深度對話

A: 我一直在閱讀 DeepSeek-V3 的技術報告，對這個模型的規模感到非常驚訝。6710 億個參數，但每個標記只激活 370 億個？這是一個巨大的 MoE 架構。它是怎麼運作的？

B: 是的，這確實是一項壯舉！DeepSeek-V3 基於 Mixture-of-Experts (MoE) 框架，允許它只為每個標記激活一部分參數。具體來說，它使用 256 個路由專家，但每個標記只激活 8 個。這使它比密集模型更高效，因為密集模型的每個標記都會激活所有參數。

A: 這說得通。但它是怎麼決定激活哪些專家的？是隨機的，還是有某種路由機制？

B: 這是個好問題！路由是基於標記到專家的親和性得分。每個標記都會為每個專家分配一個得分，並激活得分最高的前 K 個專家。DeepSeek-V3 使用 Sigmoid 函數來計算這些得分，這有助於在專家之間平衡負載。

A: 原來如此，這不是隨機的，而是在訓練過程中學到的。但這不會導致專家使用不平衡嗎？我聽說這是 MoE 模型的常見問題。

B: 確實如此！專家使用不平衡可能會成為問題，但 DeepSeek-V3 引入了一種無輔助損失的策略來處理這個問題。它不會添加一個單獨的損失項來鼓勵負載平衡，而是動態調整每個專家的偏差項。如果一個專家過載，它的偏差會減少；如果它負載不足，偏差會增加。這樣可以在不影響模型性能的情況下保持負載平衡。

A: 這很聰明。所以，沒有輔助損失意味著對主訓練目標的干擾更少。但這與使用輔助損失的傳統 MoE 模型相比怎麼樣？

B: 是的。傳統 MoE 模型通常使用輔助損失來鼓勵負載平衡，但這些損失有時會影響性能。DeepSeek-V3 的無輔助損失方法避免了這種權衡。實際上，消除實驗顯示，它在編碼和數學等任務上一致優於依賴輔助損失的模型。

A: 有趣。說到編碼和數學，我注意到 DeepSeek-V3 在 HumanEval 和 MATH 等基準測試中表現出色。這裡的秘訣是什麼？

B: 這其中一個重要因素是多標記預測（MTP）目標。DeepSeek-V3 不僅預測下一個標記，還在每個位置預測多個未來標記。這使訓練信號更加密集，並幫助模型提前計劃，這對需要順序推理的任務（如編碼和數學）特別有用。

A: 等一下，所以它一次預測多個標記？這在推理過程中是怎麼運作的？它還是使用 MTP，還是僅用於訓練？

B: 在推理過程中，MTP 模塊可以被丟棄，模型會像標準的自回歸模型一樣運行。但這裡有一個很酷的部分：MTP 模塊也可以用於推測解碼，這通過並行預測多個標記並驗證它們來加速生成。

A: 這是一個很棒的技巧。所以，它就像在訓練中獲得 MTP 的好處，然後使用它來加速推理。但注意力機制呢？我看到有關多頭潛在注意力（MLA）的內容。它是怎麼融入的？

B: MLA 是另一個關鍵創新。它通過壓縮鍵值（KV）緩存來減少記憶體佔用。它不存儲完整的注意力鍵和值，而是使用低階聯合壓縮來表示它們。這顯著減少了推理過程中的 KV 緩存大小，同時保持與標準多頭注意力相當的性能。

A: 這對效率來說是一個巨大的提升。但壓縮會導致信息損失嗎？它是怎麼保持性能的？

B: 這是個好問題。壓縮設計用來保留最重要的信息，專注於捕捉鍵和值的基本特徵的潛在向量。模型還使用旋轉位置嵌入（RoPE）來保留位置信息，這有助於減少壓縮帶來的損失。

A: 明白了。所以，MLA 就是在不犧牲太多性能的情況下提高效率。但訓練呢？訓練這麼大的模型一定非常昂貴。DeepSeek-V3 是怎麼做到降低成本的？

B: 訓練效率是一個主要焦點。DeepSeek-V3 使用 FP8 混合精度框架，這減少了記憶體使用並加快了計算。它還使用雙管道算法進行管道並行，這最小化了管道氣泡並重疊計算與通信。這些優化使模型能夠在僅 2.788 百萬 H800 GPU 小時內訓練 14.8 兆個標記。

