AI HW6

洪郡辰
R11944050

June 12, 2024

Q1

DPO 提供了一個新的獎勵模型參數化方式，使得可以通過簡單的分類損失來解決標準的 RLHF 問題，從而可以直接提取對應的最優策略。DPO 方法通過將偏好學習轉換為簡單的二元交叉熵目標，避免了在微調過程中使用 RF。DPO 通過 maximum likelihood 來實現對模型的微調，而無需建立明確的獎勵模型或進行強化學習。這使得 DPO 相對於現有的強化學習方法更加簡單且易於實施。

ORPO 也是提供了一個新的獎勵模型參數化方式，透過 odds ratio 來區分模型生成的偏好和不偏好回應，從而更有效地引導模型生成符合偏好的結果。ORPO 透過學習如何回答問題時同時學會懲罰及偏好某些答案，來達到更好的效果。

ORPO 相較於 DPO 會具有計算效率方面的優勢也更為簡單，因為 ORPO 不需要額外的參考模型，這使得它在記憶體分配和計算效率方面更為高效。相對地，DPO 需要參考模型進行優化，可能增加了計算負擔。

Q2

LoRA 是一種用於優化和加速大規模語言模型訓練的方法。其核心思想是透過降低模型參數的秩 (rank) 來減少計算資源消耗，同時保持模型性能。

LoRA 的基礎在於低秩矩陣分解。這種技術將原本高維度的參數矩陣分解成兩個低秩矩陣的乘積，從而顯著減少需要訓練和儲存的參數數量。這不僅減少了計算需求，還降低了儲存空間的使用，特別適合大規模語言模型的訓練。

通過引入低秩矩陣，LoRA 在保證模型性能的同時，大幅減少了訓練過程中的計算量和儲存需求。這對於需要處理大規模數據的語言模型尤為重要，因為它可以顯著提高訓練速度，縮短訓練時間，讓模型更快地投入實際應用。

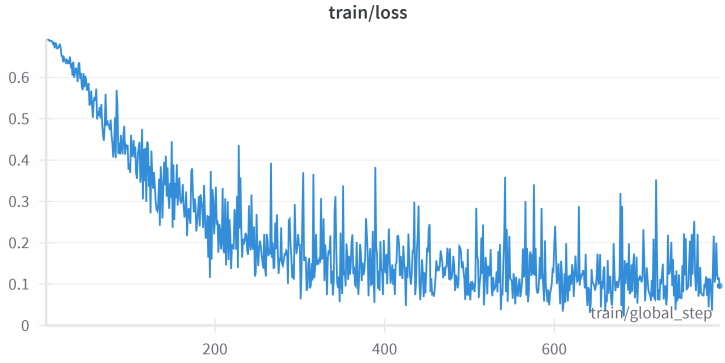
此外，LoRA 技術具有很強的靈活性，能夠適應不同大小的語言模型以及不同的應用場景。這意味著無論是小型專用模型還是大型通用模型，LoRA 都能提供性能和效率的提升。

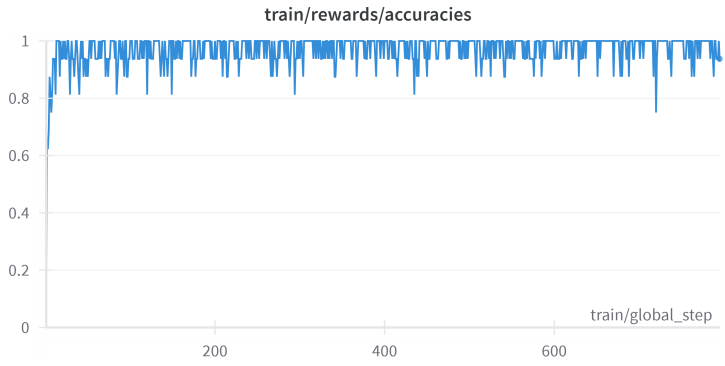
LoRA 還具備強大的兼容性，能夠與現有的優化技術和框架無縫整合。這使得使用 LoRA 時，無需對現有的訓練流程進行大規模的修改，從而降低了技術實施的門檻和成本。

