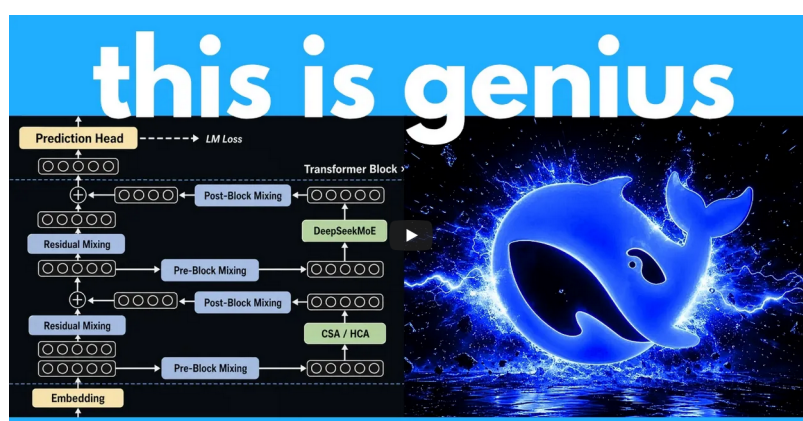


Bên trong DeepSeek V4: Làm thế nào một nhóm với nguồn lực hạn chế đã phát triển một gã khổng lồ AI với một triệu token

Bởi Frank , 21 Tháng 5 2026

Bên trong DeepSeek V4: Làm thế nào một nhóm với nguồn lực hạn chế đã phát triển một gã khổng lồ AI với một triệu token

Trong bối cảnh AI bị thống trị bởi các trung tâm dữ liệu trị giá hàng tỷ đô la và sức mạnh tính toán không giới hạn, một nhóm tương đối nhỏ của DeepSeek vừa tạo ra một cú nổ: DeepSeek V4 Pro, một mô hình Mixture-of-Experts (MoE) với 1,6 nghìn tỷ tham số trực tiếp cạnh tranh với các gã khổng lồ mã nguồn đóng trong ngành. Điều làm nên câu chuyện này đáng chú ý không chỉ là thành tích, mà còn là những giới hạn. DeepSeek được cho là hoạt động với một nhóm nhỏ hơn khoảng 40 lần so với OpenAI và không có quyền truy cập vào các chip NVIDIA tiên tiến nhất. Tuy nhiên, họ đã xây dựng một mô hình với cửa sổ ngữ cảnh 1 triệu token có thể cạnh tranh với Claude Opus 4.6 và GPT-5.4.



Điều quan trọng nhất là họ đã công khai mã nguồn mở theo giấy phép MIT và công bố một bài báo kỹ thuật với chi tiết về "kỹ thuật điên rồ" của họ.

Đây là lời giải thích về cách họ đã biến những giới hạn thành một lớp học bậc thầy về hiệu quả.

Thách thức Cốt lõi: Tại sao Bộ nhớ Một triệu Token Thường bị Phá vỡ. Trước khi thảo luận về giải pháp, chúng ta cần hiểu vấn đề. Hầu hết các LLM đều dựa vào cơ chế chú ý (attention). Khi mô hình đọc một từ (token), nó so sánh từ đó với mọi từ đứng trước nó. Đối với một triệu token, đó là một triệu phép so sánh cho mỗi token.

Điều này tạo ra hai điểm nghẽn lớn:

1. Bùng nổ Tính toán: Phép toán tăng theo cấp số nhân.
2. Phình Bộ nhớ (KV-Cache): Mô hình lưu trữ một "bảng tra cứu" của mọi từ trước đó. Với 1 triệu token, điều này lấp đầy hàng gigabyte bộ nhớ GPU đắt tiền.

Lẽ thường cho rằng cần sức mạnh tính toán thô để giải quyết vấn đề này. DeepSeek đã đặt một câu hỏi thông minh hơn: điều gì sẽ xảy ra nếu mô hình không cần nhìn vào mọi thứ?

Kiến trúc: Chú ý Lai ghép (CSA + HCA)

DeepSeek V4 không đọc văn bản một cách tuyến tính. Nó quét, nén và gọi lại theo yêu cầu. Nhóm đã giới thiệu một Kiến trúc Chú ý Lai ghép kết hợp hai chiến lược khác nhau để giải quyết vấn đề bộ nhớ.

