers to how well the model improves with a given amount of training data and computation.

RLHF:

- Requires training a **separate reward model** from scratch, adding an additional layer of training complexity.
- The PPO fine-tuning step involves significant **batch-wise computation**, leading to **inefficient updates**.
- Encountered training inefficiencies, such as **skipping batches** due to excessively high policy updates (**warnings in PPO training logs**).
- Generated responses showed only **marginal improvements** in reward scores after one epoch.

Reward Model Training Logs:

- **Epoch 1:** Loss = **0.0227** (Duration: **84m 57s**)
- **Epoch 2:** Loss = **0.0097** (Duration: **85m 07s**)
- **Epoch 3:** Loss = **0.0068** (Duration: **85m 03s**)

Reward Model Evaluation on Test Set:

- **Average Reward on More Preferred Responses: 8.22**
- **Average Reward on Less Preferred Responses: -9.69**
- **Average Reward Difference (r1 - r2): 17.9118**

- **Percentage of Pairs Where More Preferred Response Has Higher Reward: 99.73%**

DPO:

- **Directly optimizes** the model on preference data **without requiring a separate reward model**.
- Utilizes **all training samples efficiently** with a more straightforward **gradient update mechanism**.
- **Processes significantly more samples per unit time**.
- Achieved **meaningful improvement** in responses after just one epoch.

Verdict: DPO is significantly more **sample-efficient**, avoiding inefficiencies from training a separate reward model and complex policy updates.

Response Quality

Response quality is measured by the **coherence, relevance, and alignment** of the generated responses with human preferences. We evaluate this using both BLEU and ROUGE metrics, which capture different aspects of textual similarity:

- **BLEU Score:** Measures n-gram precision, indicating how many words or phrases in the generated text match the reference. Higher BLEU indicates better lexical overlap.
- **ROUGE Scores:** Measure recall and longest common subsequence (LCS) overlap, providing insight into content coverage and overall structural similarity.

RLHF (PPO) Implementation:

- **Observations:**
  - **Pre-training:** The initial outputs were largely incoherent, showing little alignment with the intended user prompt.
  - **Post-training:** There is a noticeable improvement in response alignment and coherence, although some inconsistencies remain. While there is some improvement, the RLHF PPO responses tend to behave more like a next-token predictor or perform query sentence completion rather than directly answering the question. This behavior may contribute to a slightly higher precision-based BLEU score, but the responses are less direct and contextually grounded.
- **Evaluation Results of RLHF PPO Model:**
  - **BLEU Score:** 0.063260
  - **ROUGE Scores:**
    - **ROUGE-1:** 0.1226
    - **ROUGE-2:** 0.0177
    - **ROUGE-L:** 0.0900
    - **ROUGE-Lsum:** 0.1037
- **Analysis:**

- The relatively low BLEU score suggests that the generated responses have limited n-gram overlap with the references, which is common in generation tasks with high variability.
- The ROUGE scores indicate that while there is some overlap in content (ROUGE-1), the higher-order n-grams and longer structural patterns (ROUGE-2, ROUGE-L) are less aligned.
- Overall, RLHF PPO demonstrates some improvement in aligning with human preferences but still suffers from inconsistency and limited coherence improvements.

DPO Implementation:

- **Observations:**
  - **Pre-training:** As with RLHF, initial responses were weak and not well aligned with the desired output.
  - **Post-training:** The DPO approach shows meaningful improvements, with responses that are clearer and more contextually relevant, suggesting that the direct optimization of preference probabilities leads to better fine-tuning of the model output. The DPO model produces clearer and more contextually relevant responses that better address the query directly. This direct answer generation appears to improve the overall response quality despite a slightly lower BLEU score.
- **Evaluation Results of DPO Model:**
  - **BLEU Score:** 0.060275
  - **ROUGE Scores:**
    - **ROUGE-1:** 0.1431
    - **ROUGE-2:** 0.0291
    - **ROUGE-L:** 0.1057
    - **ROUGE-Lsum:** 0.1267
- **Analysis:**
  - Although the RLHF PPO model exhibits a marginally higher BLEU score, it appears that the model is primarily engaged in next-token prediction or sentence completion rather than providing direct, contextually aligned answers.
  - The improvement in ROUGE-1 and ROUGE-2 indicates that **DPO achieves better content coverage and captures longer n-gram sequences more effectively than RLHF.**
  - Higher ROUGE-L and ROUGE-Lsum values reflect improved structural coherence and overall textual alignment with the references.
  - These improvements suggest that DPO is better at aligning generated responses with human preferences, resulting in more meaningful and contextually appropriate outputs.