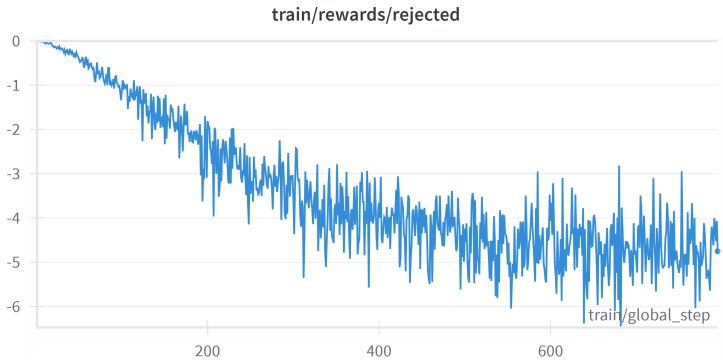
總體而言，LoRA 提供了一種高效的方法來訓練和部署大規模語言模型。通過減少模型參數的數量和計算需求，LoRA 在不犧牲模型性能的前提下，實現了資源使用的優化，特別適合在資源有限的情境下進行大規模模型的訓練和應用。這使得 LoRA 成為大規模語言模型訓練的一個重要工具，有助於推動自然語言處理技術的進一步發展。

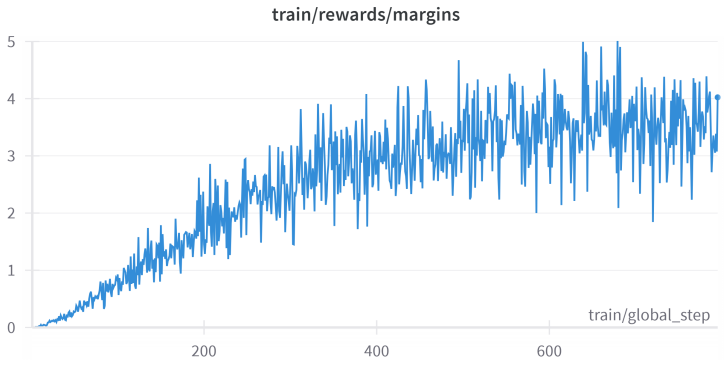
Q3

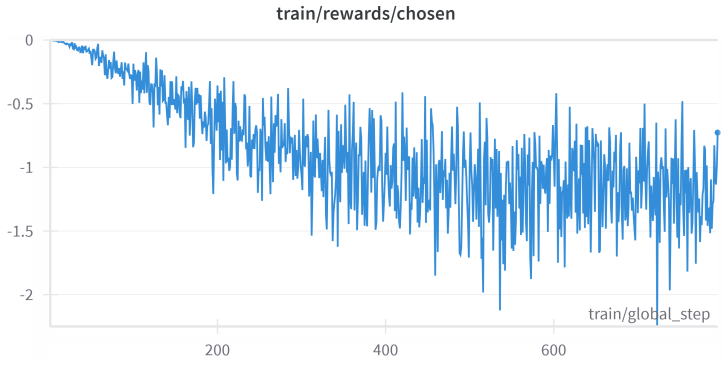
DPO



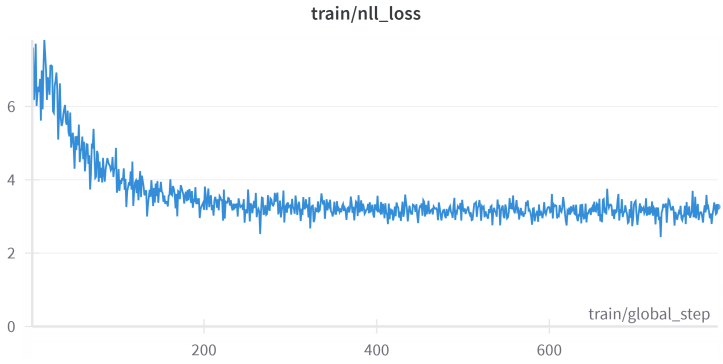


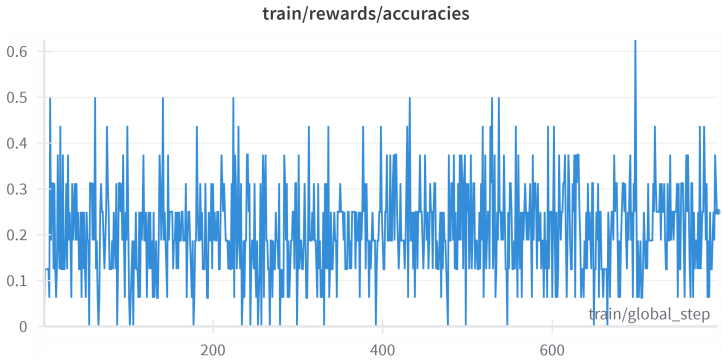

train/rewards/rejected
0
-1
-2
-3
-4
-5
-6
200
400
600
train/global_step

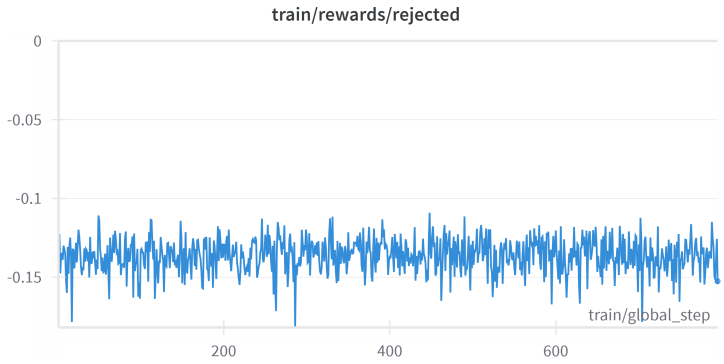

train/rewards/margins
5
4
3
2
1
0
200
400
600
train/global_step

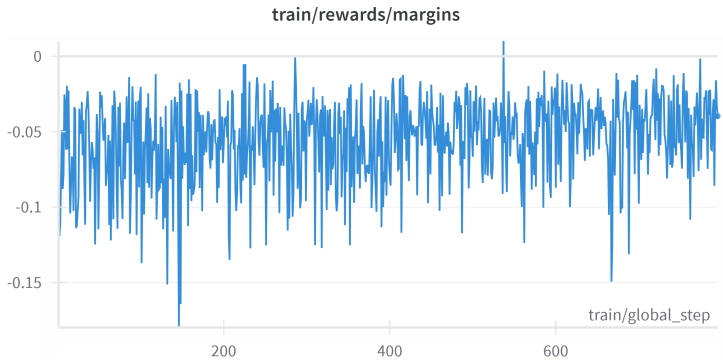

train/rewards/chosen
0
-0.5
-1
-1.5
-2
200
400
600
train/global_step

ORPO


train/nll_loss
6
4
2
0
200
400
600
train/global_step


train/rewards/accuracies
0.6
0.5
0.4
0.3
0.2
0.1
0
200
400
600
train/global_step


train/rewards/rejected
0
-0.05
-0.1
