Homework #6 Report

土木所交通組 R12521502 陳冠言

Provide a brief description and comparison of DPO and ORPO.

Direct Preference Optimization (DPO)為新的一種優化模型的演算法，主要用於訓練 Large Language Model (LLM) 以符合人類偏好。相較於現有的方法，DPO 透過對 reward model 提供新的參數，同時簡化原本複雜的過程並提升其穩定性和模型表現。透過 DPO 來訓練模型得以使其直接滿足人類偏好，無需特別使用強化學習即可符合人類期望如情感調節、摘要和對話等任務。在演算法方面，DPO 使用概率比來評估和比較不同輸出間的偏好，在梯度更新方面是以偏好輸出的概率比為依據。

Monolithic Preference Optimization without Reference Model (ORPO)也是一種新的preference alignment algorithm，可用於語言模型的微調且其特色在於無需參考模型。透過 ORPO 進行微調，能提升語言模型的性能並在各種模型大小上實現顯著的改進。在演算法方面，ORPO 是使用 odds ratio 來代替概率比，因為它在不同的生成概率分佈下具有更一致的敏感度，且 ORPO 可在單一階段中進行偏好最佳化，並在梯度計算包括兩個部分，一為懲罰錯誤預測，另一者為對比選擇與拒絕輸出。

由於兩者在演算法方面具有明顯差異，DPO 使用概率比來進行preference alignment，且通常需要多階段的preference alignment 步驟；而 ORPO 則使用賠率比，在單一階段中完成偏好最佳化，從而提高了訓練效率和穩定性。由此可見，DPO的訓練時間較長。

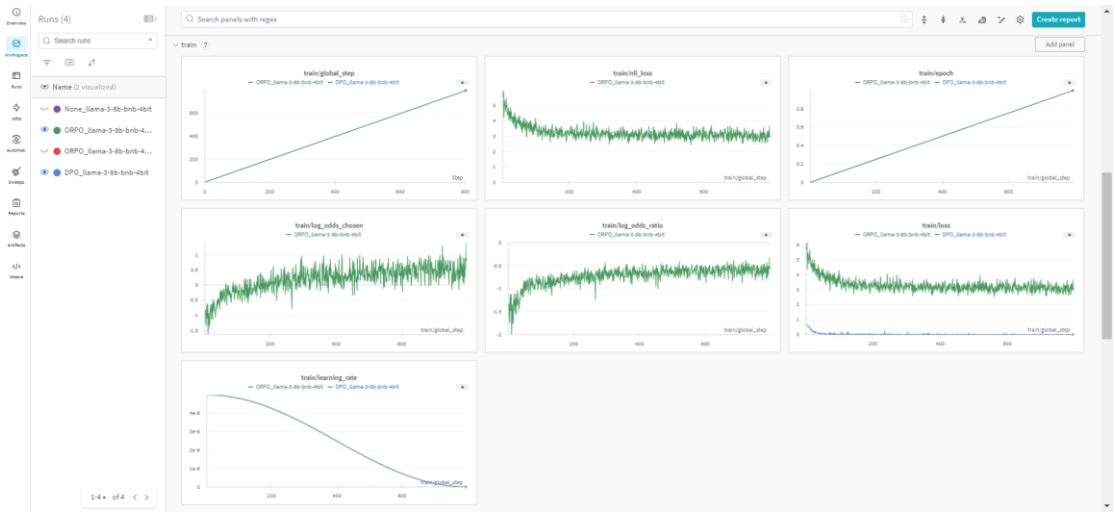
DPO (左) 和 ORPO (右) GPU使用情況

Briefly describe LoRA.

LoRA（Low-Rank Adaptation）是一種設計用來進行 parameter-efficient fine-tuning 大型語言模型的技術。在參數部分並不是進行所有參數的更新，而是僅微調一小部分的參數。在微調過程中，LoRA 將低秩矩陣添加到原始權重矩陣中，使模型能夠在不干擾預訓練知識的情況下學習特定任務的適應。

Plot your training curve by W&B, including both loss and rewards.

使用的是unsloth/llama-3-8b-bnb-4bit



Comparison and analysis of results (before & after DPO & after ORPO).