1. Bộ Lập chỉ mục (Chú ý Thừa Được Nén - CSA)

Thay vì ghi nhớ từng từ riêng lẻ, Chú ý Thừa Được Nén (CSA) nhóm các token thành các cụm nhỏ (ví dụ: nhóm 4 token). Nó hợp nhất chúng thành một đầu vào "tóm tắt" duy nhất.

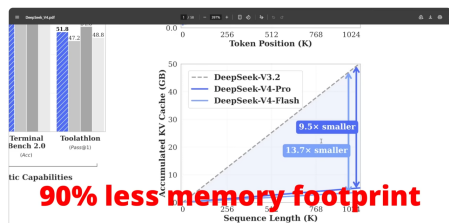
- "Bộ Lập chỉ mục Chớp nhoáng": Khi mô hình cần thông tin, nó không quét toàn bộ lịch sử. Nó thực hiện một tìm kiếm nhanh trên các bản tóm tắt này để tìm 3-4 khối phù hợp nhất.
- Kết quả: Mô hình bỏ qua 99% văn bản và chỉ tập trung vào 1% có liên quan.

2. Bức tranh Toàn cảnh (Chú ý Được Nén Mạnh - HCA)

Đôi khi bạn không cần một chi tiết cụ thể, mà là bản chất của một chương trước. Chú ý Được Nén Mạnh (HCA) tiến xa hơn bằng cách nén toàn bộ đoạn văn (128+ token) thành một vector toán học duy nhất.

- Kết quả: Độ dài chuỗi bị thu hẹp đến mức mô hình có thể đủ khả năng để xem xét mọi thứ cùng một lúc, do đó không bỏ lỡ mạch truyện chính.

Nhưng điều gì sẽ xảy ra nếu bạn cần một trích dẫn chính xác? Nếu bạn nén quá nhiều, bạn sẽ mất độ chính xác. Để giải quyết điều này, DeepSeek đã giữ lại một đường dẫn thứ ba: Chú ý Cửa sổ Trượt. Điều này giữ lại khoảng ~4.000 từ cuối cùng với độ trung thực hoàn hảo, không bị nén.



Lợi ích về Hiệu quả:

Các con số thật đáng kinh ngạc. Theo các điểm chuẩn, so với DeepSeek V3.2:

- Sức mạnh Tính toán (FLOPs): Giảm 73% (chạy bằng 27% sức mạnh tính toán).
- Bộ nhớ (KV-Cache): Giảm 90%.

Thuần hóa Quái vật: Ổn định Tín hiệu với 1,6 Nghìn tỷ Tham số

Với một nghìn tỷ tham số, quá trình huấn luyện thường thất bại. Các tín hiệu khuếch đại như micro bị rít gần loa - các giá trị bùng nổ và mô hình sụp đổ.

AI truyền thống sử dụng "kết nối dư" (đường vòng) để bỏ qua các lớp. DeepSeek cần nhiều hơn thế. Họ đã giới thiệu Manifold-Constrained Hyper-Connections (mHC) - Kết nối Siêu Hạn chế Đa tạp.

Phép loại suy: Hãy tưởng tượng nước chảy qua đường ống. Thông thường, áp suất tích tụ và đường ống vỡ. mHC buộc nước phải tuân theo một quy tắc toán học trong đó tổng áp suất luôn được bảo toàn. Tín hiệu không thể bùng nổ vì toán học ngăn cấm điều đó. Điều này đòi hỏi một thuật toán tùy chỉnh (Sinkhorn) để thực thi các quy tắc này, nhưng DeepSeek đã tối ưu hóa nó để chỉ tốn 6,7% thời gian tính toán.

Sự thay đổi "Muon": Một Trình Tối ưu hóa Mới

Trong nhiều năm, AdamW là tiêu chuẩn công nghiệp để huấn luyện mô hình. DeepSeek đã loại bỏ nó. Họ sử dụng một trình tối ưu hóa tùy chỉnh tên là Muon.