A: 這很了不起。但 FP8 訓練可能會很棘手——他們是怎麼處理精度問題的？我聽說低精度訓練可能會導致不穩定。

B: 你說得對。FP8 訓練因動態範圍有限而具有挑戰性，但 DeepSeek-V3 使用精細化量化來緩解這一問題。例如，激活值分組為 1x128 磚塊，權重分組為 128x128 塊。每組獨立縮放，這有助於處理異常值並保持訓練穩定。他們還使用高精度累積來確保關鍵操作的準確性。

A: 這說得通。所以，這是精度和效率之間的權衡，但他們已經找到了很好的平衡。但數據呢？14.8 兆個標記是一個巨大的數據集。它是什麼樣的數據？

B: 數據集多樣且高質量，重點放在英文和中文文本上。它還包括大量數學和編程數據，這有助於模型在這些領域表現出色。數據管道優化以最小化冗餘，同時保持多樣性，並使用文檔打包技術來確保數據完整性。

A: 這解釋了它在編碼和數學任務上的強大表現。但多語言性能呢？它對其他語言處理得怎麼樣？

B: 是的，DeepSeek-V3 訓練於多語言語料庫，並在包括非英文任務的基準測試（如 MMMLU）中表現良好。它在中文特別強，在中文基準測試（如 C-Eval 和 CMMLU）中超越了 Qwen2.5 等模型。

A: 這很了不起。但長上下文任務呢？我看到它支持多達 128K 個標記。它是怎麼處理這麼長的輸入的？

B: DeepSeek-V3 使用 YaRN 技術將上下文長度擴展到兩個階段：首先到 32K 個標記，然後到 128K 個標記。這使它能夠有效處理長上下文任務，如文檔摘要和檢索。它在「針在麥堆」測試中也表現良好，該測試評估長上下文理解。

A: 這是對之前模型的巨大改進。但部署呢？他們是怎麼處理這麼大模型的推理的？

B: 推理在 H800 集群上處理，GPU 通過 NVLink 和 InfiniBand 互連。部署策略將預填充和解碼階段分開，以確保高吞吐量和低延遲。他們還使用冗餘專家來在推理過程中平衡負載，這有助於保持效率。

A: 這有很多優化。但限制呢？這麼大的模型肯定有某些權衡。

B: 一個限制是部署單元大小。DeepSeek-V3 需要相對較大的集群來進行高效推理，這對較小團隊可能是一個挑戰。生成速度也有改進的空間，儘管 MTP 的推測解碼有所幫助。

A: 這說得通。但總的來說，這似乎是一個巨大的進步。DeepSeek-V3 以後有什麼計劃？他們正在探索哪些未來方向？

B: 他們正在研究幾個領域，如改進架構以支持無限上下文長度，探索額外的訓練信號來源，並增強模型的推理能力。他們還在開發更全面的評估方法來更好地評估模型性能。

A: 聽起來他們並沒有放慢腳步。謝謝你帶我走過這一切——DeepSeek-V3 確實是開源 LLM 空間中的一個遊戲改變者。

B: 絕對的！看到開源模型進步到這一步真是令人興奮。DeepSeek-V3 正在推動界限，我迫不及待地想看看他們接下來會做什麼。

A: 你提到 DeepSeek-V3 使用 FP8 混合精度訓練。我很好奇——這與 BF16 或 FP16 相比怎麼樣？FP8 真的足夠穩定來訓練這麼大的模型嗎？

B: 這是個好問題。FP8 確實更具挑戰性，因為它的動態範圍有限，但 DeepSeek-V3 使用精細化量化策略來緩解這一問題。例如，激活值分組為 1x128 磚塊，權重分組為 128x128 塊。每組獨立縮放，這有助於處理異常值並保持訓練穩定。

A: 有趣。所以，這不是簡單的 FP8 量化——這更加精細。但這不會引入管理所有這些組和縮放因子的額外開銷嗎？

B: 確實會，但這些開銷與收益相比是微不足道的。關鍵在於 FP8 降低了記憶體使用並加快了計算，這對訓練這麼大的模型至關重要。他們還使用高精度累積來確保關鍵操作（如矩陣乘法）的數值穩定性。

A: 明白了。所以，這是精度和效率之間的權衡，但他們已經找到了很好的平衡。那雙管道算法呢？它是怎麼運作的？

B: 雙管道設計用來最小化管道並行中的管道氣泡。它通過將每個工作塊分為四個組件來重疊計算和通信：注意力、全對全調度、MLP 和全對全合併。在反向傳播過程中，它進一步將計算分為「反向傳播輸入」和「反向傳播權重」，這允許更有效的重疊。