Verdict:

- **Sample Efficiency & Training Dynamics:**
  - DPO directly optimizes preference probabilities, leading to faster and more stable convergence within a single epoch, while RLHF PPO requires

additional complexity with a separately trained reward model and often faces batch-skipping issues due to KL divergence constraints.

- **Response Quality:**
    - DPO shows superior response quality with better ROUGE scores, indicating improved content relevance and coherence, despite similar BLEU scores.
- **Overall:**
    - **DPO outperforms RLHF in response quality improvements given the same training time.** This suggests that for tasks requiring efficient preference optimization, DPO is the more effective approach.

Computation Cost

Computation cost is measured in terms of **training time and GPU utilization**.

RLHF:

- **Reward model training** took **~4.25 hours** (**1:25 per epoch for 3 epochs**).
- **PPO fine-tuning** took **~6.7 hours** for just **1 epoch**.
- **Total training time exceeded 11 hours**.
- **Skipped PPO batches** indicate **inefficient computation utilization**.

DPO:

- **Completed 24,000 updates of 1 epoch** in **~5.3 hours**.
- **More stable training** without skipped updates or divergence issues.
- **Requires only a single optimization step per update**, reducing overall GPU usage.

Verdict: DPO is **computationally more efficient**, completing a full training cycle in **half the time** of RLHF while processing significantly **more samples**.

Summary of Findings

This report compares Reinforcement Learning from Human Feedback (RLHF) using Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) across three key factors: sample efficiency, response quality, and computation cost. The main findings are as follows:

- **DPO is more sample-efficient** as it directly optimizes model preferences without requiring a separate reward model, leading to faster improvements.
- **DPO achieves better response quality** after limited training, providing more coherent and contextually relevant outputs compared to RLHF.
- **DPO is computationally more efficient**, requiring significantly less training time and GPU resources while processing more updates per unit time.

Given these observations, DPO emerges as a superior approach for fine-tuning language models based on human preferences.

Conclusion

Criterion	RLHF	DPO	Winner
Sample Efficiency	Requires separate reward model; slow PPO updates	Directly optimizes preferences; faster updates	**DPO**
Response Quality	Some improvement in terms of rewards, but responses remain weak	More coherent and aligned responses	**DPO**
Computation Cost	Over 11 hours with batch inefficiencies	~5.3 hours with stable training	**DPO**

Overall, **DPO is a superior approach** for fine-tuning language models based on human preferences. It offers **better sample efficiency**, produces **higher-quality responses**, and is **computationally more efficient** compared to RLHF. Given these findings, **DPO is recommended** for scenarios where **efficient preference optimization is required without excessive computational overhead**.

Future Steps

1. **Increase Training Epochs** – Due to resource constraints, only one epoch was trained for RLHF PPO and DPO. Extending training across multiple epochs would likely enhance model performance significantly.

2. **Optimize Hyperparameters** – Further tuning of learning rates, batch sizes, and update frequencies could improve convergence efficiency and response quality.

3. **Explore Alternative Architectures** – Investigating modifications to PPO or alternative reinforcement learning approaches like TRPO or A2C could enhance RLHF efficiency.

4. **Fine-tune Reward Model** – The reward model used in RLHF was trained from scratch, which may have impacted performance. Utilizing a pre-trained reward model could improve sample efficiency.

5. **Compare on Larger Datasets** – Evaluating both methods on a more diverse dataset with longer training runs would provide a more comprehensive performance comparison.

6. **Analyze Generalization** – Assessing how well the models perform on out-of-distribution prompts can help determine robustness and adaptability.