-0.15
200
400
600
train/global_step


train/rewards/margins
0
-0.05
-0.1
-0.15
200
400
600
train/global_step


train/rewards/chosen
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
200
400
600
train/global_step

Q4

我選擇 tinyllama-bnb-4bit 進行 FineTune。

在進行 finetune 前有一些問題，模型沒有辦法成功回答，而且模型時常會出現跳針狀況，重複回答一樣的內容。

在透過 DPO 微調後，模型減少了跳針狀況，有些問題能夠在回答後成功結束對話。對於原本沒辦法回答的問題有一些也變成能夠成功回答。相較於微調前，模型對於一些問題有了不同的答案。

在透過 ORPO 微調後，對於原本沒辦法回答的問題有一些也變成能夠成功回答，但是跳針狀況仍沒有得到改善。相較於微調前，模型對於一些問題有了不同的答案。

三者比較之下，我認為透過 DPO 微調後的模型表現較佳，能夠成功回答較多的問題。

Q5

我測試了兩個不同的模型，分別為 unsloth/mistral-7b-v0.3-bnb-4bit 及 unsloth/mistral-7b-instruct-v0.3-bnb-4bit。這兩個模型相較於我用來 FineTune 的 tinyllama-bnb-4bit 的表現來得更好，模型參數量也多於 tinyllama。tinyllama 的回答品質就算是在 FineTune 後還是遠弱於 mistral-7B。mistral-7B 在 FineTune 前就能夠很好地回答問題。

mitral-7b 及 mistral-7b-instruct 在資料集上的表現我覺得看起來差異不大，應該需要在比較特定的任務上才能夠比較好地比較兩者性能上的差異。