A: 這聽起來很複雜，但說得通。所以，它基本上是通過重疊計算來隱藏通信開銷。這與其他管道並行方法（如 1F1B 或零氣泡）相比怎麼樣？

B: 雙管道比 1F1B 和零氣泡有更少的管道氣泡。它還允許雙向調度，其中微批次從管道的兩端輸送。這進一步減少了閒置時間並提高了整體效率。實際上，雙管道實現了接近零的全對全通信開銷，這對擴展 MoE 模型至關重要。

A: 這很了不起。但記憶體使用呢？雙管道是否比其他方法需要更多記憶體？

B: 它確實需要稍多記憶體，因為它保留了模型參數的兩份副本，但增加是可管理的。記憶體佔用通過技術（如重新計算 RMSNorm 和 MLA 上投影）來優化，這消除了存儲中間激活的需求。

A: 啊，所以他們用一點記憶體來換取更高的效率。這似乎是一個公平的權衡。說到記憶體，他們是怎麼處理 128K 個標記這麼長的上下文長度的 KV 緩存的？

B: 這裡 MLA 真正發揮了作用。通過壓縮 KV 緩存，它顯著減少了其大小。它不存儲完整的注意力鍵和值，而是存儲壓縮的潛在向量，這些向量要小得多。這使 DeepSeek-V3 能夠處理長上下文而不會遇到記憶體瓶頸。

A: 這是一個聰明的解決方案。但注意力質量呢？壓縮會影響模型關注到正確標記的能力嗎？

B: 压縮設計用來保留最重要的信息，所以對注意力質量的影響微乎其微。他們還使用 RoPE（旋轉位置嵌入）來保留位置信息，這有助於模型理解標記的相對位置，即使鍵和值被壓縮。

A: 這說得通。所以，MLA 是一個雙贏——它減少了記憶體使用而不犧牲太多性能。但訓練數據呢？你提到它有 14.8 兆個標記。他們是怎麼確保這麼大數據集的質量和多樣性的？

B: 數據集精心策劃，包括高質量和多樣的標記。他們優化數據管道以最小化冗餘，同時保持多樣性，並使用文檔打包技術來確保數據完整性。語料庫包括英文和中文文本的混合，重點放在數學和編程樣本上。

A: 這解釋了它在編碼和數學任務上的強大表現。但多語言任務呢？它對其他語言處理得怎麼樣？

B: 是的，DeepSeek-V3 訓練於多語言語料庫，並在包括非英文任務的基準測試（如 MMMLU）中表現良好。它在中文特別強，在中文基準測試（如 C-Eval 和 CMMLU）中超越了 Qwen2.5 等模型。

A: 這很了不起。但長上下文任務呢？我看到它支持多達 128K 個標記。它是怎麼處理這麼長的輸入的？

B: DeepSeek-V3 使用 YaRN 技術將上下文長度擴展到兩個階段：首先到 32K 個標記，然後到 128K 個標記。這使它能夠有效處理長上下文任務，如文檔摘要和檢索。它在「針在麥堆」測試中也表現良好，該測試評估長上下文理解。

A: 這是對之前模型的巨大改進。但部署呢？他們是怎麼處理這麼大模型的推理的？

B: 推理在 H800 集群上處理，GPU 通過 NVLink 和 InfiniBand 互連。部署策略將預填充和解碼階段分開，以確保高吞吐量和低延遲。他們還使用冗餘專家來在推理過程中平衡負載，這有助於保持效率。

A: 這有很多優化。但限制呢？這麼大的模型肯定有某些權衡。

B: 一個限制是部署單元大小。DeepSeek-V3 需要相對較大的集群來進行高效推理，這對較小團隊可能是一個挑戰。生成速度也有改進的空間，儘管 MTP 的推測解碼有所幫助。

A: 這說得通。但總的來說，這似乎是一個巨大的進步。DeepSeek-V3 以後有什麼計劃？他們正在探索哪些未來方向？

B: 他們正在研究幾個領域，如改進架構以支持無限上下文長度，探索額外的訓練信號來源，並增強模型的推理能力。他們還在開發更全面的評估方法來更好地評估模型性能。

A: 聽起來他們並沒有放慢腳步。謝謝你帶我走過這一切——DeepSeek-V3 確實是開源 LLM 空間中的一個遊戲改變者。

B: 絕對的！看到開源模型進步到這一步真是令人興奮。DeepSeek-V3 正在推動界限，我迫不及待地想看看他們接下來會做什麼。