Hãy coi nó như việc lên dây đàn guitar:

- AdamW liên tục thực hiện các điều chỉnh nhỏ, thận trọng.
- Muon trước tiên thực hiện các điều chỉnh lớn, thô (để tiến gần đến cao độ chính xác) và sau đó chuyển sang các điều chỉnh nhỏ, chính xác.

Điều này cho phép DeepSeek huấn luyện nhanh hơn và ổn định hơn trên tập dữ liệu khổng lồ 33 nghìn tỷ token của họ.

Điều phối Trung tâm Dữ liệu: Che giấu Độ trễ

Một mô hình 1,6 nghìn tỷ không thể vừa trên một con chip. Nó được phân tán trên nhiều kệ. Nếu Kệ A đợi dữ liệu từ Kệ B, GPU sẽ đứng yên - lãng phí tiền bạc.

DeepSeek đã giải quyết điều này bằng cách giao tiếp chéo chéo. Họ chia nhỏ dữ liệu thành các đợt. Trong khi GPU xử lý Đợt 1, mạng đã gửi Đợt 2 ở chế độ nền. Sử dụng một ngôn ngữ cấp thấp có tên Triton và "nhân hợp nhất" tiên tiến, họ đã chứng

minh bằng toán học rằng mã của họ là hoàn hảo bằng cách sử dụng bộ giải Z3. Kết quả? Độ trễ mạng "biến mất".

Khuyến nghị cho Nhà phát triển

Dựa trên các khả năng đã được xác minh của DeepSeek V4, đây là các khuyến nghị của tôi để tích hợp vào stack của bạn:

1. Tận dụng các chế độ "Suy nghĩ"

DeepSeek V4 cung cấp độ sâu suy luận có thể cấu hình. Sử dụng "Non-Think" cho trò chuyện và tóm tắt. Chuyển sang "Max/Think High" cho lập trình, toán học hoặc các tác vụ tác nhân phức tạp. Mô hình sử dụng nhiều token đầu ra hơn ở chế độ Max, nhưng nó vẫn rẻ hơn 7 lần so với Claude Opus 4.7.

2. Sử dụng Cửa sổ Ngữ cảnh cho Tác nhân

Cửa sổ 1 triệu token lý tưởng cho các tác nhân chạy dài. Bạn có thể đưa toàn bộ cơ sở mã hoặc hàng giờ lịch sử trò chuyện vào đó. Điểm chuẩn MRCR 1M cho thấy độ chính xác truy xuất hơn 83%, nghĩa là nó hiếm khi "quên" các hướng dẫn được đưa ra khi bắt đầu cuộc trò chuyện.

3. Lưu ý về Ảo giác (Vấn đề "Biết Những Gì Nó Biết")

Điều này rất quan trọng. Các phân tích độc lập cho thấy DeepSeek V4 có tỷ lệ ảo giác 94% đối với các tác vụ có câu trả lời không xác định. Nếu bạn đặt một câu hỏi ngoài cơ sở kiến thức của nó, nó có thể sẽ bịa ra một câu trả lời thay vì nói "tôi không biết". Khuyến nghị: Luôn gắn mô hình với Tạo sinh Tăng cường Truy xuất (RAG). Cung cấp văn bản nguồn trong prompt. Không tin vào các tham số nội bộ của nó cho các sự kiện ít được biết đến, nơi nó tụt hậu khoảng 18% so với Gemini 3.1.

Đánh giá: Một Sự thay đổi Mô hình về Hiệu quả

DeepSeek V4 chứng minh rằng tương lai của AI không chỉ xoay quanh việc ai có cụm máy tính lớn nhất, mà là ai viết mã thông minh nhất. Bằng cách tập trung vào phân tán thừa và nén, họ đã dân chủ hóa quyền truy cập vào trí tuệ tiên tiến. Đối với cộng đồng mã nguồn mở, đây là một cơn sốt vàng. Bài báo tiết lộ những bí mật về cơ sở hạ tầng và ổn định đào tạo mà các phòng thí nghiệm đóng thường giữ kín. Bạn có thể tải xuống trọng số ngay hôm nay từ Hugging Face và chạy cục bộ. DeepSeek đã đánh bại các gã khổng lồ không phải bằng nhiều tiền hơn. Họ đánh bại họ bằng nhiều toán học hơn.

<https://youtu.be/XJUpuOBpT-4>