深度探索 V3

概覽及主要亮點

1. 模型名稱：DeepSeek-V3，一個混合專家（MoE）語言模型，擁有 6710 億參數，其中每個標記激活 370 億參數。
2. 訓練數據集：預訓練於 14.8 兆多樣且高質量的標記。
3. 核心創新：整合多頭潛在注意力（MLA）和 DeepSeekMoE 架構，並使用無輔助損失的負載平衡以提高效率。
4. 訓練效率：僅需 2.788 萬 H800 GPU 小時即可完成全面訓練。
5. 成本效益：訓練成本估計為 5.576M USD，假設每個 GPU 小時為 2 USD。

架構創新

6. Transformer-Based 框架：保留 Transformer 架構以實現可擴展性和靈活性。
7. 多頭潛在注意力（MLA）：通過壓縮鍵值緩存而不影響性能來減少推理記憶。
8. DeepSeekMoE：利用共享和路由專家的組合，以實現成本效益的訓練和高計算效率。
9. 無輔助損失的負載平衡：引入偏差項以保持平衡的專家負載，而不影響性能。
10. 多標記預測（MTP）：按順序預測每個位置的多個標記，提高數據效率和表示預計劃。

訓練框架

11. FP8 混合精度訓練：利用細粒度量化和低精度存儲來優化記憶和計算。
12. 雙管道算法：重疊計算和通信階段，減少管道氣泡並提高並行性。
13. 高效跨節點通信：使用優化的內核進行所有到所有操作，利用 NVLink 和 InfiniBand 頻寬。
14. 低精度優化器狀態：將優化器狀態存儲在 BF16 中，減少記憶消耗而不影響性能。
15. 記憶優化技術：在反向傳播期間重新計算某些操作（例如 RMSNorm）以節省記憶。

預訓練細節

16. 穩定訓練過程：預訓練期間沒有發生不可恢復的損失峰值或回滾。
17. 上下文長度擴展：將上下文長度擴展到 32K，然後在兩個階段擴展到 128K。
18. 訓練成本：預訓練需要 2.664M GPU 小時，上下文擴展 119K GPU 小時，後訓練 5K GPU 小時。
19. 標記效率：通過最小化每兆標記的 GPU 小時來確保訓練效率。

20. 高質量數據：預訓練數據集精心策劃以實現多樣性和相關性。

後訓練增強

21. 超監督微調（SFT）：將模型輸出與人類偏好對齊。
22. 強化學習（RL）：使用群組相對策略優化進行微調。
23. 知識蒸餾：整合 DeepSeek-R1 模型的推理能力。
24. 輸出風格控制：在準確性、生成長度和風格之間取得平衡。
25. 性能精煉：後訓練進一步提高基準結果。

基準性能

26. MMLU（教育基準）：達到 88.5，超越其他開源模型。
27. GPQA（一般知識）：得分 59.1，與 GPT-4o 和 Claude-3.5-Sonnet 相當。
28. 數學基準：在數學推理任務中表現出色。
29. 編碼競賽：在 LiveCodeBench 等編碼基準中表現出色。
30. 事實知識：在英語和中文事實性基準中表現優異。

推理和部署

31. 預填充階段：結合張量並行（TP4）、序列並行（SP）和專家並行（EP32）以提高效率。
32. 解碼階段：使用 EP320 與 IBGDA 進行低延遲通信。
33. 動態冗餘：動態調整專家負載以優化資源利用。
34. 階段分離：分離預填充和解碼階段以增強吞吐量。
35. 硬體利用：優化 H800 GPU，並使用 NVLink 和 InfiniBand 互連。

載荷平衡和解碼的創新

36. 偏差基礎路由：引入偏差項以動態確保平衡的專家負載。
37. 投機解碼：使用 MTP 模塊增強生成延遲。
38. 多餘專家：複製高負載專家以平衡 GPU 工作負載。
39. 節點限制路由：將標記路由限制在最多 4 個節點以減少通信開銷。
40. 無標記丟棄：確保在訓練和推理期間保留所有標記。

技術細節

41. 集群配置：在擁有 2048 個 NVIDIA H800 GPU 的集群上進行訓練。
42. 管道並行：採用 16 路並行方案以實現可擴展性。
43. 記憶足跡：通過優化記憶使用來避免昂貴的張量並行。
44. 自定義內核：開發專門的通信內核以高效處理跨節點操作。
45. 混合精度優化：結合 FP8 和 BF16 格式以實現最佳訓練動態。

評估和結果

46. 綜合基準：在教育、編碼和推理等多個領域進行評估。
47. 開源領導：成為其類別中最強的開源基礎模型。
48. 與封閉源模型的比較：性能與 GPT-4o 和 Claude-3.5-Sonnet 相當。
49. 中文知識強項：在中文事實性基準中超越領先模型。
50. 長上下文處理：在需要擴展上下文處理的任務中表現出色。

未來方向

51. 動態冗餘探索：研究更具適應性的冗餘策略。
52. 投機解碼擴展：探索 MTP 的進一步應用以加速推理。
53. 硬體共同設計：適應下一代 GPU 以實現增強性能。
54. 更廣泛的基準覆蓋：擴展評估到更多多樣化的任務。
55. 可持續性：通過算法和硬體優化進一步降低訓練成本。

本文檔提供了 DeepSeek-V3 的全面摘要，涵蓋其架構、訓練方法、基準性能及未來展望。如果需要進一步針對特定部分或其他點進行說明，請告訴我！