

模型訓練

- 在開始實際訓練之前，確保代碼、環境和數據都是正確的。
- 檢查訓練環境的最大運行超時時間。檢查雲平台的餘額。確保沒有任何因素會中斷訓練。
- 確保如果出現中斷，我們能夠恢復任務。中間結果不會丟失。
- 如果代碼有問題，它肯定會在訓練過程中出現。
- 如果數據不乾淨，它肯定會影響後續階段。
- 使用迭代方法。按數量級進行擴展。先訓練數百萬個標記，然後是數十億個標記，最後是數萬億個標記。
- 如果訓練結束了，會發生什麼？訓練的最終處理是否正確完成？
- 從他人的錯誤中學習，比如 Llama2 所犯的錯誤。
- 不要只修復結果，而是要解決問題的根本原因。