由三個模型的回答可見，在某些有特定答案且問題不複雜的情況下，三者的回答會差不多，然而若是申論題或是詢問想法等開放式問答，ORPO 的回覆會更加詳細和全面，提供了更多的資訊和觀點，而其他兩者較簡短，以下為各自的特點：

1. Before：對於較複雜且靈活的問題回答較籠統，且比較片面並不深入，主要像是提案或是給定一些方向和策略。
2. DPO：回答較簡短，與 Before 類似，但回答的內容較深入，會給予相關舉例說明以及比較符合題意的回覆和策略。
3. ORPO：回答最詳盡，對於複雜的問題會將常見的方法與策略納入考量，同時會用不同層面的想法去思考，回答也比較有架構性。

Extra Experiments

原始版本

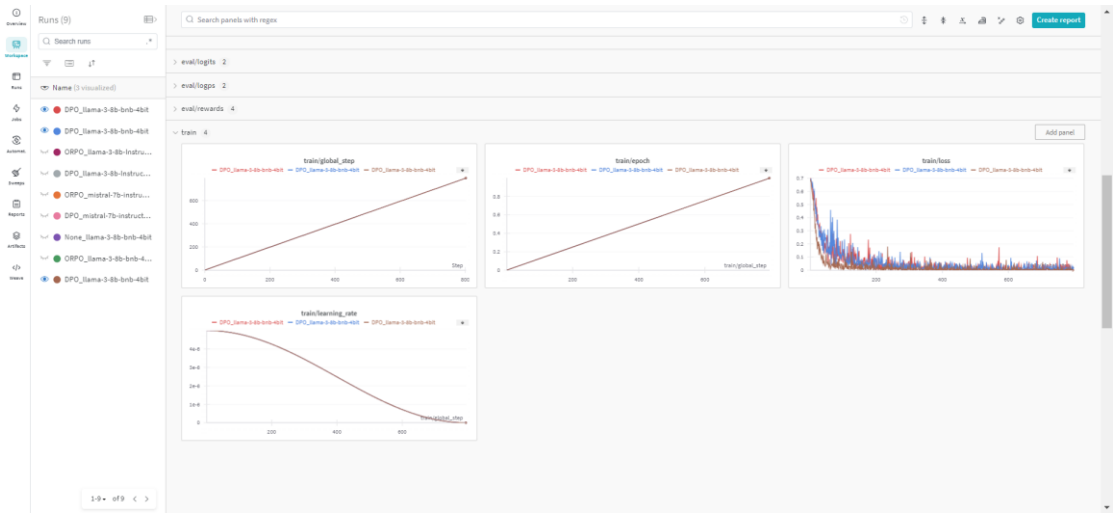
1. DPO with unsloth/llama-3-8b-bnb-4bit that epoch = 1 / beta = 0.1 (如上所示)
2. ORPO with unsloth/llama-3-8b-bnb-4bit that epoch = 1 / beta = 0.1 (如上所示)

如上圖所示

參數調整

1. DPO with unsloth/llama-3-8b-bnb-4bit that epoch = 3
2. DPO with unsloth/llama-3-8b-bnb-4bit that beta = 0.5

從上述調整可知，epoch 為訓練的回合數，越大代表訓練越多次，但也可能會有 overfit 的風險；至於 beta 值與懲罰權重有關，較高的 beta 值意味著更大的懲罰，即模型在訓練過程中更嚴格地遵從初始策略，這可能導致模型對於較大的行動空間更為保守。下圖為原始版本的 DPO 與其他兩者比較，咖啡色代表原始版本，粉色代表 beta 0.5，而藍色代表 epoch 3，從 training loss 的部分可看出，beta 為 0.5 的由於策略較為保守，因此 loss 到後面還是較其他兩者不穩定。



其他模型

1. DPO with unsloth/mistral-7b-instruct-v0.3-bnb-4bit that epoch = 1 / beta = 0.1
2. ORPO with unsloth/mistral-7b-instruct-v0.3-bnb-4bit that epoch = 1 / beta = 0.1
3. DPO with unsloth/llama-3-8b-Instruct-bnb-4bit that epoch = 1 / beta = 0.1
4. ORPO with unsloth/llama-3-8b-Instruct-bnb-4bit that epoch = 1 / beta = 0.1

透過不同的模型與原始版本進行比較，可發現在 OPPO 之間，mistral-7b-instruct-v0.3-bnb-4bit 的 loss 是變化最快的，在 reward 的部分三者差不多。在 DPO 之間 loss 差距較小難以查看，但能略為看出 mistral-7b-instruct-v0.3-bnb-4bit 的 loss 變化較快，在 reward 的部分，llama-3-8b-Instruct-bnb-4bit 的 reward 最高。而 DPO 與 OPPO 之間，loss 和 reward 差異很大。因此透過這個比對也可以發現不同 trainer 之間的相異處。

在回答方面，DPO 原始版本的回答較簡略，但使用 llama-3-8b-Instruct-bnb-4bit 後的回答較多，而 mistral-7b-instruct-v0.3-bnb-4bit 的回答最為豐富，回答內容較深入且詳細；而在 ORPO 方面，原始版本已非常詳細因此其他兩者回答的內容也都很豐富，主要差異即在回答的內容以及推理順序的不同。

Extra Experiments 的整合結果